Dealing with uncertainty

Deep learning for robust animal monitoring in uncontrolled environments

Christian Lamping

Propositions

1. Uncertainty estimation in neural networks is mandatory for achieving trustworthy AI-driven decision-making.
   (this thesis)
2. The real challenge for success of AI in agriculture is the farmers' acceptance, not the technological readiness.
   (this thesis)
3. Excluding industry from research restricts societal impact.
4. Prioritizing novelty in science leads to more solutions, not necessarily better ones.
5. Chasing perfection hinders progress.
6. A response depends more on who asks than on the question itself.

Propositions belonging to the thesis, entitled

Dealing with uncertainty: Deep learning for robust animal monitoring in uncontrolled environments

Christian Lamping
Wageningen, 01 October 2024

Dealing with uncertainty

Deep learning for robust animal monitoring
in uncontrolled environments

Christian Lamping

Thesis committee

Promotor:
Prof. Dr Peter W. G. Groot Koerkamp
Professor of Agricultural Biosystems Engineering
Wageningen University & Research

Co-promotors:
Dr Marjolein Derks
Assistant Professor, Agricultural Biosystems Engineering
Wageningen University & Research

Dr Gert W. Kootstra
Associate Professor, Agricultural Biosystems Engineering
Wageningen University & Research

Other members:
Prof. Dr I. Athanasiadis, Wageningen University & Research
Prof. Dr E. Gallmann, University of Hohenheim, Germany
Prof. Dr R. da Silva Torres, Wageningen University & Research
Dr C. Kamphuis, Wageningen University & Research

This research was conducted under the auspices of the C.T. de Wit Graduate School of Production Ecology & Resource Conservation (PE&RC)

Dealing with uncertainty

Deep learning for robust animal monitoring in uncontrolled environments

Christian Lamping

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr Carolien Kroeze,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 1 October 2024
at 10:30 a.m. in the Omnia Auditorium.

Christian Lamping
Dealing with uncertainty: Deep learning for robust animal monitoring in uncontrolled environments,
194 pages.

PhD thesis, Wageningen University, Wageningen, NL (2024)
With references, with summary in English and German

DOI: https://doi.org/10.18174/671004

Contents

Chapter 1

Introduction

1.1 Modern livestock farming and its challenges

From the first domestication of animals to the present day, livestock farming has been an integral part of human culture and history. In the early stages of human history, hunting wild animals and gathering plants were the primary methods of food procurement. As humans transitioned from hunter-gatherer societies to settled agrarian communities, livestock management evolved dramatically, with the domestication of animals starting approximately 12,000 years ago (Teletchea, 2019). The adoption of systematic animal farming for food production occurred during the Neolithic period around 8,000 to 3,500 years B.C. Since then, ongoing improvements in livestock housing and feeding have laid the foundation for establishing livestock farming as a crucial component of the human food supply (Hartung, 2013). Over the last two centuries, there has been a shift in this development from sustenance-level livestock keeping, primarily for personal consumption, to the large-scale animal production systems we witness today. This change was primarily driven by the continuously increasing demand for animal products and the resulting pressure on farmers to intensify their production in order to meet this demand. Worldwide, beef production nearly doubled over the last 50 years, rising from 40 million tons in 1970 to 77 million tons in 2021. In the same period, pork production experienced an even more significant surge, escalating from 36 million tons to 120 million tons, surpassing a threefold increase. However, the primary driver of the growth in meat production was poultry. While 15 million tons of poultry meat were produced in 1970, by 2021 production had reached 138 million tons, marking an increase of more than 800% in the past five decades. The increasing demand for poultry is also evident in the global egg production, which has increased by 355% over the past fifty years (Food and Agriculture Organization of the United Nations, 2023).

For the upcoming decade, the trend of increasing demand is expected to continue, primarily driven by a growing world population. Projections suggest that the world's population will reach 8.5 billion people, marking an increase of more than 5% within the next ten years. Additionally, changes in consumer behavior affect global demand. Although per capita food consumption in high-income countries is expected to remain at current levels, particularly the consumption in middle-income countries, such as China, is estimated to expand, contributing to the globally increasing demand. Total food consumption is expected to grow by about 4% by 2029 in these countries, with 38% of the additional calories being provided by animal products. In this context, projection studies also predict a larger proportion of poultry products in the total worldwide food consumption (OECD et al., 2023). This shift towards poultry products can be attributed to two main reasons. In high-income countries, the expanding consumer focus on environmental and health awareness is driving a transition away from high intake levels of certain livestock products, particularly red meat. Persistent high-level consumption of these products is linked to a range of cardiovascular diseases and certain types of cancer (Steinfeld et al.,

2006). Instead, consumers are turning to poultry and fish, which are perceived as healthier alternatives. In lower-income countries, the growing demand for poultry meat and eggs will mainly be driven by their affordability, attributed to low production costs and a shorter production cycle compared to other animals, as well as assumed health benefits and a wider cultural acceptability. Thus, the global poultry production is expected to continue growing, accounting for half of all additional meat produced over the next decade (OECD et al., 2023).

While demand for animal products and their production continue to increase, the livestock sector has a significant impact on the world's water, land, and biodiversity resources, and is a major contributor to climate change. This combination of continuously growing and changing demands on one hand, and limited resources on the other, is the reason why most of the increase in agricultural production has come and will come from intensification rather than expansion. The fundamental principle behind intensification is to maximize the output units per unit of inputs, such as water and feed. This productivity improvement is characterized by the industrialization of agricultural processes, involving the segmentation of production stages such as feed production, animal raising, slaughtering, and processing, with each segment strategically located to minimize operating costs (Steinfeld et al., 2006).

Especially the poultry sector has been affected by this development, so that today, poultry production is almost entirely industrialized in developed countries. Advanced breeding and feeding technologies have led to an impressive rise in productivity, which is reflected in the increase in meat and egg yield per bird. Globally, the average egg yield per bird evolved from 8 kg in 1970 to 11.12 kg in 2020, representing an increase of almost 40%. In Western Europe, the egg yield in 2020 even reached 16.70 kilograms per bird. A similar trend is observable for poultry meat. Worldwide, birds slaughtered in 2020 yielded an average of 34% more meat than those slaughtered in 1970 (Food and Agriculture Organization of the United Nations, 2023). In comparison to other sectors, poultry production is the most efficient form of animal-origin food production and has the lowest land requirements per unit of output. Its short production cycle enables producers to adapt flexibly to market demands, and ongoing advancements in genetics, animal health, and feeding practices contribute to continual improvements in efficiency. While a high concentration of poultry production may cause environmental impact at the local level, the overall damage is considerably lower compared to other livestock species. In combination, these characteristics further support its dominant role in the current development of livestock production (Steinfeld et al., 2006).

Although intensification contributes to improved efficiency, including reduced land use and higher feed conversion, the structural change introduces novel challenges. One major trend in animal farming is the concentration of production, where farms tend to concentrate geographically into specialized clusters. This trend has been accompanied by a

significant increase in the average number of animals per farm, while the overall number of farms, especially those belonging to smallholders, has decreased (Steinfeld et al., 2006). Nowadays, an average laying hen farm in Germany houses approximately 20,000 animals (Agethen, 2023), and an average broiler farm holds around 29,000 animals (Thobe et al., 2021). Despite the more concentrated occurrence of emissions, new issues arise from high animal densities and the geographical proximity between farms. The proximity of thousands of animals increases the likelihood of transferring pathogens within and between populations and has further potential consequences for the evolution of zoonotic diseases (Otte et al., 2007). For example, ongoing outbreaks of highly pathogenic avian influenza (HPAI) have affected poultry and egg production in numerous countries. Consequently, predictive studies have identified animal disease outbreaks as one of the most significant risks for livestock in the coming years (OECD et al., 2023).

Moreover, the intensification of livestock farming has raised concerns about the balance between economically efficient production and animal welfare. Animal welfare encompasses both the physical and mental well-being of the animal (Brambell, 1965) and is often subject to diverse interpretations and continuous research. Within this context, the so-called "Five Freedoms", as established by the Farm Animal Welfare Council (FAWC), provide a universally acknowledged framework for evaluating animal welfare. These freedoms include freedom from hunger and thirst, freedom from discomfort, freedom from pain, injury, or disease, freedom to express normal behavior, and freedom from fear and distress (FAWC, 1993). The framework has been highly influential and serves as the foundation for various policy statements, standards, and assessment schemes for farm animals. This has contributed to a higher emphasis on welfare aspects, which often contrast with profit-oriented production processes. Practices such as caged housing systems and high stocking densities restrict animals from exhibiting their natural behaviors and limit their ability to move freely. This is a relevant issue, particularly for laying hens, for which housing in conventional cages was long recognized as the most efficient method of housing. The limited space not only restricts the birds in behaviors such as perching, scratching, and dust-bathing but also stresses the animals, potentially leading to feather pecking and cannibalism (Madzingira, 2018). In today's context, animal welfare is no longer just an ethical concern but also a critical factor in ensuring the quality of food products. The connection between animal welfare and food quality is gaining recognition among consumers, contributing to heightened public awareness of animal welfare concerns. This has led to a rising demand for food products that are produced with a focus on the well-being of the animals involved. As a result, animal welfare is becoming an important quality indicator and selling point for food products.

The changing demands of consumers, coupled with the structural shift towards industrialized livestock farming and increased technological complexity, have been identified as primary factors driving the evolving demand for labor in agriculture (Ryan, 2023). Consequently, finding adequately skilled workers to handle the diverse needs and growing

complexity of the sector is another key challenge currently faced by livestock farming. High competition, a negative public perception of the sector, and relatively low wages contribute to a low number of new entrants to the workforce, especially in developed countries. In addition, the ongoing trend of an aging workforce exacerbates this challenge. Recent studies reveal that the issue of labor shortage is not unique to the livestock sector but impacts agriculture as a whole. In the European Union, 2.5 million workers have left the agricultural sector over the last decade, and it is projected that the agricultural workforce will further decline by about 2% annually until 2030 (Ryan, 2023).

Although the challenges faced by livestock farming are complex and multifaceted, they have often sparked innovation and the development of sustainable solutions, driving continuous progress in the sector. In this regard, the role of technology has been significant, with technological advancements playing a crucial role in addressing the challenges.

1.2 Technologic answer

Early developments in agricultural intensification primarily focused on enhancing efficiency and simplifying animal handling, particularly addressing housing conditions. Characteristic of this development was the introduction of year-round indoor housing and caging for poultry. The adoption of battery cages for laying hens in the 1930s (Kawamura et al., 2023), for instance, enabled increased animal density on farms while simultaneously enhancing animal control.

In response to the increasing awareness of animal welfare, housing conditions and regulations have evolved towards more animal-friendly systems. In the case of laying hens, traditional cages have been developed into so-called enriched cages, offering more space and opportunities for natural behaviors. Since 2012, the use of conventional battery cages has been prohibited in the European Union (Directive, 1999), making it mandatory for all laying hens to be kept in enriched cages or cage-free systems. While enriched cages still severely restrict the animals' locomotion, cage-free housing systems provide hens with greater freedom of movement and behavior, thereby substantially reducing the pain experienced by the animals (Alonso and Schuck-Paim, 2022). In this regard, aviary systems are the most popular housing type. These systems consist of multiple layers and compartments that provide hens with a nesting box and perches, as well as chain feeders and drinkers. The multi-level structure maximizes space efficiency compared to traditional free-range systems, allowing the housing of large flocks while improving animal welfare by offering more space to express natural behaviors.

Although such systems represent a significant improvement from cages in terms of animal welfare, the freedom of movement for birds poses management challenges, especially concerning the handling and monitoring of the animals. Details of the monitoring-related challenges will be further elaborated in Section 1.4, as all methods developed in this thesis

consider the application in cage-free systems, specifically addressing challenges associated with this type of housing.

A significant driving force transforming all sectors of livestock farming has been, and continues to be, automation. Through innovations such as automated feeding, manure removal, and egg transportation via conveyor belts, it has become possible to handle large herds of animals, reduce manual work, and consequently increase farm productivity. Over the last century, automated solutions have been continuously developed and improved. Today, a significant portion of the tasks related to animal husbandry is automated, with an ongoing drive toward greater automation. Traditionally, this drive is mainly motivated by increased efficiency and simplified handling of animals and produced outputs. However, automated solutions have the potential to address multiple challenges currently present in livestock farming, such as labor shortage, biosecurity and animal welfare, which are discussed below.

Labor shortage
Historically, tasks like feeding and egg collection in laying hen farms were manually performed by stockmen. By automating routine tasks, such as with automated egg belts, the dependence on manual labor can be reduced, a crucial consideration given the ongoing decline in the available workforce. This technological development not only decreases the overall requirement for personnel in farm management but also alters the composition of the workforce. Higher levels of automation primarily replace low-skilled and time-consuming tasks, enabling human workers to focus on more complex activities, such as animal care, process management, or machinery maintenance, requiring different sets of skills (Gallardo and Sauer, 2018). Moreover, automation can improve working conditions for workers by reducing physically demanding tasks, potentially enhancing the attractiveness of employment in the livestock farming sector.

Biosecurity
Automation also plays a crucial role in enhancing biosecurity in livestock farming, primarily due to two aspects. First, automated solutions restrict the introduction and spread of pathogens within a farm and between farms by minimizing human contact with animals. This creates a more sterile environment, reducing the likelihood of disease transmissions and protecting the health of both animals and humans. Second, automated control of environmental conditions on the farm ensures optimal conditions for animal health, thereby reducing the risk of infectious outbreaks. For example, modern ventilation and manure removal systems in poultry farms decrease the concentration of ammonia, which otherwise could increase susceptibility to respiratory diseases (Koerkamp et al., 1995).

Animal welfare
Studies have shown that frequent contact between farm personnel and livestock, along with the associated discomfort and disturbances, often results in stress for the animals and regularly leads to injuries (Hemsworth, 2003). Minimizing human interaction, in

combination with the prevention of diseases, is therefore a crucial benefit of automated solutions to enhance animal welfare. Simultaneously, the automated provision of food, water, and fresh air, sometimes even optimized for the individual animal, can enhance living conditions more effectively than if these were manually ensured by human workers. Recent developments further focus the automated assessment of welfare and health conditions to ensure the well-being of the animals and allow early warnings in case of issues. An example of such technology is wearable sensors for cows to record feeding and milking behavior (Neethirajan and Kemp, 2021). While these systems are becoming increasingly commercially available for livestock animals in smaller herds, most assessment tools for animals housed in large groups, such as poultry, are still the subject of ongoing research (Li et al., 2020b).

In the domain of automated solutions for livestock farming, especially the automation of animal monitoring methods has recently gained attention in research and development as animal vital parameters are critical for both productivity and welfare (Ben Sassi et al., 2016). However, most of the monitoring and assessment work is currently conducted manually by farm personnel, making it labor-intensive, subjective and costly, thereby opening up an immense potential for innovative solutions.

1.3 Automated animal monitoring: From group-level observation towards individual assessment

The continuous monitoring of various production-related metrics and environmental conditions has become a common practice in modern livestock farms. Sensors and automated programs can consistently gather and store data, offering direct access to farmers and enabling advanced analyses based on the collected information. This ongoing data stream, available around the clock, allows for the identification of anomalies and potential issues by comparing current data with past measurements and predefined standards. However, traditionally, most of the evaluated data focuses on environmental conditions or productivity metrics, such as water and feed consumption, without considering information about the animals' well-being. While productivity data can indirectly provide insights into the health and well-being of the animals, this occurs at a later stage when health or welfare problems already affect eating behavior or weight development. This delay complicates the early detection of diseases or welfare-related issues.

For livestock animals housed in large groups, such as poultry, the assessment of health and welfare conditions typically occurs at a group level and is primarily executed manually by the farmer. This makes the quality of the assessment dependent on human perception, leading to a lack of standardization. Additionally, the assessments are only performed periodically through regular, manual checks in the barn, without the capability of continu-

ous monitoring. Current research tackles these issues, aiming to automate the monitoring of animal-related data. An example from the poultry domain is sound analysis, which utilizes the sounds produced by the animals as indicators of health and welfare. Studies have demonstrated that stressors, such as feather pecking and disease outbreaks, can significantly alter chicken vocalizations. This correlation enables the detection of health and welfare issues through sound recording and subsequent analysis (Astill et al., 2020).

While the development of such solutions marks a first step towards automated animal welfare monitoring in poultry, they still involve observing the animals at the flock level. Similar to analyzing consumption or productivity metrics, decisions are made based on the condition of the entire group. This implies that measures are only taken when health or welfare issues reach a magnitude significant enough to manifest across the entire population. At this point, the conditions of individual animals may have already crossed a critical threshold. Therefore, effective animal monitoring requires a shift from the group to the individual animal, as welfare is an individual experience.

Due to labor shortages and a declining workforce per individual animal, the available time and attention for each animal decreases. This is where the concept of Precision Livestock Farming (PLF) becomes significant. PLF aims to manage individual animals through automated and continuous monitoring of health, welfare, production, and environmental parameters, enabling real-time detection of abnormalities (Berckmans, 2017). This concept has been particularly facilitated by the increasing availability of sensor technologies and gained considerable attention across diverse livestock sectors in recent years (Banhazi et al., 2012).

Although scientific studies have explored the application of PLF for various animal species in livestock farming, the practical adoption varies significantly among these species. In dairy cattle, for example, numerous wearable devices for the identification, tracking, and behavior analysis of individual cows are commercially available and well-established in the market (Stygar et al., 2021). By measuring and transmitting individual health and performance-related data, these sensors have become important tools for management in modern dairy farming. In contrast, the poultry sector appears to lag behind, as commercially available monitoring solutions still focus on the entire flock rather than the individual animal (Rowe et al., 2019). Nevertheless, there are various prototype-stage monitoring tools in development that operate at the per-animal level, primarily focusing on health and welfare assessments in broilers or laying hens. A popular monitoring approach for animals kept in groups is the use of body-worn sensor technologies, where sensors are attached to the animals for continuous tracking and monitoring. These sensors can be either active, meaning they actively broadcast signals and data, or passive, without their own power supply, obtaining power from the reading device. For example, RFID-tags have been utilized to track the nesting and egg-laying behavior of hens (Rowe et al., 2019; Chien and Chen, 2018). By computing the time difference between an RFID-

tagged bird passing two readers, it is possible to compute feeding and resting times or analyze the general activity level of a chicken flock (Feiyang et al., 2016).

These body-worn sensors offer precise measurements for individual and continuous monitoring, which is valuable for pattern detection and behavior analysis. While early developments included larger devices that restricted the animals in their natural behavior, modern solutions are lightweight and do not impact the animals' freedom (Ellen et al., 2019). Therefore, these sensors are particularly suitable for monitoring and studying poultry behavior, especially in research environments where the animals are housed in small groups. Although the effectiveness of these systems has been successfully tested under conditions of commercial farms (Baxter and O'Connell, 2020), the primary drawback of body-worn sensors for poultry, hindering their application on a larger scale, is the substantial effort required to mount these systems on each bird. While the effort for large animals like cows is manageable and justifiable, it becomes infeasible for the extensive number of birds in poultry farms. The need to equip each bird with a sensor, required maintenance and the associated costs in terms of labor and materials make the application economically unattractive on a larger scale.

Nevertheless, it cannot be neglected that for a meaningful impact on animal welfare in poultry farming, it is crucial to develop solutions for automated individual monitoring that are applicable outside of research environments and transfer the technology to actual farms with large flocks. This challenge is addressed by vision-based monitoring solutions, which utilize remote sensing to monitor animals in large groups without the need for attaching sensors to each individual, making them well-suited for commercial applications. As this technology plays a key role in the methods developed in this thesis, the subsequent section will delve further into the utilization of computer vision for the automated monitoring of animals.

1.4 Computer vision for animal monitoring

Vision-based solutions offer several benefits for the automated monitoring of group-housed animals on a large scale. Firstly, employing camera sensors enables continuous measurements in a non-invasive manner. In contrast to manual methods for assessing animal conditions or wearable devices, vision-based methods neither require direct handling of the animals nor the mounting and carrying of sensors. This is crucial for ensuring the welfare and comfort of the animals on a permanent basis.

In addition to the reduced impact on the animals, scalability is another key advantage of computer vision methods. Unlike per-animal sensors, which necessitate additional installation effort and material costs for each additional animal, camera systems can monitor multiple groups after being installed once. While the cameras simply serve as sensors, acquiring the raw image data, the actual analysis happens at the software level.

This enables the utilization of group-level recordings with multiple animals while shifting to the individual for subsequent analysis.

Moreover, this separation of data collection and utilization makes vision-based systems highly flexible. Instead of developing a whole system for a single task, different assessment tasks can be performed on the same image data without requiring additional hardware. For instance, simply detecting the presence of a bird, as possible with on-body sensors, can be combined with assessment tasks that go far beyond localization, such as the classification of activity, injuries or diseases. This flexibility and scalability make vision-based monitoring approaches interesting for commercial applications in livestock farming. Since the effort and costs for installation do not linearly increase with the number of animals, economic use of these systems is possible even for large groups of animals.

While offering huge potential, the relevance of vision-based systems for poultry monitoring has emerged in recent years, mainly attributed to major technical advancements in the research field. Initial approaches primarily focused on tasks that were relatively simple compared to modern applications, particularly the segmentation of animals within images. These aimed, for example, to track individual birds (Sergeant et al., 1998; Fujii et al., 2009), estimate weights (De Wet et al., 2003), or classify behavior phenotypes (Leroy et al., 2005). As research evolved, tasks became more sophisticated , leading to the development of methods for detecting sick chickens (Okinda et al., 2019) or assessing the activity level of chickens (Aydin, 2017). However, these methods relied on traditional computer vision techniques such as edge detection or segmentation based on colors or distances obtained from 3D images. While those methods might be sufficient for straightforward tasks such as detecting animals against a clear contrasting background, they are not very robust image variations as they require manually predefined thresholds or specific image features. This makes the approaches very sensitive to image-related changes, such as varying illumination. Due to this lack of robustness, these traditional methods were quite limited in their applicability.

This changed with the advent of learning-based approaches, which facilitated new application possibilities and improved robustness. Instead of relying on hand-crafted features, these methods automatically learn to extract relevant features and patterns from training data. In this regard, supervised deep learning using convolutional neural networks (CNNs) has become the predominant approach in computer vision (Krizhevsky et al., 2017). Training a CNN involves feeding a large dataset of labeled images to the network, each image associated with a corresponding label, such as the object class or a segmentation mask. During training, the network adjusts its weights and biases, gradually reducing the error between the predicted and desired output. In this process, features relevant to the specific task are given greater importance than less relevant ones. By using large amounts of data that incorporate variations like diverse image qualities, illumination changes, or artificial image modifications, the networks learn to be resilient to such

variations, enhancing the method's robustness for practical applications. Moreover, compared to traditional approaches, trained neural networks can often generalize better to new situations, meaning they can make predictions on images they have not been trained on. This characteristic is especially beneficial for real-world applications, given the inherent variability in input images. In poultry monitoring, deep learning has been utilized in various approaches, either to enhance the robustness of existing applications or to explore new application domains that were impractical with traditional methods. For instance, CNNs have been used to classify behaviors of individuals in a group of multiple chickens (Wang et al., 2020), determine their gender (Wu et al., 2023), or detect and count laying hens in crowded cages (Geffen et al., 2020).

Despite significant advancements in automated animal monitoring through deep learning approaches, their practical applications, especially in the poultry domain, still encounter several challenges. This finding is underscored by a recent study from the laying hen sector (van Veen et al., 2023), which identified the effectiveness and validity of existing solutions as the biggest obstacle for their widespread implementation. Here, both terms relate to technological readiness, with effectiveness describing a system's ability to perform its intended task and validity referring to the accuracy of the delivered results. While there is a diverse range of obstacles named in the literature and industry, most of them can be classified under one of the following overarching challenges for computer vision for animal monitoring:

Variations in the environment and among animals
One of the most difficult challenges for vision-based systems is the ability to deal with variations. In the case of animal monitoring, these variations exist in both the environment and the appearance of the animals. While this is relevant for all real-world applications outside of controlled environments, the harsh conditions in livestock husbandry further intensify this challenge. Image recordings in a farm are influenced by factors such as widely varying illumination conditions or high animal densities, all of which must be addressed by the monitoring systems. For instance, an animal might appear in completely different colors when in shadow or in direct sunlight. However, reliable assessment of this particular animal is expected independent of the illumination conditions. Despite the effect of the environment, the visual appearance of the animals themselves can vary depending on the breed, age, or health condition. In this context, it is important to note that the mentioned factors represent only a subset of the multifaceted determinants influencing the visual appearance of an animal in an image. This diversity introduces an infinite range of potential appearances, adding an additional layer of complexity that further complicates the monitoring process.

Image quality
Environmental conditions not only cause variations in the images but also affect the overall image quality. For example, poor lighting conditions can result in overexposed or

underexposed images, making it difficult for algorithms to detect relevant features and distinguish between animals. Similarly, dust or water particles in the air disturb the sensor data and compromise the image quality. This has a significant impact on the quality of monitoring and assessment results. However, while these factors influence technical image integrity, the value of an image for animal monitoring is not solely dependent on technical characteristics. Other aspects, like motion blur, occlusions, or the pose of the animal, can compromise the usefulness of an image for a specific monitoring task, even if the image is of technically good quality. Consequently, not all images are equally useful for the task. While low quality may complicate assessments, for example, due to hardly visible features, there are also situations in which assessing an animal is not possible at all. For instance, assessing the conditions of a chicken's feet is obviously impossible if the bird is sitting and thus occluding its feet. For the training of a neural network, datasets are typically carefully curated, excluding images that are not beneficial to the task. Likewise, research results reported in publications are often based on such datasets. However, when applied in a real-world scenario such as poultry farms, recorded images are less controlled, which might lead to unusual or low-quality inputs. Therefore, automated monitoring systems must be robust enough to handle image degradation and also capable of dealing with situations where the content captured in an image is insufficient for accurate predictions.

Prediction interpretability

The interpretability of predictions made by learning-based methods is another crucial aspect to consider for the implementation of monitoring applications. In contrast to traditional computer vision methods, which rely on explicit rules for measuring predefined features, deep learning models are often considered to be "black boxes", as it is difficult to understand how they make decisions. Consequently, there is no indication of the reliability for a certain prediction. This becomes problematic, for instance, in the context of the previously described occurrence of low-quality inputs, where making a reliable prediction is not possible. In such cases, automated animal monitoring systems must be able to flag such predictions as unreliable or abstain from making a decision. As these systems are intended to have a significant impact on management decisions on a farm, it is crucial that predictions are trustworthy. If users cannot comprehend why a model makes a particular decision, they may be less likely to trust the decision, which hinders the commercial success and therefore the establishment of such systems in poultry farming.

The challenges outlined above are evident in the context of automated animal monitoring in livestock farming, but they are also relevant issues for the general application of deep learning-based computer vision. In this thesis, efforts have been made to address these challenges with a primary emphasis on poultry monitoring while also considering the adaptability of the developed methods to other domains. To illustrate the practical significance of these challenges, the assessment of plumage condition in laying hens was chosen as a use case and served as a representative example guiding the developments in the following chapters.

1.5 Plumage condition assessment of laying hens as a use case

In this thesis, the focus was on the use case of plumage condition assessment in laying hens, which is of dual relevance from both scientific and societal perspectives. It encompasses several critical challenges of computer vision, while also having significant implications for animal welfare and productivity.

The plumage condition of laying hens is a crucial indicator of their overall welfare and behavior. Poor condition can be a sign of stress, diseases, or underlying health issues (Milisits et al., 2021). Especially feather pecking and cannibalism are common plumage-damaging behaviors, representing major welfare concerns in laying hens. Thus, early detection of such issues is crucial to take effective corrective measures. The longer these issues remain undetected, the more they may exacerbate. For instance, bleeding due to feather removal has been reported to stimulate cannibalism (Blokhuis, 1989). Therefore, regular and comprehensive plumage examination is an essential part of flock management. Typically, this is done manually by handling and assessing the animals using a defined scoring scheme (Campe et al., 2018). Figure 1.1 displays exemplary plumage conditions, illustrating a three-point scoring scheme. However, this manual method has its limitations, as the quality of examination and evaluation depends on the qualifications, experience, and motivation of the assessor. Moreover, the manual process introduces stress for the animals, as it requires individual handling and assessment.

Figure 1.1: Illustration of plumage condition scores in laying hens based on a three-point scoring scheme (Knierim et al., 2016).

In contrast, the standardization and 24/7 availability of automated systems, combined with the potential for non-intrusive assessment, make computer vision systems a promising alternative to manual assessment. However, automated plumage condition assessment in a farm environment is a complex task, encountering various challenges that are also highly relevant to other types of visual assessments beyond the poultry or agricultural domain.

Firstly, plumage assessment involves an evaluation at the individual animal level, requiring the detection and visual isolation of each hen from other animals in the flock. This initial step is a fundamental prerequisite for any assessment system focusing on individual objects of interest. Additionally, the assessment of a hen's plumage goes beyond merely identifying the presence or absence of specific features, as there is no deterministic relationship between the existence of such features and a particular score. Instead, the assessment necessitates the comprehensive evaluation of each animal. This level of complexity cannot be effectively addressed using traditional computer vision methods that are based on predefined features.

Moreover, in our use case, plumage condition assessment is conducted within an uncontrolled, cage-free farm environment, where laying hens have freedom of movement. This dynamic setting includes variations in the environment, animal appearance, and image quality, demanding a robust and adaptable assessment system. Addressing these challenges, the handling of low-quality inputs and the interpretability of the provided assessments also become significant.

Consequently, plumage condition assessment faces the typical challenges outlined in Section 1.4, making it an ideal proxy for individual animal monitoring in general, but also for other vision-based monitoring systems in real-world applications. Therefore, the methods developed in this thesis were evaluated for the use case of plumage condition assessment but designed to be applicable to a broader range of visual assessment tasks.

1.6 Research objective and thesis outline

While plumage condition assessment in laying hens served as the concrete application for developing and evaluating the different methods, the broader objective of this thesis was to contribute to the advancement of automated animal monitoring in general. To advance the establishment of computer vision-based solutions for automated animal monitoring in livestock farming, it requires research that addresses the weaknesses of current systems. This involves enhancing the effectiveness and robustness of these systems, with a particular focus on their commercial application. Currently, numerous monitoring solutions exist as prototypes and demonstrate promising results in controlled environments. However, they encounter challenges in adapting to the complexities of uncontrolled commercial farms. As previously elaborated, these complexities involve monitoring individual animals within large groups, variations in environment and animal appearance, and handling diverse image qualities. Additionally, factors such as a lack of transparency in decision-making processes and limited scalability often render current systems impractical for widespread adoption.

This thesis aimed to further advance the field of individual animal monitoring by developing effective, robust, and practical methods for real-world applications in commercial

livestock farms. It focused on addressing the challenges that hinder the practical implementation of existing computer vision methods and improving the approaches' adaptability to real-world settings.

Therefore, the objective of this thesis was:

"to develop robust computer vision methods allowing the practical implementation of individual animal monitoring with a use case on plumage condition assessment in laying hens."

From this objective it was hypothesized that:

"the use of state-of-the-art deep learning methods enables robust individual animal monitoring in uncontrolled farm environments."

To test this hypothesis, a monitoring framework was developed, integrating four distinct modules primarily founded on deep learning methods. Each of these modules was explored in a dedicated chapter within this thesis, targeting challenges present in animal monitoring and thereby contributing to the overall objective. The following sections provide an outline of the chapters, while a simplified illustration of the relationships between them, presented as modules within a monitoring system, is given in Figure 1.2.

Chapter 2 presents the development of ChickenNet, a convolutional neural network for detection and assessment. The primary goal of this work was to achieve the transition from group-level monitoring to individual animal analysis, while providing assessments that are robust against environmental and animal variations. To achieve this, a learning-based approach was employed, integrating simultaneous detection and segmentation with an additional regression output for assessment tasks within an end-to-end convolutional neural network. During model training, real-world image data from a commercial farm was augmented with various techniques to enhance generalization. It was further investigated whether high image resolution and the use of depth information improve the assessment performance of the model. This aimed to quantify the trade-off between accuracy and computational efficiency, a crucial consideration for the commercial viability of an automated monitoring system. In the overall context of this thesis, ChickenNet can be seen as the backbone of the developed framework, as it initially provides animal-level predictions that are subsequently processed in the following modules.

ChickenNet, as it was developed in the initial step, provided an assessment for each image once an animal was detected. Those assessments did not yet consider the quality of the given data, nor could the model indicate the reliability of its predictions. This limitation is particularly critical in uncontrolled monitoring environments, characterized by diverse conditions and fluctuating image quality. If unknown or low-quality inputs lead to erroneous predictions, those remain unrecognized. Therefore, in Chapter 3, methods to quantify the uncertainties of predictions were investigated, which aimed to identify unreliable assessments. This involved the implementation of three different uncertainty

estimators within the ChickenNet architecture. The first approach employed an indirect estimation of uncertainty by predicting the occlusion level of a detected animal. Additionally, two methods for directly quantifying data-related and model-related uncertainty were incorporated. The hypothesis was that the estimated uncertainties would be higher for incorrect assessments compared to correct ones. Furthermore, it was hypothesized that rejecting uncertain predictions could enhance the overall assessment accuracy. To evaluate the generalizability of the developed methods beyond the use case of plumage condition assessment, they were additionally evaluated on a dataset for human age estimation.

Thorough monitoring of laying hens' plumage condition requires a comprehensive assessment. Relying on single, potentially uncertain captures from a single perspective can be limiting. Being able to estimate the uncertainty of individual assessments, however, allows for meaningful comparison and prioritization when multiple assessments are available for an animal. This principle is the foundation of the work presented in Chapter 4. Rather than relying on single images, a method that utilizes entire image sequences was developed to integrate observations from multiple viewpoints and thus generate more robust assessments. Considering the estimated uncertainties associated with each image-level assessment within the sequence, these assessments were selectively fused to generate a final output. This selective approach aimed to ignore predictions with high uncertainty, such as those arising from blurred frames or unfavorable animal poses. In this chapter, different approaches for the fusion of individual predictions were evaluated and compared to the conventional assessment on image level. It was hypothesized that the assessment results of the proposed methods would outperform image-level assessments. Furthermore, the validity of this hypothesis for different uncertainty estimators and a reduced number of available assessments within a sequence was investigated. In line with Chapter 3, the developed methods were evaluated on both plumage condition assessment and human age estimation to test their general applicability.

Animal assessment based on multiple images requires assigning the different observations to a specific individual. While straightforward for image sequences of a single animal, this task becomes challenging when dealing with simultaneous recording of multiple animals or observations captured at different times. In such cases, re-identification becomes critical, requiring the ability to distinguish individual animals from others. For laying hens, this task is particularly difficult due to the high number of individuals in a flock and their similar appearance. Chapter 5 explores the use of deep learning-based methods for animal re-identification. Here, the goal was to provide a method suitable for application in large groups where gathering training images for each individual animal is impractical. To achieve this, a neural network for similarity learning in laying hens was employed. This network was designed to recognize images belonging to the same animal by learning representations that emphasize similarities within and differences between individuals. It was hypothesized that an approach based on a transformer architecture would be able to

re-identify individual hens within the uncontrolled farm environment while outperforming traditional CNN-based architectures. Beyond the architectural comparison, the chapter investigated the effects of different data-sampling strategies during training. Moreover, to assess the practical applicability, it was evaluated how the number of distinct animals and available images per animal affect the method's performance.

Finally, Chapter 6 concludes the thesis with a general discussion, reflecting the findings of the previous chapters and their contributions to the overall objective. It also explores the broader significance of the developed methods for individual animal monitoring, discussing both their scientific and societal relevance. In addition, limitations of the current work and recommendations for future research directions are addressed.

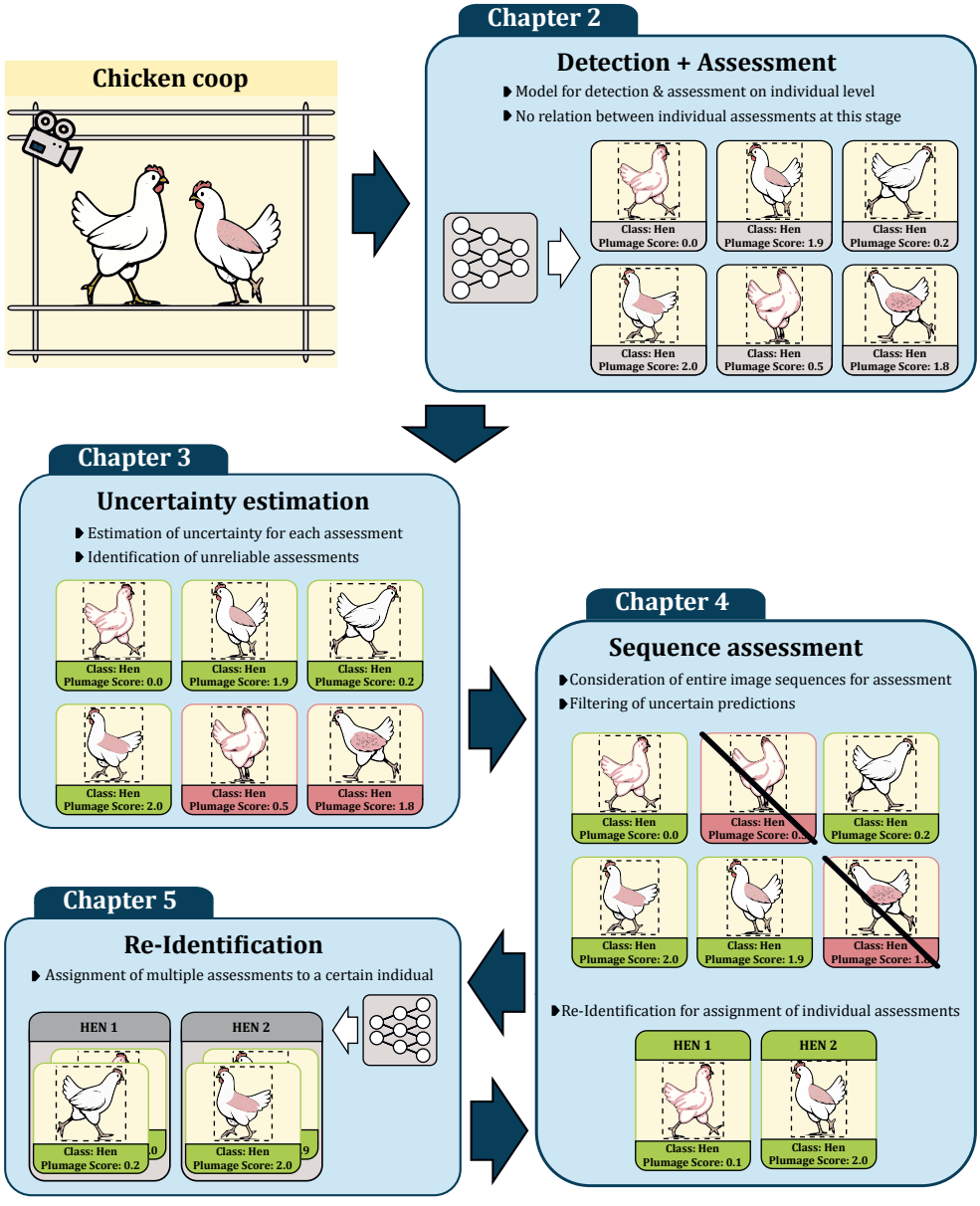


Figure 1.2: Overview of the different modules developed in the dedicated chapters of this thesis.

Chapter 2

ChickenNet - an end-to-end approach for plumage condition assessment of laying hens in commercial farms using computer vision

This chapter is based on:

Lamping, C., Derks, M., Groot Koerkamp, P., and Kootstra, G. (2022). Chickennet - an end-to-end approach for plumage condition assessment of laying hens in commercial farms using computer vision. *Computers and Electronics in Agriculture*, 194. https://doi.org/10.1016/j.compag.2022.106695

Abstract

Regular plumage condition assessment in laying hens is essential to monitor the hens' welfare status and to detect the occurrence of feather pecking activities. However, in commercial farms this is a labor-intensive, manual task. This study proposes a novel approach for automated plumage condition assessment using computer vision and deep learning. It presents ChickenNet, an end-to-end convolutional neural network that detects hens and simultaneously predicts a plumage condition score for each detected hen. To investigate the effect of input image characteristics, the method was evaluated using images with and without depth information in resolutions of 384 × 384, 512 × 512, 896 × 896 and 1216 × 1216 pixels. Further, to determine the impact of subjective human annotations, plumage condition predictions were compared to manual assessments of one observer and to matching annotations of two observers. Among all tested settings, performance metrics based on matching manual annotations of two observers were equal or better than the ones based on annotations of a single observer. The best result obtained among all tested configurations was a mean average precision (mAP) of 98.02% for hen detection while 91.83% of the plumage condition scores were predicted correctly. Moreover, it was revealed that performance of hen detection and plumage condition assessment of ChickenNet was not generally enhanced by depth information. Increasing image resolutions improved plumage assessment up to a resolution of 896 × 896 pixels, while high detection accuracies (mAP > 0.96) could already be achieved using lower resolutions. The results indicate that ChickenNet provides a sufficient basis for automated monitoring of plumage conditions in commercial laying hen farms.

2.1 Introduction

Feather pecking is a common issue in commercial laying hen flocks, which negatively impacts both animal welfare and production performance (Dixon, 2008). Underlying reasons are complex and affected by multiple factors such as nutrition, environment and genetics (Rodenburg et al., 2013). Since feather coverage of hens is reduced through feather pecking, it is compromising thermoregulation and behavior of the birds (McAdie and Keeling, 2000). Although feather damage and injuries are not exclusively caused by feather pecking, regular assessment of plumage condition provides valuable information about the overall welfare situation in a flock (Knierim et al., 2016). While modern livestock industry focuses on efficiency, which resulted in highly automated feeding and climate systems, the assessment of the chickens is usually still a manual task. The current flock situation is evaluated by examining individual animals that are randomly chosen from the flock. Regarding the large number of animals in modern laying hen farms, this is labor intensive and can lead to a lack of care for the individual. As manual assessment is time consuming, it cannot be executed continuously. Instead, it is usually based on snapshots of single situations and therefore, it does not allow for early and permanent identification of threats or negative changes on flock level. However, early and reliable detection of feather pecking is important as it increases the chance of corrective actions being effective. The later a negative development is detected, the more the associated issues exacerbate. For instance, bleeding due to feather removal has been reported to stimulate cannibalism (Blokhuis, 1989). Moreover, manual assessment results are dependent on qualification, experience, and motivation of the observer (Döhring et al., 2020) resulting in subjectivity and inter-observer differences. Therefore, reliable and standardized monitoring of individual plumage conditions cannot be guaranteed.

To improve both the efficiency as well as the quality of plumage condition assessment in laying hens, there is a strong need for automated solutions. Vision-based systems are a common approach for a variety of applications in poultry monitoring, as a single camera unit can cover multiple animals and allows continuous operation without human involvement. Most of the existing work focuses on recognition of behavioral traits or individual welfare indicators based on image analysis. For example, Leroy et al. (Leroy et al., 2005) aimed to measure the behavior of individually recorded laying hens by identifying six different behaviors (standing, sitting, sleeping, grooming, scratching, pecking). Similarly, Zhuang et al. implemented machine vision to detect sick broilers by detailed posture analysis (Zhuang et al., 2018). However, in both studies, birds were recorded in prepared experimental environments to evaluate their features. These artificial environments allow the spatial isolation of individual animals as well as the standardization of the animal pose during recording. Further, environmental conditions such as illumination can be controlled, which is not the case in a realistic farm scenario that we target in this paper.

Automated poultry monitoring in commercial farm environments is challenging due to uncontrolled conditions that give rise to a lot of variation in the images. Commonly, automated poultry monitoring is addressed in two steps (Okinda et al., 2020); detection and segmentation of individual animals, and the assessment of the animal. The first step is challenging in a commercial farm, as the animals in the camera image need to be distinguished from the background and from other individuals under changing circumstances. This requires robustness against changes in lighting as well as varying animal densities. Furthermore, birds are able to move freely, so recordings can be blurred, individual animals vary in pose, and animals frequently occlude each other. Early studies mainly used color features to detect the animal. Methods such as an ellipse-fitting model were applied to identify the bird's silhouette and extract it from the background (Kashiha et al., 2014; Okinda et al., 2019). Additionally, a common approach to address the detection of individual birds is the use of 3D vision technology. By obtaining depth information from the recorded images, the animal of interest is segmented. This method was for instance used to detect broilers in order to identify lameness (Aydin, 2017), sick animals (Okinda et al., 2019), or to predict the weight of individual birds (McAdie and Keeling, 2000). The second step, assessment, is based on information that can be acquired from the visual appearance of the detected bird. Also this step is challenging in a farm environment, as the environmental conditions influence the image lighting and quality, which complicates the recognition of relevant characteristics, such as the bird's behavior (Leroy et al., 2005) or health condition (Zhuang et al., 2018).

The above-mentioned approaches rely on manually designed image-processing steps to acquire image features for detection and assessment. To improve performance and robustness, recent poultry-monitoring studies are based on end-to-end deep learning instead, which has the advantage that both features detection as well as decision making are jointly optimized based on labeled training data. Li et al., for instance, trained a Mask Region-Convolution Neural Network (Mask R-CNN) to detect poultry preening behavior (Li et al., 2020a) Other approaches based on deep learning have been used for behavior classification of egg breeders (Wang et al., 2020), or counting of laying hens in battery cages (Geffen et al., 2020). The task of plumage condition assessment is especially challenging as it goes beyond detecting the appearance of certain distinct features that clearly specify the bird's condition. Instead, injuries and feather damages can vary in appearance and size, which requires a holistic assessment of the plumage. The existing work on automated assessment of plumage condition in laying hens is very limited. To our knowledge, there is only one study addressing this topic. (Döhring et al., 2020) used differences in color and contrast to assess plumage conditions of brown laying hens and assigned a plumage condition score to each detected hen. This approach allowed a holistic assessment but relied on the characteristic that down feathers of brown hens are lighter than the exterior feathers, which simplifies the detection of feather losses. This advantage could not be used when dealing with white hens in the study. Alternatively, color-

and contrast based methods were applied to thermal images which showed promising first results but made the evaluation much more expensive and also prone to changes in the ambient temperature (Döhring et al., 2020).

The primary objective of our study was to address the named challenges of detection and assessment and to provide an approach for automated plumage condition assessment in commercial farm environments. By introducing ChickenNet, a convolutional neural network extending Mask R-CNN, we present an integrated approach that combines detection and plumage condition assessment in a single neural network that can be trained end-to-end. This enables joint optimization of both, hen detection and plumage condition assessment in one model. Developed as a learning-based approach, it is expected to provide robustness against variation of visual appearances of birds and environmental conditions. Further, instead of detecting certain characteristics such as injuries or naked spots on the plumage, ChickenNet predicts an individual plumage condition score for each hen directly from the image. This architecture is designed to be generic, allowing the application to any other task that combines simultaneous detection and assessment. Considering the goal of applicability in commercial farms, a robust assessment performance is required while simplicity of the image acquisition system is preferred. Therefore, we evaluated ChickenNet using different input image settings. The use of conventional color images as input was compared to the addition of depth information. Moreover, different image resolutions were assessed in order to examine the trade-off between performance and complexity. Higher resolutions where expected to be more detailed, allowing an improved assessment while increasing the computational costs. Development and evaluation of our approach were done using white laying hens as example.

2.2 Material and methods

In the following paragraphs, the material and methods will be presented. The data collection methods are described in Section 2.2.1 Section 2.2.2 presents ChickenNet, our developed neural network for chicken detection and plumage condition assessment. Finally, in Section 2.2.3, the experimental setup will be described.

2.2.1 Data collection

All images for training and testing of the developed algorithm were collected in a commercial farm environment. The following sections describe the environment, the image-acquisition system and the labeling of the recorded data.

2.2.1.1 Animals and farm environment

Experiments were conducted in a free-range barn with 18,000 Dekalb White laying hens in Garrel, Germany. The barn was separated in three compartments, equipped with a

Big Dutchman NATURA Step aviary system. During data collection, the hens were able to move around freely within the barn and outside of it. The animals were not manually selected, but all birds appearing in front of the image acquisition system were recorded. To obtain images of birds at different ages with different plumage conditions, data was collected from two consecutive flocks in this barn. Recordings of the first flock were made in August and September 2020 at 59, 65 and 68 weeks of age. Images of the second flock were recorded in November 2020 at 19 and 23 weeks of age so that the overall plumage quality was higher compared to the older animals of the previous flock.

2.2.1.2 Image acquisition setup

All images have been recorded using a Stereolabs ZED 2 stereo camera. The camera simultaneously provides color and depth frames with a resolution of 2208 × 1242 pixels. The ZED 2 camera was attached by an USB port to a NVIDIA Jetson TX2 development board and operated using an Ubuntu 18.04 operating system and the ZED Software Development Kit 3.4 (SDK). Using depth restoration features of the SDK, holes in the depth images resulting from the stereo-matching of the camera were filled after recording. Thus, a fully dense depth map with a distance value for every pixel in the image was obtained for all recorded frames as shown in Figure 2.1.

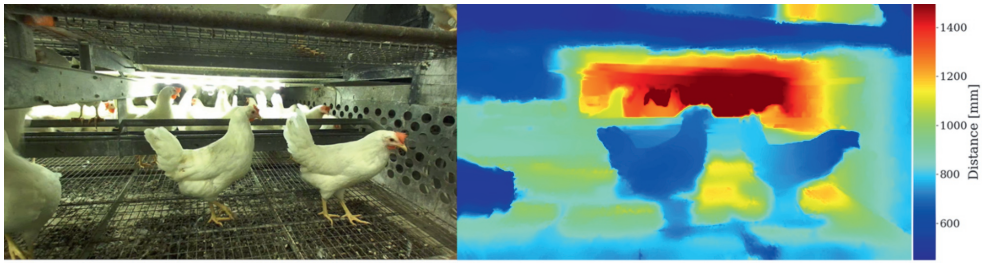


Figure 2.1: Example of a recorded scene with white hens on a tier in the aviary system. RGB-image (left) and depth image (right).

The camera setup used in the barn is illustrated in Figure 2.2. Using a portable height- and angle adjustable mount, the camera was placed in front of the aviaries' feeding line to record the hens passing by. All animals were recorded from a distance of 40 to 120 cm, depending on the animal's position in the aviary.

2.2.1.3 Image data

Each recording had a maximum length of 50 s to avoid multiple images of certain hens and overrepresentation of these hens in the dataset. In case that all animals left the field of view, the recording was stopped earlier. This resulted in videos of 15 to 50 s including 1–8 simultaneously recorded birds per video. After 2–4 recorded videos, the

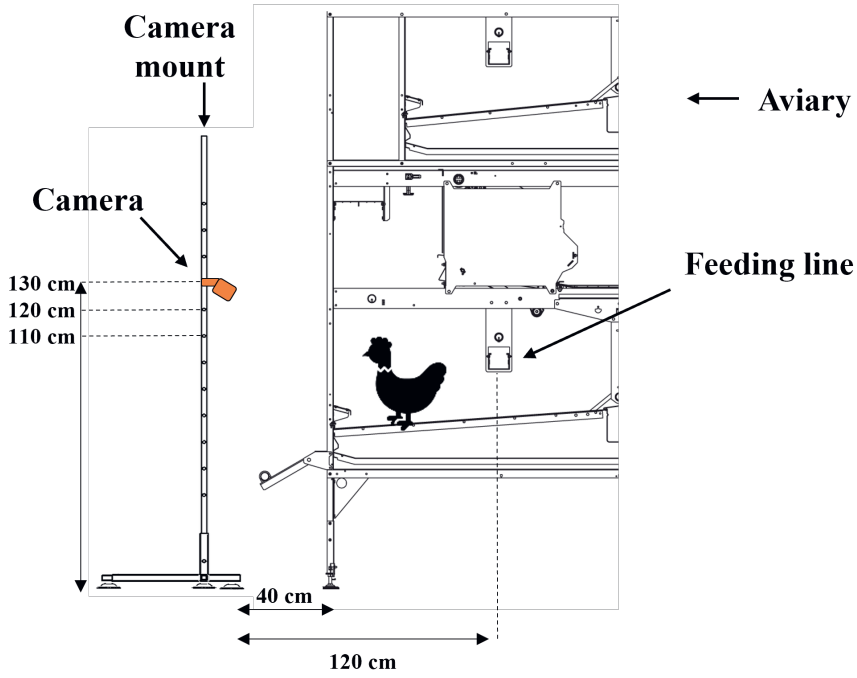


Figure 2.2: Cross-section of the aviary system with the image acquisition setup

camera was moved to another compartment of the farmhouse. During all recordings, the camera height was randomly varied between 110, 120 and 130 cm with respect to the floor to increase the variety of the collected images. This procedure was repeated for all five days of recording, resulting in a total of 52 recorded video sequences. From the raw data, individual images were removed according to the following criteria:

1. Consecutive frames with high similarity: Since recordings were made with 15 frames per second, many almost identical images were produced. Images without visible movement of the hens between consecutive frames were therefore removed from the data.

2. Images without birds: The birds were able to freely move in front of the camera and could leave the field of view. Recorded images without any birds were removed.

3. Blurred images: Due to rapid movements of the hens, some recordings were blurred. These were removed from our dataset.

For the recorded depth images, the absolute distance values were normalized so that the minimum distance of each image corresponded to 0 while the maximum distance value was 255.

2.2.1.4 Training, validation and test sets

Assignment of the images to the training, validation and test dataset was not purely random. Instead, three rules were followed:

1. Complete sequences only: All image data was obtained from video sequences leading to multiple consecutive images of a certain hen. Therefore, all images of one recorded sequence were used if the sequence was picked to guarantee that a hen from one dataset was not included in another dataset.
2. Temporal distribution: For each dataset, image sequences from all recording days were selected. Considering the influence of the bird's age on the plumage condition, this was done to reduce the imbalance of plumage condition scores in the datasets as good as possible.
3. Split ratio: Following the restrictions of rule 1 and 2, images were randomly assigned to the training, validation and test set in order to obtain an 80–10-10-split.

This procedure resulted in 1221 training images, 137 validation images and 185 images for testing.

2.2.1.5 Ground truth labeling

The manual labelling of the images was conducted using the V7 Darwin Image Annotation Tool. For each visible hen in front of the feeding line, an individual segmentation mask was drawn and a plumage condition score was assigned. The scoring was based on a manual assessment of each bird in an image, considering the three-point scale developed in (Knierim et al., 2016). Following the scoring criteria defined there, birds without any plumage damages (no featherless area) were assigned to score 0, minor damages (featherless area diameter < 5 cm) resulted in a score of 1 and for heavy damages (featherless area diameter ≥ 5 cm) a score of 2 was given. Table 2.1 illustrates the resulting distribution of the score annotations and the total number of hens among the three sets.

Table 2.1: Distribution of score annotations among the datasets.

Dataset	Score 0 annotations	Score 1 annotations	Score 2 annotations	"Undecided" cases	Total
Training	746	371	691	658	2466
Validation	119	113	29	93	354
Test	89	81	85	115	370

To annotate the scores, each frame was assessed individually based on information that was visible in the image itself. Therefore, it was possible that a different plumage conditions score was assigned to the same bird if it moved between two frames. For instance, if

a large featherless spot was clearly visible in one pose of the bird, the plumage condition would be scored as 2, while it might receive a score of 1 if the spot is not fully visible anymore in a different pose. If the plumage of a hen was not clearly assessable from the images, e.g. due to occlusions or the bird's posture, the plumage condition was marked as "undecided" (score = -1).

To quantify the reliability of the given ground truth labels, the plumage condition scores in the test dataset were labeled by a second annotator in addition to the original labeling. Then, the inter-observer reliability was measured by calculating Cohen's kappa (Cohen, 1960) for the labels. Kappa is calculated as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2.1}$$

where p_0 denotes the relative observed agreement among both raters and the hypothetical probability of chance agreement. A kappa of 1 indicates perfect agreement between the original labels and the second annotator, whereas a kappa of 0 indicates agreement equivalent to what would be expected by chance (Landis and Koch, 1977). Table 2.2 shows the original labels of assessor 1 in comparison to the labels of the second observer.

Table 2.2: Confusion matrix of original score labels and second assessment labels for a total of 255 hens in the test dataset.

		Assessment 1			**Agreement (%)**
		0	**1**	**2**	
Assessment 2	**0**	84	12	0	87.5
	1	5	57	17	72.15
	2	0	12	68	85.0

From this data, a kappa coefficient of 0.73 was calculated, which indicates a substantial agreement of both observers (Landis and Koch, 1977). In order to determine the effect of ambiguous human labels on the networks predictions, performance of the developed method was evaluated considering this double labeling of the test dataset. First, the plumage condition scores predicted by ChickenNet were compared to the original assessments of one human annotator. In a second evaluation, deviations of the predictions from the ground truth were measured considering only those annotations where both annotators agreed.

2.2.2 Chicken detection and plumage-condition scoring

In this study, a deep convolutional neural network, ChickenNet, was developed as an extension of Mask R-CNN (He et al., 2017). The network was then trained to detect laying

hens and assess their plumage condition. The following sections describe the extended network as well as the training procedure we used on our collected image data.

2.2.2.1 Network architecture

Figure 2.3 illustrates the network architecture of ChickenNet for RGB-D input images. The original Mask R-CNN consists of a backbone network for feature extraction from an RGB input image, a region proposal network (RPN) to propose regions of interest (ROI) and a detection head including a mask prediction branch in parallel with a branch for classification and box regression that process the proposed ROIs to perform instance segmentation (He et al., 2017). As a starting point for algorithm development, we used the open source Mask R-CNN implementation of Matterport (Abdulla, 2017). Our proposed method adds an additional output layer for the prediction of the feather score, as well as an additional input dimension to use RGB-D data as input. The modifications to the original architecture are described in detail in the following section.

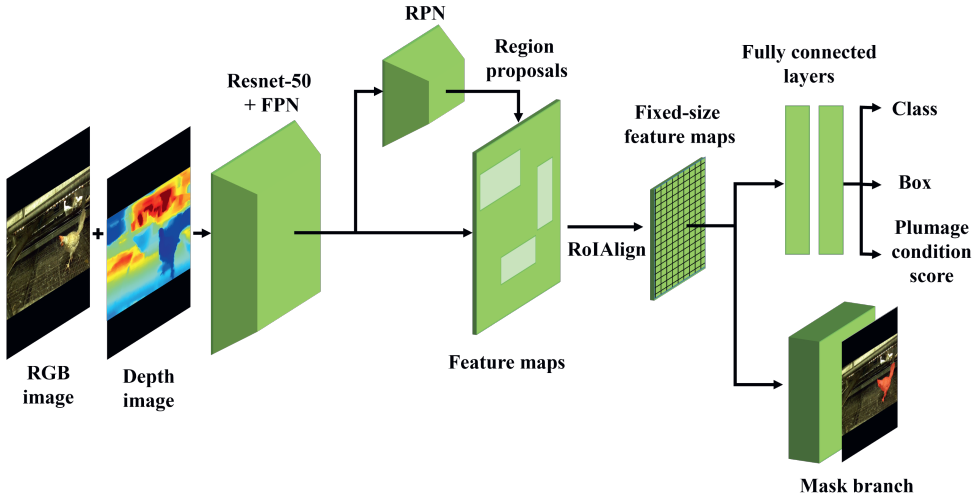


Figure 2.3: ChickenNet architecture. The model extends Mask R-CNN by an input channel for depth information and an additional output layer for the plumage condition score.

RGB-D input data

In our approach, each input image was first resized to a fixed $n \times n$ resolution with an aspect ratio of 1:1. In the experiments n was set to 384, 512, 896 and 1216. To obtain a squared form, the original 2208 × 1242 pixels images were resized and zeros were used for padding of blank areas. As feature extraction network, a ResNet-50 was used, which is the combination of a residual neural network and feature pyramid networks with 50 layers (He et al., 2016). To incorporate depth data of the four-channel RGB-D images, an additional

input channel was added to the first convolutional layer of the backbone network. Thus, depth information was treated as an additional input similarly to the three color channels. In contrast to the original Mask R-CNN implementation, we also changed the number of anchors per image generated from the RPN to 128 instead of 256 in order to speed up the processing. This could be done since the images contained a small number of birds per image. For the same reason, we reduced the number of maximum detections to 20. If the model was trained without using depth information, the first convolutional layer of ChickenNet was identical to the original Mask R-CNN implementation.

Score regression

Intuitively, the score prediction that is based on a three-point scale could also correspond to a classification problem with three different classes. However, this would ignore the ordinal relationship of the plumage condition scores. The scores are related to each other and should therefore not be seen as independent classes. In this case, the order information would have been ignored, meaning that for instance a plumage score of 2 is closer to score 1 than it is to score 0. Therefore, we formulated the plumage condition assessment as a regression problem using an additional output layer in our model. After generation of region of interest (ROI) proposals through the region proposal network (RPN), RoI alignment is applied following the original structure (He et al., 2017). This results in fixed-size feature maps for each region of interest. Subsequent to these feature maps, the original Mask R-CNN comprised two branches, one for classification and bounding box regression and one for predicting the segmentation mask. The former branch contains two fully connected (FC) layers followed by the output layer for classification and the output layer for box regression. The classification layer consists of two neurons, one for the hen class and one for background class, while the regression layer consists of four neurons for bounding box prediction. To predict a score label for each detected target, we extended this branch with an additional FC output layer *in parallel* to the existing ones. This layer consists of a single neuron and receives the same input as the classification- and box layers. A smooth ReLU activation function is applied on the layer's output to make sure the predicted score is positive. To compute the score loss of each prediction while only considering training examples with known plumage condition, two cases were differentiated. For all cases with known plumage score, we used a smooth L1 loss. If a hen's plumage condition was not clearly assessable and therefore labeled with a ground truth score of -1, the loss of this example was set to 0, independent of the predicted score. Let x denote the error between predicted value and ground truth and y denote the ground truth condition score. The score loss l was calculated as follows:

$$l(y,x) = \begin{cases} 0, & \text{if } x = -1 \\ s_{L1}(x), & \text{otherwise} \end{cases} \tag{2.2}$$

$$s_{L1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{2.3}$$

By using this case distinction, it was avoided that undetermined plumage conditions of training examples negatively influenced the loss value and thus the score prediction. For all other outputs of the network, the standard Mask R-CNN losses were used. The total loss was then calculated as the sum of the score loss, the losses corresponding to the class-, box, and mask heads as well as the RPN bounding box loss and RPN class loss.

Figure 2.4 illustrates a sample output of our algorithm for an image of a single hen and for multiple hens. For better visibility, only RGB channels without the depth channel were visualized. In both images, each individual hen was segmented and correctly classified while indicating the probability score of the class prediction. Likewise, the predicted plumage condition score for each hen was assigned to the corresponding segmentation mask.

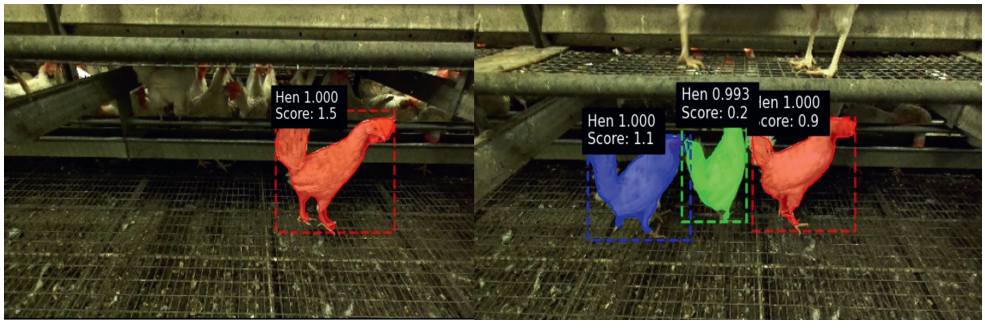


Figure 2.4: Output images including masks, class labels, class probabilities, and plumage condition scores for an image of a single hen (left) and for multiple hens (right). Different mask colors were used to visualize different individuals in each image.

2.2.2.2 Training procedure

All images from the training dataset were augmented during training in order to increase the generalization performance and robustness of the algorithm. Considering the applicability in a non-controlled farm environment, especially robustness to varying light conditions and different animal poses is important. Therefore, orientation, scale, and brightness of the training images were varied using the parameters listed in Table 2.3. For each image, between one and three of the listed augmenters were randomly chosen and combined.

Further, to avoid focus on unintended features in the background of an image for the plumage condition score, images with hens only were created in addition to the original

Table 2.3: Image augmentation parameters that were used during training.

Augmentation parameter	Parameter change and description
Reflection	50% probability (horizontal reflection only)
Scale	Uniformly random selection from range [-50%, 50%]
Brightness	Uniformly random selection from range [-40%, 40%]

training images. Based on the manually annotated segmentation masks, the birds were cropped from the original image. Background information was removed from the RGB-channels (Figure 2.5) and from the depth channel. These images were added to the training set. Note that during validation and testing, only the original images with background were used.

Figure 2.5: RGB-image from the training dataset with background (left) and without (right).

Prior to training, the model was initialized with weights obtained from pre-training the Mask- R CNN network on the Microsoft Common Objects in Context (COCO) dataset (Lin et al., 2014). Due to the fact that an additional channel for the depth information and an additional head were built, which differs from the original architecture, weights of the modified first convolutional layer and the scoring head were randomly initialized using Xavier Initialization (Glorot and Bengio, 2010). In the same way, weights of the classification head were initialized, since our model predicts only a single class (hen), while the COCO dataset contains 80 different classes. During training, the weights of the networks were optimized simultaneously using stochastic gradient descent (SGD) as optimizer. We used a learning rate of 0.02, weight decay of 0.0001 and momentum of 0.9 and trained our model for 100 epochs with a batch size of 1 on our training dataset while validating it on the validation dataset once per epoch. After training, the model was set into inference mode using the network weights that minimized the total loss on the validation set. This model was then evaluated on the test dataset.

2.2.3 Experimental setup

In our experiments, we investigated the capability of automated plumage condition assessment. Also, we examined the influence of image resolution and the addition of depth information in order to determine the required complexity of the image acquisition system.

2.2.3.1 Experiments

The experiments were conducted using images with and without depth information. First, ChickenNet was trained and validated using the RGB-D images from our datasets as described in Section 2.2.2.2. This procedure was repeated four times while varying image resolutions between 384 × 384 pixels, 512 × 512 pixels, 896 × 896 pixels and 1216 × 1216 pixels in order to examine the effect of changes in the input image resolution. After each training, detection and plumage condition assessment were evaluated on the images from the test set.

Second, the input channel modifications as described in Section 2.2.2.1 were reversed in order to exclude all depth information and only use the RGB-channels of the images. Thus, random initialization of the first convolutional layer was not needed and pre-trained weights could be used for this layer as well. Then, training and validation were repeated for the different resolutions compared in stage one. Detection and plumage condition assessment were evaluated using our test data without depth information.

2.2.3.2 Evaluation methods

ChickenNet was evaluated on two tasks. First, the instance segmentation of the hens and second, the prediction of the plumage condition. These two evaluations are described in more detail below.

Hen detection
To determine if a segmentation of a hen was correct, the intersection over union (IOU) between the ground truth mask and the predicted mask was used. In our evaluation, an IOU larger than the threshold of 0.5 indicates that the algorithm segmented the hen correctly and the prediction is marked as a true positive (TP). If the IOU was below this threshold, the detection was marked as false positive (FP). If the algorithm did not detect a hen corresponding to a ground truth mask, this mask was marked as false negative (FN). Based on the number of TP, FP and FN, the precision p and recall r were calculated following equations 2.4 and 2.5. Precision indicates which proportion of detections was actually correct, while the recall is the proportion of true hens that was detected by the network.

$$p = \frac{TP}{TP + FP} \tag{2.4}$$

$$r = \frac{TP}{TP + FN} \tag{2.5}$$

Both metrics are dependent on the selected confidence threshold of the predictions. A confidence threshold considers all predictions with a confidence level larger than or equal to this value. Thus, precision and recall usually have a trade-off relationship.

The mean average precision (mAP) and the F1-Score were used as measures that combine precision and recall to comprehensively evaluate both aspects. The mAP (Everingham et al., 2015, 2010), is the mean of Average Precision (AP) obtained for each class. As our model includes only one class, mAP is defined as:

$$mAP = AP = \sum_{k=0}^{K} (r_{\mathrm{k+1}} - r_k) p_{\mathrm{interpr}}(r_{\mathrm{k+1}}) \tag{2.6}$$

$$p_{\mathrm{interpr}}(r_{\mathrm{k+1}}) = \max_{r \geq r_{\mathrm{k+1}}} \tag{2.7}$$

Here, $p(r_k)$ denotes the precision value at a given recall value r_k considering all k confidence levels.

As shown in equation 2.8, the F1-Score is defined as the harmonic mean of precision and recall.

$$F1 = \frac{2 \cdot p \cdot r}{p + r} \tag{2.8}$$

For our evaluation, the confidence threshold that optimized the F1-score on the validation set was selected.

Plumage condition scoring

Further, we evaluated the accuracy of the plumage-condition scores predicted by ChickenNet. We wanted to determine how accurately the predicted scores match the manually assigned scores. Since the score prediction of our model is a regression problem, we evaluated how close the prediction is to the ground truth score, using the root mean square error measure (RMSE) and coefficient of determination (R^2). The assessment was analyzed for the true positive detections. Birds with unknown plumage condition (ground truth score = -1) were excluded from the analysis. In the following equations, the number of remaining detections will be defined as N.

Let y_i denote the ground truth score, $\bar{y}$ the average of the ground truth plumage scores and $\hat{y}_i$ the predicted plumage condition score by ChickenNet, we can compute the RMSE and R^2 as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{2.9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \tag{2.10}$$

To match the traditional 3-point scoring system used by the ground truth assessment, we also evaluated the classification accuracy of our method. To do this, we introduced score classes. The score class $\hat{s}_i$ of each predicted score $\hat{y}_i$ was calculated as

$$\hat{s}_i(\hat{y}_i) = \begin{cases} 0, & \text{if } \hat{y}_i < 0.5 \\ 1, & \text{if } 0.5 \leq \hat{y}_i < 1.5 \\ 2, & \text{if } \hat{y}_i > 1.5 \end{cases} \tag{2.11}$$

The accuracy a for the score prediction is thus defined as the proportion of correct score classifications where a score classification is denoted as correct if $\hat{s}_i(\hat{y}_i) = y_i$.

$$a = \frac{1}{N}\sum_{i=0}^{n} c(\hat{s}_i(\hat{y}_i), \hat{y}_i) \tag{2.12}$$

$$c(\hat{s}_i(\hat{y}_i), \hat{y}_i) = \begin{cases} 1, & \text{if } \hat{s}_i(\hat{y}_i) = \hat{y}_i \\ 0, & \text{otherwise} \end{cases} \tag{2.13}$$

Also, we analyzed the predicted classes using a confusion matrix, in which the ground truth scores were compared to the predicted score classes.

2.3 Results

The following section describes the results of our experiments. The hen detection performance as well as the plumage assessment performance of ChickenNet are presented.

2.3.1 Hen detection

The hen detection performance of our method, independent of the assigned score, is given in Figure 2.6. It presents mAP and F1-scores for four different image resolutions and using input with (RGB-D) and without (RGB) depth information.

The mAP among all tested settings was on average 97.72%, the average F1-Score was 95.78%. Using 512 × 512-pixel RGB-D images resulted in the highest mAP (98.59%)

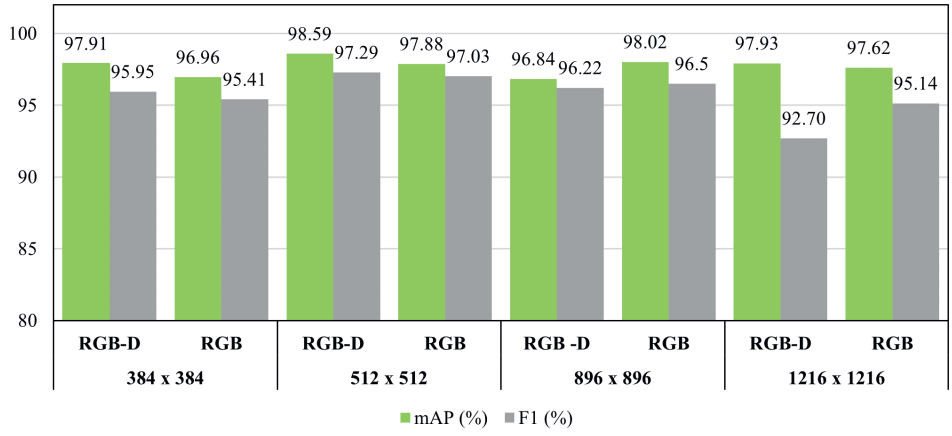


Figure 2.6: Detection results (mean average precision mAP and F1) of hens for four resolutions and depth configurations.

and F1-Score (97.29%) while RGB-D images in a resolution of 896 × 896 pixels had the lowest mAP (96.84%). The lowest F1-Score was obtained from RGB-D images with a resolution of 1216 × 1216 pixels. It was observed that, except from the latter setting, all performance differences among the resolutions were less than three percentage points. Comparing images of the same resolution, the maximum difference between images with and without depth information was 1.18 and 2.44 percentage points for mAP and for the F1-Score, respectively. It should be noted that neither color-only images nor color and depth in combination performed consistently better than the other configuration. Figure 2.7 shows predictions of the compared methods for a sample image. While the predicted plumages scores matched among all compared configurations, differences in the detection performance were visible. For this example, only the 512 × 512 RGB-D setting resulted in a correct detection of all hens in the image.

In Figure 2.8, the precision-recall curves for the compared methods are shown. Except from the RGB-D images with a resolution of 1216 × 1216 pixels, all settings had similar precisions as the recall increased. None of the evaluated resolutions outperformed the others at all levels of recall. Similarly, the incorporation of depth information did not result in a better overall performance for all settings. However, all evaluated combinations in our experiments achieved a maximum precision of at least 99% and a maximum recall of at least 97%.

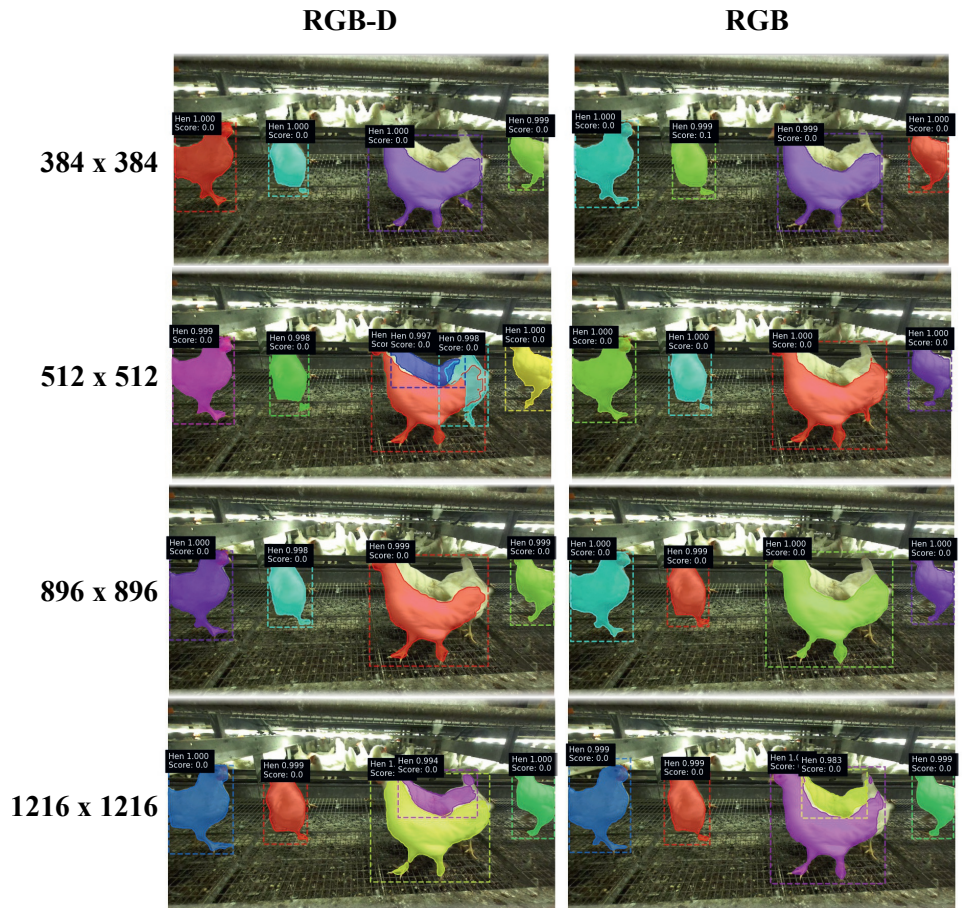


Figure 2.7: Predictions for an example image for four camera resolutions with and without depth information. The number of correctly detected hens varied among the configurations.

2.3.2 Plumage condition scoring

In our experiments, the plumage condition assessment of ChickenNet was evaluated for the different resolutions and depth settings. As described in 2.1.5, predicted plumage condition scores were compared to the original assessments of the first manual observer as well as to the manual assessments where both observers agreed. Considering both ground truth labels, Figure 2.9 presents RMSE, R2, and the accuracy for each of the evaluated settings.

Except for the experiments with a resolution of 1216 × 1216 pixels, an improved assessment performance was obtained with increasing image resolution. For both RGB-D and RGB images, the RMSE of the predicted plumage condition score decreased and R2 as

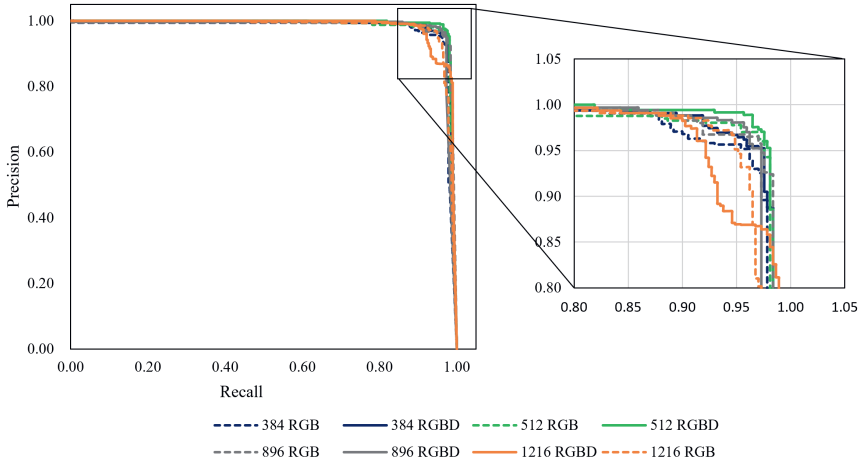


Figure 2.8: Precision-Recall curves for the different resolutions and depth settings.

well as the accuracy increased.

Furthermore, an improved assessment performance was observed if only those hens were considered where both human observers agreed about the plumage condition score. Among all evaluated settings, performance metrics were equal or better than the ones based on annotations of the single observer. The lowest RMSE of 0.26 was reached with RGB and RGB-D images in a resolution of 896 × 896 pixels. Compared to the evaluation based on labels of one observer, the accuracies in this case increased from 84.25% to 90.38% (RGB-D) and from 85.43% to 91.83% (RGB).

In the experiments, adding depth information did not improve the general scoring performance. While the incorporation of the additional depth channel resulted in a lower RMSE for a resolution of 384 × 384 pixels, the RMSE was higher for resolutions of 512 × 512 pixels and 1216 × 1216 pixels and equal to the RMSE obtained from RGB images for a resolution of 896 × 896 pixels.

For the resolution of 896 × 896 pixels, which showed the best assessment performance, the individual predictions with and without depth information were further analyzed. In Figure 2.10a, the normalized confusion matrices of the predicted score classes for both settings are shown against the ground truth annotated by observer 1. Figure 2.10b shows the evaluation considering only the matching score annotations of both observers. The diagonal elements in the matrices show the percentage of correctly predicted scores and each row sums up to 100%.

Most wrong predictions occurred when severe plumage damages (score 2) were recognized as light damages (score 1). Considering the ground truth labels given by one annotator,

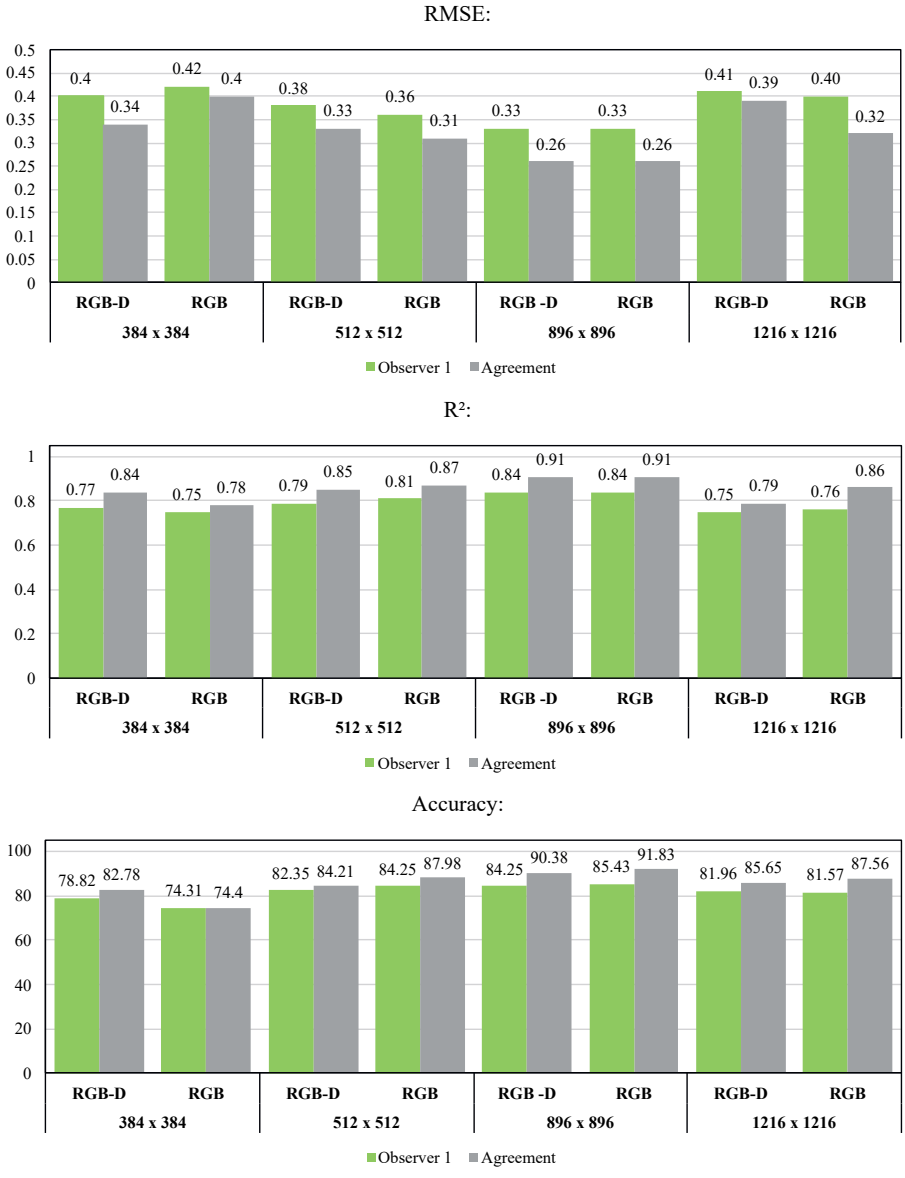


Figure 2.9: Plumage condition assessment results. RMSE, R^2 and accuracy were evaluated for the ground truth scores annotated by observer 1 as well as for the ground truth scores where observer 1 and observer 2 agreed.

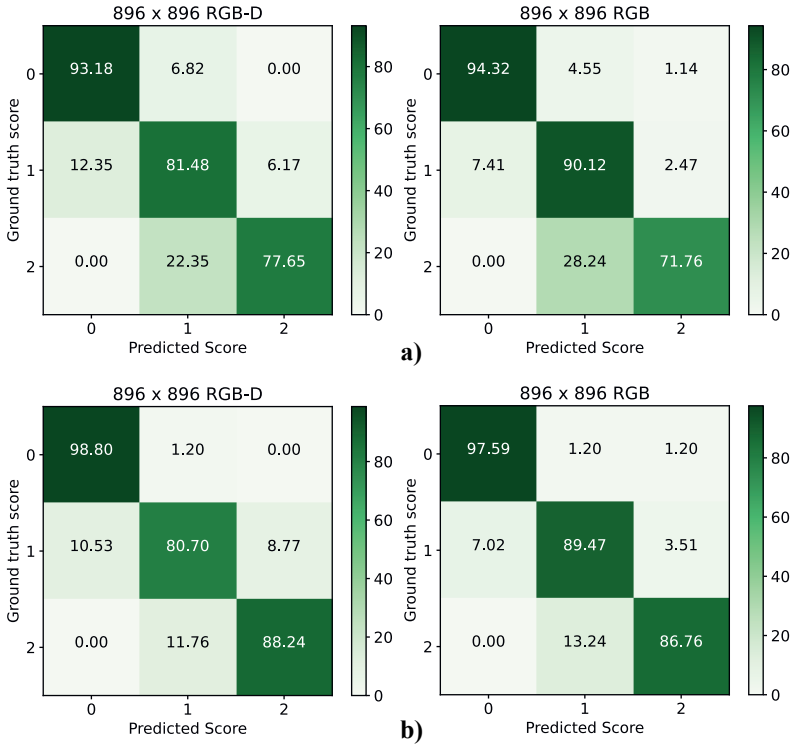


Figure 2.10: Confusion matrix of predicted score classes against the ground truth for RGB images with and without depth information. (a) Evaluation based on score annotations from observer 1. (b) Evaluation based on matching annotations of both observers

this was observed in 22.35% of cases using depth information and in 28.24% of cases without depth information. Score classes 0 and 1 were confused less frequently. There, 12.35% (RGB-D) and 7.41% (RGB) of the score 1 annotations were identified as score 0 while 6.82% (RGB-D) and 4.55% (RGB) of the score 0 annotations were identified as score 1. Differences of more than one score between ground truth and prediction were very rare during the experiments. The only occurrence was observed in the RGB setting when 1.14% of the score 0 annotations were predicted as score 2. We also noted that the proportion of birds without any plumage damages (score 0) that were correctly identified as such was 93.18% and 94.32% for RGB-D and RGB, respectively.

If only the matching annotations of both observers were considered, the percentage of correctly predicted plumage conditions increased for ground truth scores 0 and 2. Birds without any plumage damages were correctly identified as such in 98.8% (RGB-D) and 97.59% (RGB) of cases, respectively.

For birds with severe plumage damages, 88.24% (RGB-D) and 86.76% (RGB) of the

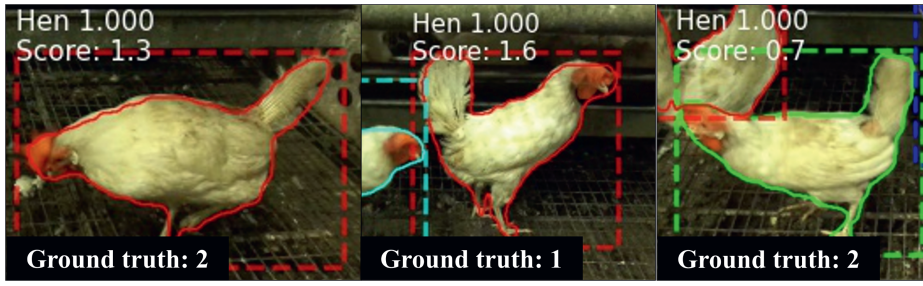


Figure 2.11: Examples for false score predictions including class predictions, class probability, predicted plumage scores and ground truth scores. For better visibility, masks were disabled and detections were cropped.

predictions were correct. The proportion of score 1 annotations that were identified as score 2 increased for both RGB-D and RGB, which resulted in a slightly reduced percentage (-0.78 for RGB-D and -0.65 for RGB) of correctly classified birds with light plumage damages. Visualized examples for false assessments are given in Figure 2.11

The results given in Figure 2.10 indicate that false score predictions were mainly due to the determination of the plumage damage severity and not due to the distinction between hens with plumage damages and those with intact plumage. Therefore, Figure 2.12 shows the normalized confusion matrices of the predicted score classes if the damage-indicating scores 1 and 2 were combined. Results based on one observer are presented in Figure 2.12a while results based on matching annotations of both observers are presented in Figure 2.12b. For both settings, RGB-D and RGB, misclassifications were significantly reduced compared to the original evaluation. Using ground truth labels of one observer, hens with any plumage damage (so score 1 or 2) were correctly identified in 93.98% (RGB-D) and 96.39% (RGB) of the cases. If only matching labels of two observers were considered, this proportion increased to 95.2% (RGB-D) and 96.8% (RGB), respectively.

2.4 Discussion

2.4.1 Influence of input image characteristics

The results showed that the influence of the image resolution was stronger on the plumage condition assessment than on the detection. Hens could be reliably detected in all evaluated resolutions for which a minimum mAP of 96.84% was observed. Regarding the assessment performance, the RMSE of the score predictions and the score class accuracy improved up to the optimal resolution of 896 × 896 pixels, after which the performance decreased again. This drop in accuracy is consistent with the results of studies such as

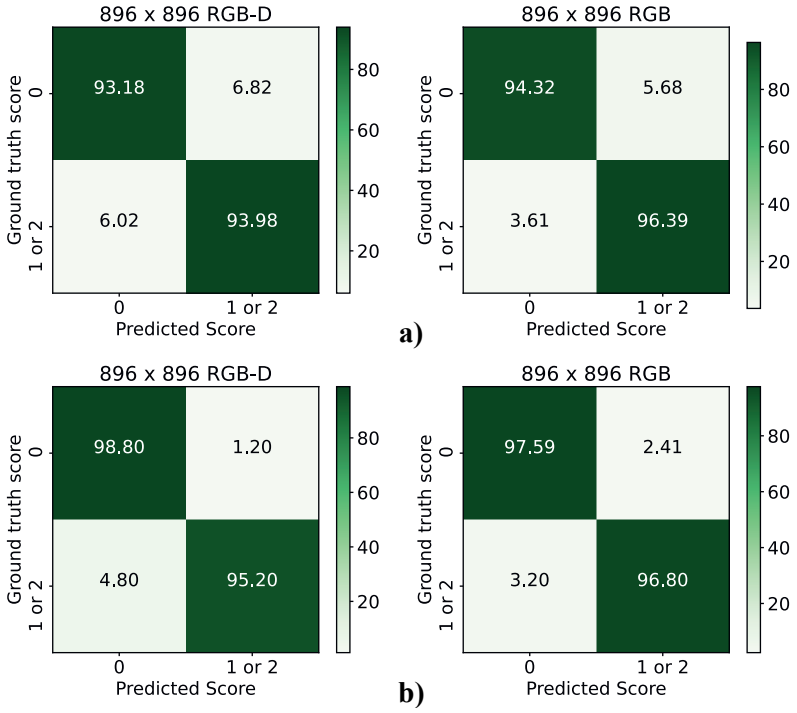


Figure 2.12: Confusion matrix of predicted score classes with summarized classes 1 and 2. (a) Evaluation based on score annotations from observer 1. (b) Evaluation based on matching annotations of both observers.

Sabottke and Spieler (2020) and Shao et al. (2020) in which it was found that increasing image resolution for CNN training improves performance of the network only up to a certain level. For instance, identification of different radiology findings based on images showed best results at resolutions between 256 and 448 pixels per dimension. Higher and lower image resolutions decreased detection performance. This was reasoned to be caused by a trade-off between the detection of small features in the images and the generalization capability of the model. In general, receptive field coverage of the convolutional layers reduces as the image resolution increases which results in capturing less high-level information (Luke et al., 2019). This leads to a drop in accuracy beyond a certain resolution. The observation that detection performance was much less influenced by image resolution during our experiments might be explained by the recorded scenarios. While the assessment is often based on small image features that are not visible in lower image resolutions, detection focuses the full bird which is much larger in size. In addition, there was a clear contrast between birds and background in the recorded images and no other objects similar to a hen were around. This makes the detection task considerably simpler compared to the plumage condition assessment.

Further, our experiments showed that neither the detection of laying hens nor the individual plumage condition assessment were clearly improved by using depth information in addition to conventional color images. Even though previous studies have shown that depth information can increase object detection performance (Eitel et al., 2015; Lenz et al., 2015), others also revealed that this benefit was not observable for all types of objects and dependent on the way that color and depth information were fused (Ophoff et al., 2019).

For instance, improved object detection through the usage of depth information might be achieved if depth is not treated as an additional channel but processed in a separate CNN-branch. Studies revealed that the perceptual structures of RGB and depth images are different so that geometric silhouettes of objects hidden in depth data may not be fully revealed if used as an additional channel along with the color channels (Bo et al., 2011; Zhu et al., 2020). In addition, the use of RGB-images in our experiments already resulted in a minimum mAP of 96.84% over all evaluated image settings, providing little potential for improvement. This could explain the observed deviations in detection performance between RGB and RGB-D images that were less than three percentage points and in both directions.

2.4.2 Plumage condition scoring

Considering annotations where both observers agreed, it was revealed that 91.83% of all plumage score predictions were correct and an RMSE of 0.26 was observed using the optimal image configuration with RGB-images in a resolution of 896 × 896 pixels. This experiment resulted in an R2 of 0.91, which indicates that 91% of the variance in the ground truth scores is explained by the predictions of ChickenNet.

Results of the plumage condition assessment in our experiments showed a better performance of ChickenNet for hen plumages labeled with a score of 0, while more confusion was observed for ground truth labels 1 and 2. A cause for this variation might have been ambiguous human scoring of the plumage conditions. This was indicated by the inter-observer variations test in which a kappa coefficient of 0.73 for the test dataset was calculated. While annotating the ground truth labels, borderline cases occurred, meaning that for some plumage damages is was hard to manually distinguish between light or severe. In contrast, hens with completely intact plumages were easier to recognize as such which led to more reliable ground truth labels. This finding is in line with the assessment evaluation which distinguished between birds annotated by one human observer and those with matching annotations of two observers. Disagreement of the two observers indicates an unreliable plumage score annotation. In this case, ChickenNet cannot be expected to agree with the annotation in this case. In contrast, if both observers agreed about the plumage condition, the ground-truth is expected to be more certain. It was assumed that the reduction of unreliable annotations in the test data would increase the

assessment performance of ChickenNet. Results showed this to be true as predictions of ChickenNet matched better with the annotations where both observers agreed than with the ones based on annotations of the single observer. For the optimal configuration using RGB images in a resolution of 896 × 896 pixels, accuracy was increased from 85.43% to 91.83

2.4.3 Practical applicability

As this study aimed to provide an approach for automated plumage condition assessment that can be applied in a practical farm scenario, the feasibility and benefit of the developed method need to be considered while discussing the results.

In our experiments, similar performances for RGB and RGB-D images were observed. As discussed in 2.4.2, processing depth data in a separate CNN-branch might increase the benefit of this information.

However, this approach would result in higher computing power as color and depth images are processed in two parallel network branches. Considering the already high detection performance of our model, this additional complexity would be contradictory to a simple on-farm application without providing significant performance improvements. In contrast, using RGB only avoids the additional input channel of the neural network and allows the application of traditional color images. This reduces camera and computation costs as well as data processing which is required to calculate depth data from recorded images.

Further, our approach was developed in order to support farmers by providing continuous information about plumage conditions in the flock. This particularly requires the detection of changes on flock level. For instance, a suddenly increasing proportion of birds with severe plumage damages in the farm needs to be identified. In our experiments, ChickenNet was able to predict the correct plumage condition of individual birds with an accuracy of up to 91.83% over all plumage condition classes. Most false predictions were due to the confusion of plumage condition scores 1 and 2. However, a reliable differentiation between hens with (score 1 and 2) and without (score 0) plumage damages would already provide a benefit for the identification of an increasing number of plumage damages in a flock. Considering only these two plumage condition states, accuracies of ChickenNet were even higher. Using RGB images in a resolution of 896 × 896 pixels, 97.59% of birds with intact plumage and 96.8% of birds with feather damages were correctly identified. Provided that enough birds from a flock are being recorded, ChickenNet is able to support the plumage condition assessment on flock level.

2.4.4 Future improvements

To provide a reliable tool for the assessment of a whole flock, the assessed birds need to form a representative sample of the flock. Overrepresentation of certain birds in the

recordings have to be avoided. This could be obtained by either using a mobile recording system which can be moved within the barn, or by re-identification of individual animals. In this case, plumage condition assessments could be assigned to the individual which would indicate that a bird has already been assessed.

Another future improvement is related to ambiguous and unknown plumage conditions. In the experiments it was observed that even for humans, it may be hard to manually assess the hens' plumage score based on single images. Some bird detections might be not suitable for plumage condition assessment, meaning that birds can be occluded or plumage conditions might be unclear due to the pose of the animal. During training, birds were manually annotated and these unclear cases were marked. However, as ChickenNet predicts a score for all detected hens, occluded and unrecognizable plumages would compromise the assessment in a practical farm application. To reduce these cases, multiple cameras with multiple perspectives might be considered to capture different views of one hen. An alternative to address this issue might be the alteration of the model architecture to additionally determine the suitability of each detection for plumage condition assessment. Currently, our approach is designed as an end-to-end solution that detects, segments and assesses hens simultaneously. Although this is an efficient design, it does not allow the identification or filtering of detections that are inadequate for assessment. Separating detection and assessment using a two-stage approach would allow filtering prior to assessment. Furthermore, as we observed that an accurate detection of the hen is feasible using lower resolutions, detection and filtering could be executed using low input resolution and a less complex model. In a second step, the plumage condition of the filtered hens could be assessed using a more sophisticated model. Also, to further evaluate the general applicability of the presented method, extending the experiments to hens of other colors is essential.

2.5 Conclusions

n this study, we presented ChickenNet, a deep convolutional neural network that detects and segments hens while simultaneously providing a holistic assessment of the plumage by computing a plumage condition score for each detected hen.

The best overall performance in our experiments was obtained using RGB images in a resolution of 896 $\times$ 896 pixels. This configuration resulted in a mAP of 98.02%, while 91.83% of all score predictions were correct with an RMSE of 0.26. We revealed that using color information only was sufficient to fulfill the task of hen detection and assessment, as no relevant improvements were obtained through incorporation of depth information. In our experiments, detection performance was barely influenced by the image resolution. Among the evaluated resolutions, differences in assessment accuracy were less than three percentage points. Assessment performance was more dependent on the image resolution and the plumage assessment improved as the image resolution increased until 896

× 896, with a small decrease for 1216 × 1216 pixels. Errors in the prediction of the condition score were mainly due to the determination of the plumage damage severity (differentiation between score classes 1 and 2) and less to the distinction between hens with plumage damages and those with intact plumage (class 0 vs. class 1 and 2). Finally, we conclude that ChickenNet provides a basis that can be used to establish automated plumage condition monitoring of white laying hens in commercial farms.

Chapter 3

Uncertainty estimation for deep neural networks to improve the assessment of plumage conditions of chickens

This chapter is based on:

Lamping, C., Kootstra, G., and Derks, M. (2023). Uncertainty estimation for deep neural networks to improve the assessment of plumage conditions of chickens. *Smart Agricultural Technology*, 5. https://doi.org/10.1016/j.atech.2023.100308

Abstract

The combination of computer vision with deep learning has become a popular tool for automation of labor-intensive monitoring tasks in modern livestock farming. However, uncontrolled and varying environmental conditions, which usually prevail in farmhouses, influence the performance of vision-based applications. Image quality can be reduced, for instance by occlusions, illumination or motions of the animals, which can influence the reliability of those applications. To address this issue, this study proposes an approach for the identification of uncertain neural-network predictions to improve the overall prediction quality. It proposes the direct quantification of aleatoric and epistemic uncertainty on the one hand and indirect estimation of uncertainty through the prediction of occlusions on the other hand. Our approach simultaneously integrates the different methods into an end-to-end trainable instance segmentation and regression model. The objective of this study was to first investigate how well the different measures can quantify the uncertainty of a prediction by comparing them to human uncertainty assessments. Then, it was analyzed whether the uncertainty estimations are capable to identify and reject erroneous predictions by evaluating the correlation between the predictive error and the uncertainty estimations. Finally, individual predictions were rejected based on the estimated uncertainties to analyze the effect on the overall accuracy. As a use-case, the developed methods were applied to the prediction of plumage conditions of chickens but also examined in a separate domain. The results showed that the outputs of our approaches for the estimation of aleatoric and epistemic uncertainty correlate to the predictive error of the model, and lead to increased performance when uncertain predictions are rejected. In contrast, the indirect method to identify occluded samples did not serve as a reliable indicator for uncertainty and could therefore not be used to improve the accuracy of the model outputs.

3.1 Introduction

In poultry farming, systematic monitoring of the animal's health and welfare is a crucial task. As manual checks are labor-intensive, costly and might stress the animals, more and more applications based on computer vision and deep learning have found their way into the farms to automate this work (Okinda et al., 2020). However, using vision-based approaches in uncontrolled environments, such as poultry farms, is challenging. Varying illumination, occlusions, motion of the animals, and other environmental factors can influence the image quality, which in turn influences the predictions made by the systems. Most deep learning models include a prediction of confidence, which can be taken into account when making the predictions, however, this tends to suffer from over or under confidence, if poorly calibrated (Gawlikowski et al., 2023).

This paper focuses on the identification of unreliable neural network predictions to improve the overall prediction quality. Knowing whether a model's prediction can be trusted or not would allow the indication of unreliable predictions for further (manual) verification or to reject them automatically from the final model output. Such a reject option has become particularly relevant in safety-critical applications, such as in medical imaging (Tian et al., 2024; Ge and Wang, 2021). Identification of unreliable predictions can be addressed in multiple ways; either indirectly by defining and measuring indicators for (un-)reliability, or by directly estimating the uncertainty of a prediction.

3.1.1 Indirect uncertainty estimation through feature recognition

One approach to distinguish reliable from unreliable predictions in neural networks is the addition of a network output that indirectly indicates the uncertainty of a prediction and is trained on explicit ground-truth labels. These can be outputs quantifying the general quality of an image (Bosse et al., 2016), obtained from human perception or derived from a comparison with reference images. Alternatively, specific predefined features that serve as an indicator for reliability can be extracted from each image. Those features can be related, for instance, to the pose, occlusion, or motion of the object of interest, but also to the environmental conditions of the recorded scene (Hernandez-Ortega et al., 2019). For example, an overexposed object might indicate a high level of uncertainty for predictions made on this object. In this case, networks are developed to measure these features and, depending on the individual feature characteristics, an estimation of the trustworthiness of a prediction can be made.

3.1.2 Direct quantification of uncertainty

An alternative to the methods based on ground-truth labels is to directly quantify the uncertainty of a prediction. There are two types of uncertainty that are commonly distinguished; aleatoric and epistemic uncertainty (Kiureghian and Ditlevsen, 2009).

Epistemic uncertainty
Epistemic uncertainty captures uncertainty which is caused by a lack of knowledge (Hüllermeier and Waegeman, 2021). In machine learning, it is often referred to as model uncertainty, as this type of uncertainty is dependent on the model and the quality of training data. Thus, epistemic uncertainty can be reduced by extending the training data, improving the machine-learning model, or better data analysis. Epistemic uncertainty is especially important to identify limits of the model's capabilities. For example, if a model receives an input which is very different from the training data, epistemic uncertainty of the predictions would be high.

Aleatoric uncertainty
On the other hand, aleatoric uncertainty refers to the uncertainty that is presumed to be intrinsic randomness of an observation (Kiureghian and Ditlevsen, 2009), that is noise and uncertainty in the input data. This includes, for example, sensor noise or ambiguities in the data itself. As this type of uncertainty is a property of the data and not a property of the model, it cannot be reduced by the model, even if the amount of training data is increased (Kraus and Dietmayer, 2019). Aleatoric uncertainty can further be categorized in homoscedastic uncertainty, which is constant for all inputs, for example caused by a biased sensor, and heteroscedastic uncertainty, where the uncertainty is different for different inputs. The latter is especially relevant for computer vision applications, as it is specifically dependent on the content of the input image (Kendall and Gal, 2017). Even if the same sensor is used, there are inputs causing more uncertain outputs than others, for example, due to a lack of visual features, motion blur of objects, or (partial) occlusion of the object of interest.

Uncertainty estimation in deep learning for computer vision
Estimation and distinction of different types of uncertainty have been subject of multiple studies for various use cases in deep learning-based computer vision. For example, epistemic uncertainty of segmentation masks was used to analyze brain scans (Roy et al., 2019; Jungo et al., 2018) and diabetic retinopathy (Filos et al., 2019). Other applications focused methods to capture aleatoric uncertainty in human pose estimation (Gundavarapu et al., 2019) and per-pixel uncertainty in depth estimation (Liu et al., 2019).

In addition to the estimation of either aleatoric or epistemic uncertainty, some approaches specifically addressed the decomposition of the total uncertainty into epistemic and aleatoric components. By measuring the total uncertainty and one of its components, the other component could be obtained as the difference of the two known values (Depeweg et al., 2018; Prado et al., 2019). For example, this approach was followed to capture uncertainty in regression and reinforcement learning tasks (Depeweg et al., 2018).

Moreover, recent studies also integrated uncertainty estimation in deep learning models to reject uncertain predictions. However, this was mostly restricted to classification tasks in which predictions were rejected considering their uncertainty about the predicted class

(Tian et al., 2024; Cicalese et al., 2021; Pires et al., 2020).

3.1.3 The proposed work

Our approach addresses the above named challenge of identifying unreliable neural network predictions. It was developed with a primary focus on the use case of plumage-condition assessment in laying hens, a vision-based task susceptible to the impacts of uncontrolled farm environments. Regular assessment of plumage condition is an essential task in laying-hen farming and provides valuable information about the overall welfare of the birds. As the manual assessment is labor-intensive and not standardized, a method to automate this task by using computer vision and a neural network called "ChickenNet" was proposed in Lamping et al. (2022). The proposed model in that work segmented detected chickens and predicted a regression-based plumage-condition score (see Section 3.2.1). However, predictions were given also in situations where the quality of the images was not sufficient to make the assessment, resulting in reduced performance. In the current work, we aimed to deal with this by estimating uncertainties for each plumage-condition assessment as an integral part of the deep neural network. Our approach used a model for plumage condition assessment proposed in Lamping et al. (2022), however, it was intended be applicable on other tasks for object detection or instance segmentation. Overall, the main contributions of this research are as follows:

- Our proposed approach presents a flexible method that integrates three different uncertainty measures into an end-to-end trainable instance-segmentation and regression model to identify unreliable model predictions.
- This research addresses the challenge of unknown reliability in neural network predictions in a dual manner: It involves and compares both the direct estimation of aleatoric and epistemic uncertainty for predictions, and the indirect estimation of uncertainty through the prediction of the level of occlusion for detected instances.
- We evaluate how well our model's estimation of occlusions, epistemic and aleatoric uncertainty indicate the reliability of a prediction. This evaluation encompasses both qualitative and quantitative analyses, with comparisons drawn against human assessments.
- Leveraging our method's estimations, we implement a strategy to reject uncertain predictions, in order to enhance overall prediction accuracy.
- Effectiveness of our method is demonstrated not only within the prioritized use case of plumage condition assessment but also in an unrelated domain.

3.2 Material and methods

In the following, the material and methods will be presented. Section 3.2.1 explains the architecture of ChickenNet, the neural network for plumage-condition assessment that we used as an example to implement uncertainty predictions. In Section 3.2.2 the image data will be presented, while Section 3.2.3 introduces the three different approaches to estimate uncertainty. Finally, Section 3.2.4 describes the training parameters and Section 3.2.5 elaborates the experimental setups and the corresponding evaluation metrics.

3.2.1 Introduction to ChickenNet

he developed methods were designed to extend ChickenNet, an end-to-end trainable convolutional neural network that detects and segments individual chicken while predicting a plumage-condition score for each detected hen, which was presented in Lamping et al. (2022). ChickenNet extends Mask R-CNN (He et al., 2017) by predicting the condition score in addition to the mask, bounding box and class of the detected object.

It incorporates a region-proposal network after a ResNet50 backbone to generate region of interest (RoI) proposals. Then, these RoIs are realigned by a RoI alignment layer and transformed into fixed-sized feature maps, which are further processed in two branches. The first branch performs mask segmentation while the second one is designed for classification, bounding box regression and plumage-condition assessment. The latter branch contains two fully connected (FC) layers followed by the output layer for classification as well as the regression layers for box and plumage-score prediction. The plumage-condition scoring was formulated as a regression problem, with scores ranging from 0 to 2, where a score of 0 indicates a perfect condition and a score of 2 indicates heavy plumage damage. Contrary to the original implementation of ChickenNet, in this work, three-channel RGB-images were used without the additional depth channel that was implemented in the original work, as there were no benefits found of including depth information for the assessment (Lamping et al., 2022). For the details of the ChickenNet architecture, we refer to Lamping et al. (2022).

3.2.2 Image data

For this study, the image datasets of the original ChickenNet paper were used for training, validation, and testing. All images were obtained from video sequences recorded in a commercial laying hen farm following the procedure described in Lamping et al. (2022). Using a Stereolabs ZED 2 camera, which was placed in front of the aviaries' feeding line, hens were recorded from a distance of 40 to 120 cm, depending on the animal's position in the aviary. While the images were identical to the original dataset, image annotations were updated for the purpose of the present study. These annotations were made by a human expert and included labels for the class ("hen"), segmentation masks, bounding

boxes, plumage-condition scores, and occlusion levels for each individual hen. In total, 2621 hen instances were included in the training set, with 371 instances in the validation set and 387 in the test set.

Plumage conditions of the hens were annotated in two steps. First, each video frame was annotated individually based on information visible in that specific frame of the video sequence. A score of 0 was provided when a plumage without any damages was observed, 1 when minor damages (featherless area diameter < 5 cm) were visible and 2 in case of heavy damages (featherless area diameter ≥ 5 cm). If a hen was not clearly assessable by the annotator due to occlusions or the bird's posture, it was marked as unclear and a score of -1 was given. However, to predict uncertainty and to evaluate the effect on the overall assessment performance, including the ground-truth scores of the unclear cases, was required. Therefore, in a second step, these unclear cases were annotated again by using information from the whole video sequences. Thus, all hens obtained a ground-truth score of either 0, 1, or 2 while scores that were unclear for a human annotator were still indicated as such.

As our approach extends ChickenNet by an output to predict the level of occlusion for each detected hen, additional ground-truth labels for training of the neural network are required. These labels were provided through manual assessment of the image data. For annotation, occlusions caused by other animals as well as by the aviary system itself were considered. Also, all birds in an image were annotated, meaning that occlusion information was also assigned to birds that were fully visible. In this case, an occlusion level of zero was indicated. To simplify the annotation, occlusions were categorized using five different ground truth labels: 0%, 1–25%, 26–50%, 51–75% and 76–100%. One of the five classes was assigned to each visible hen based on the human estimation of the occlusion. Table 3.1 illustrates the resulting distribution of the occlusion annotations among the training, validation and test set.

Table 3.1: Distribution of occlusion annotations among the datasets.

Dataset	0%	1-25%	26-50%	51-75%	76-100%
Training	1196	664	353	277	131
Validation	144	126	54	40	7
Test	233	96	22	19	17

3.2.3 Three methods to predict uncertainty

We developed three different methods that extend ChickenNet to estimate uncertainty of the predicted plumage-condition score and integrated them into one model: (1) prediction of the level of occlusion, (2) prediction of the epistemic uncertainty, and (3) prediction

of the aleatoric uncertainty. These three methods are described in detail in the following sections.

3.2.3.1 End-to-end learned prediction of occlusion levels

The first approach considers the level of occlusion of each individual hen as an indicator for unreliable plumage-condition assessments. The hypothesis is that the more a bird is occluded, the more erroneous the prediction of the plumage condition is. Thus, this is an indirect uncertainty measurement based on pre-defined image characteristics, and independent of the prediction of the plumage-condition score. We designed the prediction of occlusions as an end-to-end trainable solution incorporated in the ChickenNet architecture, as described below.

In addition to the existing outputs of ChickenNet for detection, segmentation, and plumage-condition assessment of an individual laying hen, the existing model architecture was supplemented with an output to predict the level of occlusion as shown in Figure 3.1. Even though the occlusion annotations were made using five distinct categories, we formulated the occlusion estimation as a regression problem. This was done to obtain a continuous value for the occlusion level and to consider the order in the levels of occlusion.

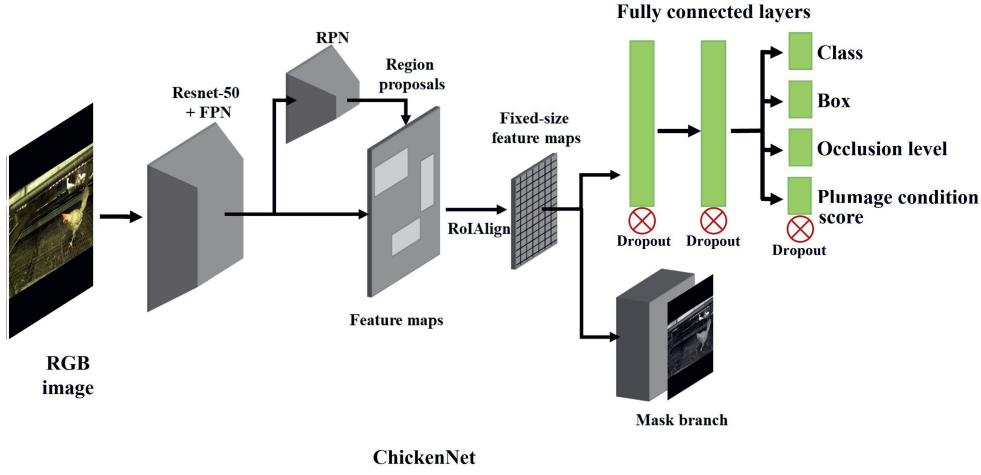


Figure 3.1: Network architecture. ChickenNet is extended by dropout layers to predict epistemic uncertainty and by additional output layers for prediction of occlusions and aleatoric uncertainty.

To implement an additional output for occlusion prediction, we extended the second branch of the network with an additional FC output layer, in parallel to the three existing ones. Thus, the output layers for classification, bounding boxes, plumage scores, and

occlusion share the same input features. To obtain values between 0% and 100% for the predicted occlusion level, y^{occ}, of instance i, this layer consists of a single neuron with a sigmoid activation function. During training, the loss of each occlusion prediction was calculated using the smooth L1-loss.

3.2.3.2 Monte-Carlo dropout for epistemic uncertainty estimation

To estimate epistemic uncertainty of the plumage-condition predictions, we applied the Monte-Carlo dropout method, which was introduced by Gal and Ghahramani (2016) and has been frequently implemented to estimate epistemic uncertainty in multiple computer vision applications (Kraus and Dietmayer, 2019; Lamping et al., 2022; Wang et al., 2019; Le et al., 2018), including instance-segmentation tasks using Mask R-CNN (Blok et al., 2022). Traditionally, dropout has been used as a regularization technique for training of neural networks. During the training process, some units are randomly disconnected from the network, which promotes multiple pathways for prediction by the network and has been shown to prevent overfitting on the training data. In contrast to standard dropout, Monte-Carlo dropout is applied during inference in multiple forward passes. For each forward pass, the dropout pattern varies, which results in different predictions for each pass. The resulting predictions replicate a Query-by-Committee of different activated network weights, and thus can approximate a probability distribution with the variance related to the uncertainty of the prediction of the mean of these samples.

To estimate epistemic uncertainty within ChickenNet, we aimed to generate multiple score prediction samples to compute the predictive variance. Monte-Carlo dropout was applied to all fully connected layers after the generation of the fixed-size feature maps (see Figure 3.1). This placement was previously shown to be successful in similar network architectures (Kendall and Cipolla, 2016). It is important to note that only the predictive variance of the plumage score is used to calculate uncertainty, while the variance in the outputs of bounding-box regression, occlusion prediction, mask segmentation, and classification were not included in the uncertainty estimation in our work.

Changes to the ChickenNet architecture were made in consideration of the computation time. Therefore, each input was passed through the network only once until it reaches the dropout layer(s). After these layers, T forward passes are performed with a dropout probability of p for each neuron to be disconnected from the network. This resulted in T samples for the plumage condition output. In our approach, we used a dropout rate of $p = 0.2$, which has been a common choice in literature and was shown to perform well for the estimation of epistemic uncertainty (Avci et al., 2021; Verdoja and Kyrki, 2021; Zhang et al., 2022). The number of forward passes was set to $T = 30$, based on results in Gal and Ghahramani (2016); Laakom et al. (2021). By computing the mean of the generated samples, $\overline{y}_i^{score}$, the final plumage-condition prediction was made. The epistemic uncertainty $\hat{y}_i^{epistemic}$ was estimated using the obtained variance of the samples:

$$\hat{y}_i^{epistemic} = \frac{1}{T-1}\sum_{t=0}^{T}(\hat{y}_{t,i}^{score} - \bar{y}_i^{score})^2 \tag{3.1}$$

3.2.3.3 Loss attenuation for aleatoric uncertainty estimation

In addition to the estimation of epistemic uncertainty, we focused the prediction of heteroscedastic aleatoric uncertainty, which varies dependent on the input data. For simplicity, heteroscedastic aleatoric uncertainty will be referred to as aleatoric uncertainty in the remainder of the paper. Our approach followed the method presented by in Kendall and Gal (2017), which combines the estimation of both epistemic uncertainty and aleatoric uncertainty in one model. For regression tasks, such as the plumage-score prediction in ChickenNet, the authors showed that aleatoric uncertainty can be interpreted as so-called learned loss attenuation. This approaches focuses the loss function of the neural network, which measures the difference between predicted and actual regression output, allowing the network to learn during training. Instead of only predicting a single regression output $\hat{y}_i^{score}$, the presented model simultaneously predicts a measure of aleatoric uncertainty, given by the variance σ_i^2. As both $\hat{y}_i^{score}$ and σ_i^2 are dependent on the data, both outputs can be learned directly from the inputs. For our implementation, this results in the following loss function:

$$L_{score} = \frac{1}{N}\sum_{i=0}^{N}\frac{(y_i^{score} - \hat{y}_i^{score})^2}{2\sigma_i^2} + \frac{1}{2}\sigma_i^2 \tag{3.2}$$

Contrary to the formulation in Kendall and Gal (2017), this loss function removes the logarithmic function from the second term, which had shown to improve training stability of a neural network in previous studies (Le et al., 2018) and rectified the unstable training behavior we experienced while using the original formulation. Note that learning to predict σ_i^2 does not require explicit ground-truth labels about the uncertainty. Instead, it is learned implicitly through the loss function. The first term of the function encourages the model to reduce the predictive error, while predicting a high variance also reduces the contribution of this term to the overall loss. However, the second term penalizes large variances for all inputs. This teaches the network to predict a higher variance for predictions that might be erroneous and to predict a low variance for correct answers, providing a measure of aleatoric uncertainty. This formulation of the loss function allows the network to learn to attenuate effects of false predictions or even false labels, making it more robust to noisy data. In ChickenNet, this loss function was used to replace the initially implemented smooth L1 loss of the plumage score prediction.

To combine both aleatoric and epistemic uncertainty estimation in one model, the probabilistic modeling for epistemic uncertainty estimation must be incorporated in the loss function above. In line with the approach presented in Section 3.2.3.2 and following the

approach in Filos et al. (2019), we used Monte-Carlo dropout to generate multiple predictions of a sample. Thus, during test time, both predicted outputs $\hat{y}_i^{score}$ and the variance σ_i^2 were calculated as mean values of T forward passes. Accordingly, to solely predict the aleatoric uncertainty and no epistemic uncertainty, only one forward pass without dropout would be performed during test time instead of T passes.

To implement the aleatoric loss function into the plumage-score prediction of ChickenNet, we readjusted the weighting of the individual components of the overall model loss. The equal weighing of the score loss and the losses corresponding to the classification, bounding box, occlusion and mask segmentation outputs in the original ChickenNet resulted in a decreased detection and classification performance. Therefore, we reduced the weight of the score loss by a factor of ten which maintained the original detection and classification performance, while allowing to predict the uncertainties. Thus, the total loss of ChickenNet was defined as:

$$L = L_{class} + L_{box} + L_{mask} + L_{occlusion} + 0.1L_{score} \tag{3.3}$$

3.2.4 Training parameters

During training, all layers of the network except for the final outputs for classification, bounding boxes, plumage scores, and occlusions (see Figure 3.1) were initialized with weights obtained from pre-training on the Microsoft Common Objects in Context (COCO) dataset (Lin et al., 2014). For weight optimization, we used stochastic gradient descent (SGD) with a learning rate of 0.0001 and momentum of 0.9.

During training, image augmentations of the original ChickenNet were applied which included variations in scale, brightness and horizontal reflection as described in Lamping et al. (2022). To avoid consideration of any information in the background of an image for plumage-condition assessment, we additionally augmented the image backgrounds. While images including the original background were still used for training, different backgrounds were added to create artificial images, extending the training data. Here, we used recordings that were obtained during data collection but did not contain any chicken. Based on the manually annotated instance segmentations, hens were cropped from the original image and the background was randomly replaced. An example for influencing background information might be the condition of the farm environment. As a farm house is disinfected before a flock moves in, the aviary system tends to be cleaner on images of young birds. Over time, the cleanliness of the housing system as well as the plumage conditions decrease. Therefore, a clean image background might be identified as an indicator for good plumage condition. To exclude this potential risk, images with random backgrounds were included in the training data.

The training was performed using standard dropout for regularization, in line with the

approach in Kendall and Gal (2017). All other training parameters were set analogously to the training procedure of ChickenNet and can be obtained from Lamping et al. (2022).

3.2.5 Experiments

Two sets of experiments were set up to evaluate the performance of the three developed methods for estimating the different types of uncertainty. First, the general effectiveness of the proposed approaches was evaluated in three different experiments. Subsequently, it was investigated whether uncertainty estimation is beneficial for the identification of false predictions in order to improve the overall assessment performance of the ChickenNet model.

3.2.5.1 Individual analysis of uncertainty estimation methods

In the first set of experiments, we tested how well our model estimates epistemic uncertainty, aleatoric uncertainty, and occlusions.

Experiment 1: Occlusion prediction
Regarding occlusion estimation, our image data was annotated to train the model in a supervised manner, which allowed a direct comparison of predictions and the corresponding ground-truth occlusion levels. For evaluation, an occlusion prediction was considered correct if it was within the borders defined by the ground truth labels which were 0%, 1–25%, 26–50%, 51–75% and 76–100%. Following this definition, the accuracy of the occlusion prediction was defined as the proportion of correct predictions among the total number of predictions.

Experiment 2: Uncertainty estimation vs. human assessment
The epistemic and aleatoric uncertainty predictions needed to be evaluated indirectly, since no ground-truth uncertainty is known. We did, however, have a categorical assessment by the human annotator of the possibility to assess the chicken in the image with the 'unknown' labels, which could be interpreted as an indicator of uncertainty (Lamping et al., 2022). The 'unknown' labeled chickens were used to test the consistency of the human uncertainty with the predicted occlusion, epistemic and aleatoric uncertainty by the network. An unpaired two-sample t-test was carried out with $\alpha = 0.05$ to compare between uncertainty predictions made on samples labeled as "unknown" and those labeled as "known" by the human assessor to see if the predicted uncertainties significantly differ for the chickens labeled as 'known' and 'unknown'.

Experiment 3: Effects of artificial image modifications
The third experiment provides a qualitative analysis of the epistemic and aleatoric uncertainty predictions. We aimed to evaluate the changes in the uncertainty predictions through artificially modified images, manipulating the aleatoric uncertainty. We chose

a batch of ten images containing different chickens with intact, slightly damaged, and heavily damaged plumages and added varying occlusions and blur to them. Occlusions were added either to the head, center or the back of the chicken's body. To occlude the specific body parts of a chicken, black boxes were manually added to the particular regions of the images. For all samples of birds with clearly visible naked spots (score = 2) in the plumage, an occlusion of the chicken's back also means an occlusion of those naked spots, as illustrated in Figure 3.2. By occluding the parts of a chicken's body that are relevant for the assessment of the plumage condition, we expected an increase in aleatoric uncertainty of the prediction while occlusion of less relevant regions, such as the head, was expected to have a smaller effect on the prediction's uncertainty. Compromising the image quality by applying 10-pixel motion blur was also supposed to increase aleatoric uncertainty as it decreases the quality of the data, which might impact the certainty of a prediction. Regarding epistemic uncertainty, we expected no systematic changes caused by the manual image modifications. Figure 3.2 demonstrates the modifications that were applied to the images in this experiment for two exemplary images.

3.2.5.2 Investigating the use of uncertainty estimations to improve assessment performances

In this set of experiments, we tested the suitability of the different uncertainty predictions to automatically identify false assessments made by the ChickenNet model. Here, the objective was to determine whether high predicted uncertainties relate to false predictions and could be used to reject them from the final prediction output in order to improve the overall plumage-condition assessment. Therefore, in experiment 4, we compared the predicted numeric values for occlusions, epistemic uncertainty, and aleatoric uncertainty to the assessment errors made by the model. We hypothesized that uncertainties would be much higher for false predictions with a high error. Then, in experiment 5, the assessment accuracy of the plumage-condition assessment was evaluated while rejecting samples based on their predicted uncertainties. This was repeated in experiments 6 but for a different tasks and a different dataset to evaluate the effectiveness of our method in other domains. For both experiments, it was hypothesized that the rejection of uncertain predictions would lead to an improved assessment.

Experiment 4: Uncertainty predictions vs. predictive error
First, we compared the predicted numeric values for occlusions, epistemic uncertainty, and aleatoric uncertainty to the assessment errors made by the model. For all samples in the test set, the plumage-condition score, $\hat{y}_i^{score}$, was predicted and deviations from the ground-truth scores, y_i^{score}, were calculated using the absolute error:

$$e_i = |y_i^{score} - \hat{y}_i^{score}| \tag{3.4}$$

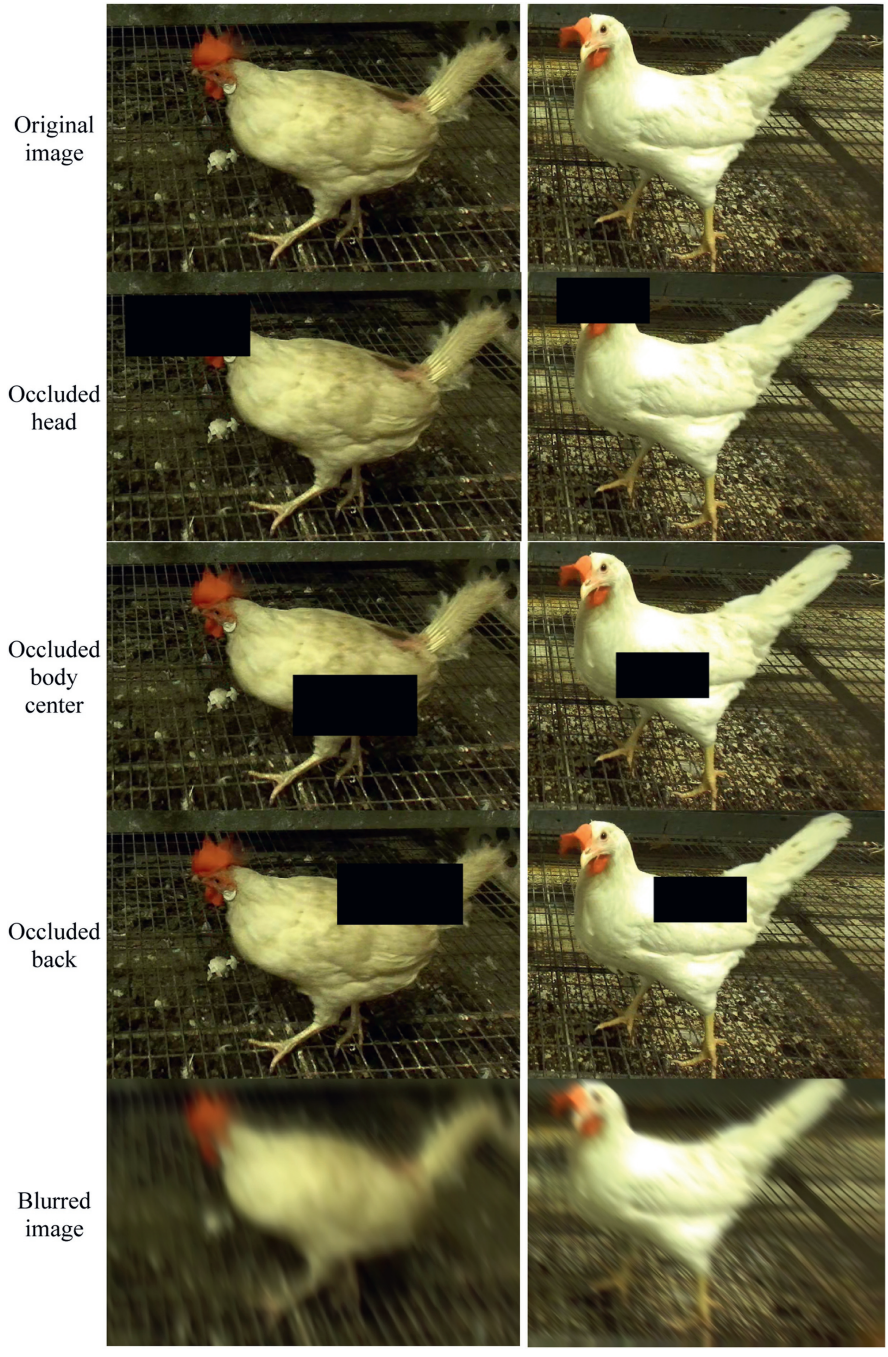


Figure 3.2: Artificial image modifications illustrated for a hen with clearly visible plumage damage (left) and for a hen without plumage damage (right).

Then, the correlation between the absolute assessment error and the predicted uncertainties was calculated using Pearson's correlation coefficient.

Experiment 5: Uncertainty-based rejection compared to the base model
In this experiment, we used the predicted uncertainties of each sample to reject uncertain score predictions and evaluate the overall assessment accuracy of ChickenNet. This allows a direct comparison between the method proposed here and the standard ChickenNet without uncertainty estimation. We quantized the predicted plumage condition, $\hat{y}_i^{score\_cls}$, to allow the calculation of the accuracy by comparing to the ground-truth condition classes:

$$\hat{y}_i^{score\_cls} = \begin{cases} 0, & \text{if } \hat{y}_i^{score} < 0.5 \\ 1, & \text{if } 0.5 \leq \hat{y}_i^{score} \leq 1.5 \\ 2, & \text{if } \hat{y}_i^{score} \geq 1.5 \end{cases} \tag{3.5}$$

For each of the three predicted uncertainty types, all N samples from the test set were sorted by the respective numerical uncertainty value in descending order. Thus, we obtained three differently sorted lists, each containing N test samples, with the most uncertain starting sample. Based on these lists, we evaluated the prediction accuracy on the test set. Then, we removed the first most uncertain sample from each list and evaluated the assessment performance again, based on the remaining $N - 1$ test samples. This process was iterated until only one test sample remained. Furthermore, the results from the uncertainty-based rejection were compared to the performance that was obtained if samples were manually rejected by a human observer. This manual selection corresponds to removal of all 'unknown' labelled instances. For evaluation, we used rejection-accuracy curves, where the assessment accuracy is calculated as a function of the percentage of the rejection. We hypothesized that for all three types of uncertainty, the overall assessment performance would increase with an increasing proportion of rejections.

Experiment 6: Cross-Domain Applicability
The present approach focuses the estimation of uncertainty for the identification of unreliable neural-network predictions to improve in plumage condition assessment in laying hens. However, to demonstrate the applicability of our method as a general framework for incorporation uncertainty into assessment tasks, we finally evaluated our approach on data beyond the domain of plumage condition assessment. Therefore we tested the rejection based on aleatoric and epistemic uncertainty using our approach on the MARS-Attribute dataset (Chen et al., 2019). This dataset consists of pedestrian recordings from different cameras which were additionally labeled with 32 person-related attributes such as gender, clothing-color or age. Instead of plumage condition labels, we trained ChickenNet on the age attributes which ranged from 0 to 3, indicating children, teenager, adults and elderly

people. Analogues to the plumage condition estimation, we modeled the age estimation as a regression task so that it is suitable for the architecture of our model. Thus, for each age prediction, estimations of epistemic and aleatoric uncertainty could be obtained from our architecture. As the MARS-Attributes dataset does not provide annotation of the level of occlusion for each person, this experiment solely focused the estimation of aleatoric and epistemic uncertainty for each sample.

After training the model, predictions were made on the test set followed by a step-by-step rejection according to the procedure described in Experiment 5. After each rejection, the assessment accuracy was evaluated and compared to the accuracy of the standard model without uncertainty-based rejection.

3.3 Results

The results are summarized following the order of the experiments. First, the individual performance evaluation for each of the three uncertainty estimation approaches is presented. Then, the use of the methods for an improved assessment performance of ChickenNet is focused.

3.3.1 Individual analysis of uncertainty estimation methods

Experiment 1: Occlusion prediction
In our first experiment, we evaluated the occlusion prediction of our model compared to the ground-truth annotated occlusion level. The corresponding results are given in the normalized confusion matrix in Figure 3.3, showing a high accuracy for the lower levels of occlusion, 68% for small occlusions and 88% for birds without any occlusion. Most wrong predictions occurred for the two highest occlusion levels, corresponding to hens with a ground-truth occlusion above 50%. In 58% of cases, occlusions between 75% and 100% were incorrectly classified as occlusions between 50% and 75%. Vice-versa, misclassification occurred in 38% of cases. Furthermore, the results show that misclassification by more than one class was very rare and only occurred in 1% of cases for ground truth occlusion between 1% and 25%. Over all occlusion level classes, we obtained a prediction accuracy of 78.3% and an RMSE of 9.0%, based on the means of the class ranges.

Experiment 2: Uncertainty estimation vs. human assessment
We compared the predictions of occlusions, aleatoric uncertainty, and epistemic uncertainty to uncertainty labels given by a human annotator. Table 3.2 shows the mean and standard deviation for the three uncertainty predictions for samples that were marked as "unknown" and for the samples that were marked as "known" by the human annotator. The p-values, resulting from an unpaired two-sample Student's t-test, show no significant difference for known and unknown for epistemic uncertainty ($p = 0.16$), a small significant

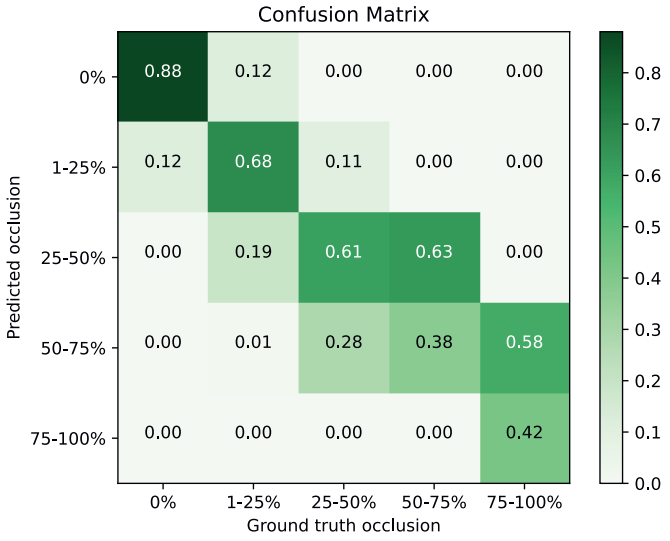


Figure 3.3: Confusion matrix of predicted occlusion levels against the ground truth occlusion.

difference for aleatoric uncertainty ($p = 0.04$) with higher uncertainty for the unknown class, and a strong significant difference for occlusion prediction, with higher predicted occlusions for the unknown class.

Table 3.2: The mean and standard deviation for predicted occlusions, aleatoric uncertainties and epistemic uncertainties on samples marked as known and unknown by a human annotator. The p-value results from an unpaired two-sample Student's t-test.

	Epistemic uncertainty		**Aleatoric uncertainty**		**Occlusion prediction**	
	Known	**Unknown**	**Known**	**Unknown**	**Known**	**Unknown**
μ	0.12	0.11	0.22	0.26	0.05	0.25
σ	0.08	0.07	0.16	0.20	0.09	0.26
p	0.16		0.04		< 0.01	

Human uncertainty assessments were made based on the available image data and therefore affected by factors such as the quality of the image, occlusions, and the pose of a detected chicken. As aleatoric uncertainty corresponds to uncertainty in the data itself, we would expect a significant difference between the predicted aleatoric uncertainties made on samples that were indicated as known and those that were labeled as unknown. This significant difference was shown in our experiment.

For epistemic uncertainty, mean values of 0.12 and 0.11 were obtained for the known and unknown samples, respectively. Since epistemic uncertainty is related to model-specific uncertainty, we cannot anticipate a correlation with assessments given by a human. Hence, we would expect both groups, i.e., known and unknown samples, to exhibit comparable distributions, which is consistent with the obtained results.

A clear difference between both groups was observed for the predicted occlusion levels. For most known samples, low occlusion levels were predicted, with a mean of 5% and a standard deviation of 9% among all predictions. In contrast, the predicted occlusions vary much more for unknown samples, for which a mean occlusion of 25% and a standard deviation of 26% were obtained. The significant difference between both distributions was clearly confirmed by the performed t-test. These results indicate that the human uncertainty estimation is heavily affected by occlusions. In our experiment, a sample was never marked as known if high occlusion was predicted for this sample. However, occlusions were one, but not the only reason for an "unknown" label, which is confirmed by the observed high variation in occlusion predictions among the unknown samples.

Experiment 3: Effect of artificial image modifications
Table 3.3 presents the effects of occlusions and noise on the prediction of aleatoric and epistemic uncertainty for ten chicken detections from the test dataset. It shows the absolute change in the uncertainty predictions for each image modification compared to the original images. Figure 3.4 visualizes the means of both predicted uncertainties for each image modifications among the test samples.

Table 3.3: Absolute changes in predicted epistemic and aleatoric uncertainty in ten samples for different image modifications compared to original images.

Sample	GT	Occluded head		Occluded center		Occluded back		Blurred	
	Score	Epistemic	Aleatoric	Epistemic	Aleatoric	Epistemic	Aleatoric	Epistemic	Aleatoric
Sample 1	1	-0.07	+0.06	-0.14	+0.18	-0.15	+0.18	-0.12	+0.08
Sample 2	0	0.00	+0.01	+0.13	+0.46	+0.14	+0.38	+0.11	+0.59
Sample 3	2	+0.03	+0.04	-0.01	+0.03	0.00	+0.17	-0.02	+0.06
Sample 4	0	-0.03	0.00	+0.02	+0.22	+0.02	+0.22	-0.05	+0.03
Sample 5	0	0.00	0.00	+0.13	+0.23	+0.01	+0.03	+0.02	+0.02
Sample 6	0	0.00	0.00	0.00	+0.01	0.00	0.00	0.00	+0.01
Sample 7	2	+0.01	+0.01	+0.01	+0.02	-0.05	+0.15	+0.06	0.00
Sample 8	2	+0.09	-0.03	+0.06	0.00	+0.02	+0.08	+0.03	+0.05
Sample 9	2	+0.01	+0.04	-0.01	+0.06	+0.01	+0.04	+0.02	+0.12
Sample 10	1	-0.02	+0.01	+0.02	0.00	-0.10	+0.08	-0.02	+0.09

First, it was observed that the effect of occlusions on the predicted aleatoric uncertainties depended on which part of a hen's body was occluded. Occlusion of the head region caused smaller changes in aleatoric uncertainty than the occlusions of body center or back. A median increase of 0.01 compared to the original image was observed among the test samples if the head of a bird was occluded. Here, aleatoric uncertainty increased in six out of ten cases and decreased in one case. Occlusion of the body's center and

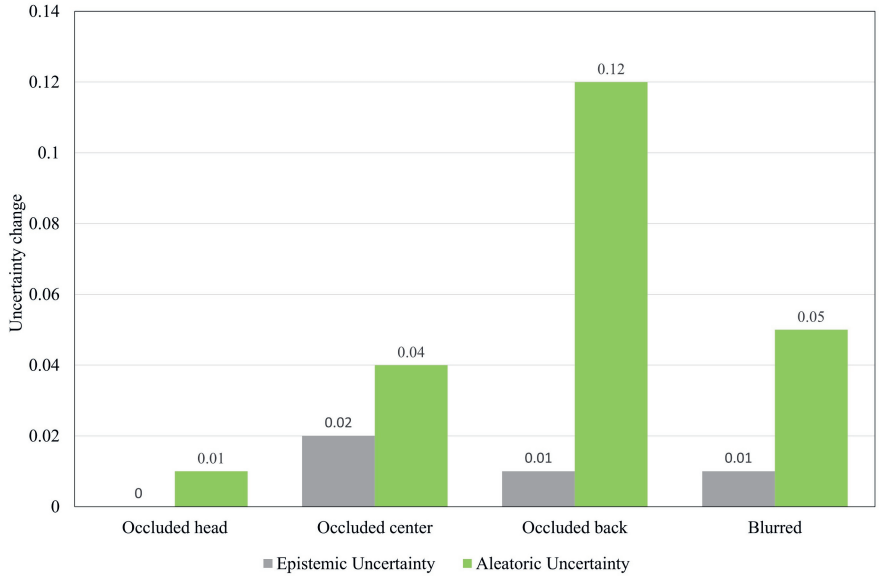


Figure 3.4: Means of absolute changes in predicted epistemic and aleatoric uncertainties for each image modification compared to original images.

back resulted in equal or higher aleatoric uncertainty compared to the original image for all test samples, with median increases of 0.04 and 0.12 respectively. Also, a difference between birds of different plumage conditions was observed. For hens with clearly visible plumage damages (score = 2), occlusion of the body center only had small effects on the aleatoric uncertainty. In contrast, when the back of the bird and therefore the visible naked spot in the plumage was occluded, a significantly higher aleatoric uncertainty was predicted for the plumage-condition assessment. If no naked spots were visible on the plumage (score = 0 or 1), occlusion of body and center both clearly increased the aleatoric uncertainty of the plumage score prediction. Adding artificial blur to the images also increased aleatoric uncertainty for most test samples. Here, a median increase of 0.05 was observed. The intensity of the change varied between the samples without a clearly recognizable dependency on the plumage condition.

The observed dependency of uncertainty predictions on the plumage condition could be explained by the varying relevance of each body part for the plumage assessment. Plumage damage can be identified as soon as naked spots are visible, without a need to consider the full body of a chicken. This is different for chickens without visible naked spots, as the entire plumage needs to be considered to give a good estimation of the condition. Therefore, occlusions or modifications applied to the image are less relevant for the assessment of a damaged plumage, as long as the damages are visible. A highly certain assessment can be made, even though parts of the body might be invisible. On the other hand, aleatoric

uncertainty of intact plumage predictions is much more affected by changes. Each body region can potentially contain indicators for plumage damages, thus the aleatoric uncertainty of the prediction increases if these parts are occluded or blurred. Furthermore, through the three different scores for plumage conditions, ChickenNet differentiates between heavy and light damages. Plumage damages can vary in shape, size and position, which makes images of damaged plumages much more ambiguous than intact plumages. This might explain the generally higher aleatoric uncertainty which was observed for the damaged plumage sample.

While a clear dependency between aleatoric uncertainty on the image quality and occlusion was observed, changes in epistemic uncertainty were less clear. All types of occlusion and image blur resulted in median increases up to 0.02. However, for all experiments, also decreases in epistemic uncertainty were overserved among the test samples. This is in line with the expected results, as epistemic uncertainty relates to model-dependent uncertainty and should therefore not directly correlate with changes in the image data. Indirectly, manual modifications might affect epistemic uncertainty, as these artificially created images differ from the samples the model was trained on. For example, situations in which the body center of a hen is occluded exclusively without any occlusions of the back or head rarely occur in reality, which leads to an underrepresentation of those samples in the training data.

3.3.2 Investigating the use of uncertainty estimations to improve prediction performance

Experiment 4: Uncertainty predictions vs. predictive error
Figure 3.5 compares the absolute error of all samples from the test set to the model's prediction of occlusion, epistemic, and aleatoric uncertainty, expressed on a logarithmic scale. The epistemic and aleatoric uncertainty predictions show a similar distribution of points. For both, the lowest uncertainties were predicted for the samples with the smallest predictive error. Also, a high concentration of data points could be observed for high-error and high-uncertainty values. This indicates that both aleatoric and epistemic uncertainty are generally higher for plumage score predictions that strongly differ from the ground-truth score, compared to predictions with small errors. No clear relation between the predicted occlusion and the error was observed.

For aleatoric uncertainty, the Pearson's correlation coefficient between the logarithmic error and the logarithmic uncertainty was 0.95, while it was 0.81 for epistemic uncertainty. This indicates a strong positive correlation between both types of uncertainty and the prediction error. The relationship between predicted occlusions and the assessment error was not clear, with a correlation coefficient of -0.12.

Experiment 5: Uncertainty-based rejection compared to the base model
The previous results suggest that the epistemic and aleatoric uncertainties can be used

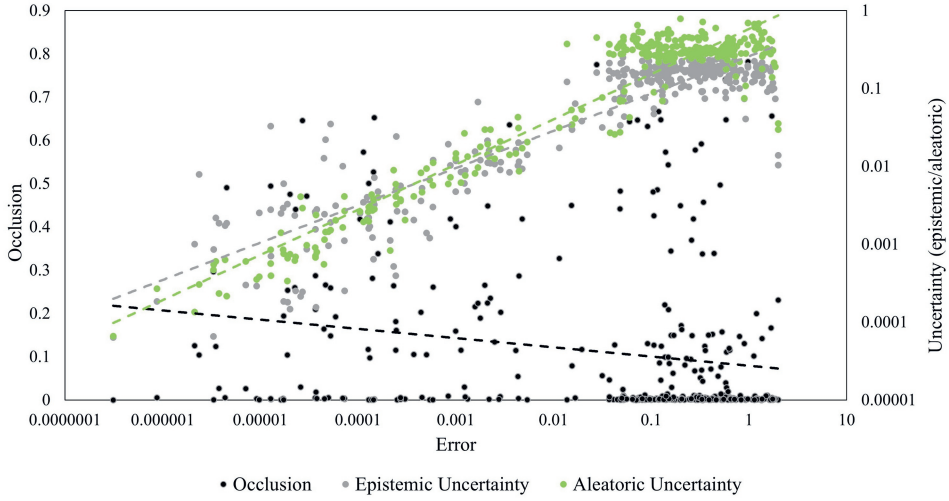


Figure 3.5: Predicted occlusions, aleatoric uncertainties and epistemic uncertainties compared to the absolute error for all samples from the test dataset, expressed on a logarithmic scale. The dashed lines show the linear regression line through the logarithmic data. This relates to a Pearson's correlation coefficient of 0.95 for aleatoric uncertainty, 0.81 for epistemic uncertainty, and -0.12 for occlusion.

to filter out erroneous predictions of the plumage-condition. This hypothesis was tested in experiment 5. Figure 3.6 presents the rejection-accuracy curves for rejection based on aleatoric uncertainty, epistemic uncertainty, and predicted occlusion. The curves were generated by evaluating the assessment accuracy while varying the rejection threshold, where the most uncertain samples were rejected first. As a reference, the assessment performance of standard ChickenNet obtained after manually excluding samples classified as unknown by the human assessor is visualized at 0.85. This exclusion pertains to images rejected by humans, which make up 31.45% of the test set. All three curves have the same starting accuracy of 0.77, since this is the assessment accuracy of ChickenNet based on the complete test set which forms the reference to the uncertainty-based rejection. As expected from the previous results, rejection based on aleatoric and epistemic uncertainty increased the overall accuracy with an increasing rejection rate. Initially, accuracy increased faster for aleatoric uncertainty, reaching the level of human rejection at a rejection rate of 34.67%. For epistemic uncertainty, this accuracy level was obtained the first time at a rejection rate of 49.73%. Thus, both uncertainty based rejections did not surpass the human-based rejection rate of 31.45%. An accuracy of 1.0 on our test dataset was observed at a rejection rate of 73.38% for aleatoric uncertainty and at a rejection rate of 74.46% for epistemic uncertainty. These results show that both aleatoric and epistemic uncertainty can be used to filter out erroneous predictions.

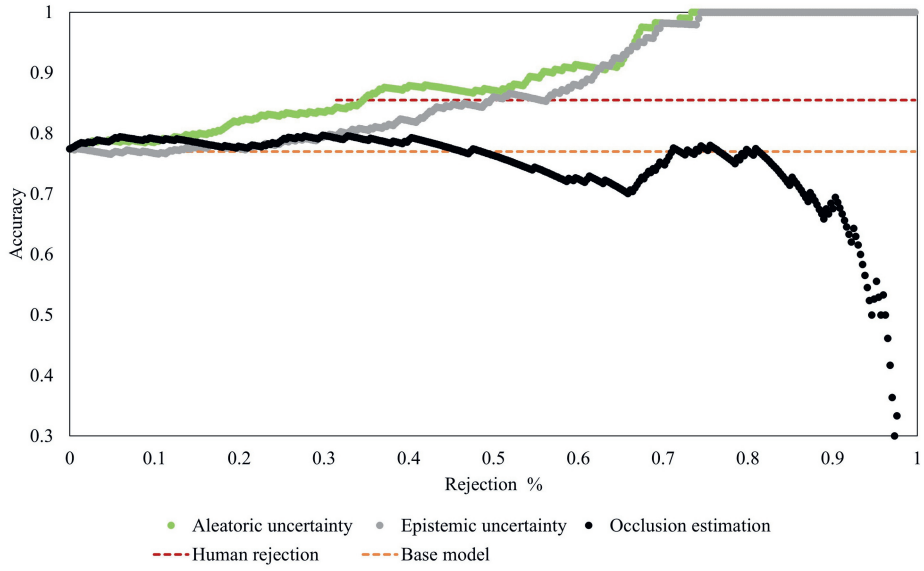


Figure 3.6: Rejection-accuracy curves for rejection based on predicted occlusion (black), aleatoric uncertainty (green) and epistemic uncertainty (gray). The red dashed line shows the accuracy obtained when rejecting all images that were classified as 'unknown' by the human annotator. Note that for the latter, no different rejection rates are indicated as the rejection rate in the annotated data was fixed to 31.45%. The line is shown for clarity. The orange line indicates the accuracy of the standard ChickenNet model without any rejections, neither by a human annotator nor by the model itself.

In contrast, rejection based on the predicted occlusion level did not result in an increase in accuracy compared to the base model. An initial increase in accuracy was observed when heavily occluded samples were rejected. However, with an increasing rejection rate, the accuracy dropped. For lower levels of occlusions, the errors vary much more and thus the rejection of those samples does not improve the overall accuracy. This shows that the predicted occlusion level is not a valuable feature to filter out erroneous predictions.

Experiment 6: Cross-domain applicability

In addition to the uncertainty-based rejection for enhancing plumage condition assessment, we evaluated the rejection of uncertain predictions on the MARS-attributes dataset, concentrating the task of age estimation through ChickenNet. Analogous to the previous experiment, Figure 3.7 illustrates the prediction accuracy with an increasing proportion of samples rejected either based on aleatoric or epistemic uncertainty. Both were compared to the accuracy of the standard ChickenNet which did not include any sample rejections. In this setting, an age-estimation accuracy of 0.76 was obtained. This is lower than the state-of-the-art accuracy showcased in Chen et al. (2019), which may be attributed to the

comprehensive nature of the ChickenNet architecture, which encompasses object detection, segmentation, assessment, and uncertainty estimation, rather than being exclusively designed for age estimation. Furthermore, we evaluated image-based predictions independently while the state-of-the-art models additionally incorporate temporal information. Nonetheless, our experiment showed that the performance of standard ChickenNet on this cross-domain task can be drastically improved by rejecting samples identified as uncertain by our method without rejections. Similar to experiment 5, we observed a permanently higher accuracy compared to the base model which increased with an increasing proportion of rejected samples. The effectiveness of aleatoric uncertainty-based rejections surpassed those based on epistemic uncertainty at the same proportion of rejections. In contrast to the previous experiment, the accuracy-rejection curves displayed smoother trends, potentially attributed to the larger scale of the MARS-Attributes dataset relative to the plumage condition dataset. This dataset size disparity might also explain why an accuracy of 1.0 was obtained at a much higher rejection rate compared to the previous experiment. Overall, this experiment showed a successful filtering of erroneous predictions on the MARS-attributes dataset, which indicates the cross-domain applicability of our method.

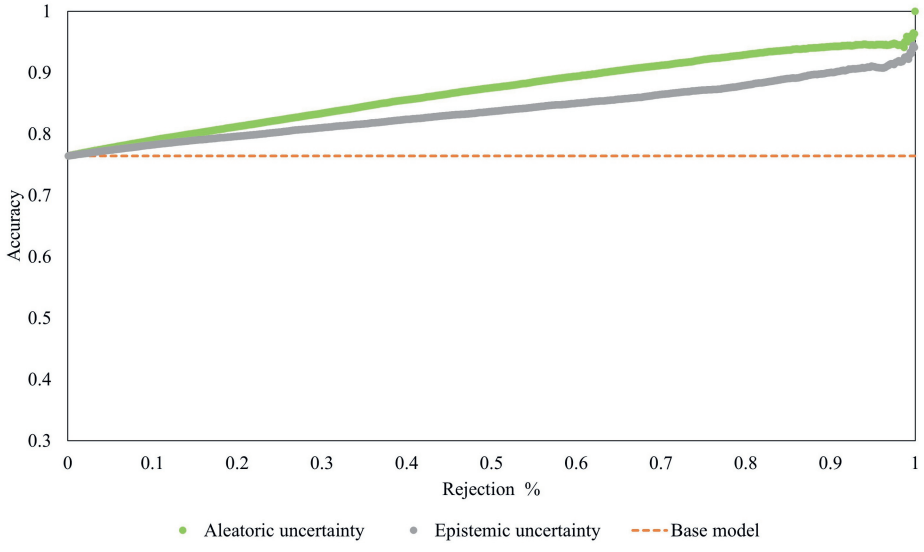


Figure 3.7: Rejection-accuracy curves for the task of age-estimation, evaluated on the MARS-attributes dataset. Rejections of samples were based on predicted aleatoric uncertainty (green) and epistemic uncertainty (gray). The orange dashed line indicates the age estimation accuracy of the standard ChickenNet model without any rejections.

3.4 Discussion

3.4.1 Effectiveness for an improved assessment performance

Results of the fourth experiment showed a strong correlation between the predictive error of plumage-condition assessments and the epistemic and aleatoric uncertainties estimated by our methods. This makes both metrics useful indicators to identify images that the neural network cannot handle well, either due to insufficient model capabilities or due to the quality of the input image. As shown in experiments 5 and 6, rejection of highly uncertain samples increased the overall accuracy compared to the base model, not only for plumage condition assessment but also for age estimation from images, a task from a totally different domain. With an increasing rejection rate, our method's accuracies outperformed those of current state-of-the-art approaches on the MARS-attributes dataset, however here it is crucial to acknowledge that the number of samples was significantly reduced through rejection, thereby impeding a fair comparison to other approaches. In our experiments, highest accuracies were obtained if the rejection rate was close to 100%, which means that only few samples were considered for assessment. However, the utilization of uncertainty estimations given by our methods for rejecting or prioritizing individual predictions was demonstrated to be an effective approach to enhance the performance of assessment tasks.

Nevertheless, it's noteworthy that if uncertain predictions are not equally distributed among the different classes of plumage conditions (or ages), an increasing rejection rate also leads to a change in the distribution of the remaining samples. This needs to be considered for applications in which the distribution of the investigated attributes is of importance. For instance, plumage-condition assessment of a chicken flock requires a good estimation of the whole flock situation. If this is obtained from samples recorded in the flock, the distribution is essential. A high rejection rate results in high accuracies, but might not represent the actual situation of the flock due to the low sampling rate and bias in the distribution due to the rejection of samples. Dealing with this could be a topic addressed by future research.

3.4.2 Accordance of uncertainty predictions and human uncertainty estimation

During our experiments, we compared the uncertainty estimations made by our methods to human determined uncertainty. In experiment 2, hens of which the plumage condition was not clearly assessable were manually identified and marked as unknown by a human assessor. For these unknown samples, we observed slightly more predictions with high aleatoric uncertainty on unknown samples than for the known samples, while no significant difference between both groups was found for epistemic uncertainty predictions. This is in line with findings from experiment 3 in which manual image modifications

that increased uncertainty from a human perspective were applied and resulted in higher aleatoric uncertainty predictions while epistemic uncertainty was not affected. However, despite the difference in aleatoric uncertainty predictions between unknown and known samples, experiment 2 also revealed a large overlap of both distributions. Thus, samples marked as unknown by the assessor were indicated as uncertain by our method and vice versa. This is likely due to the characteristics of the human uncertainty labels, which are significantly affected by occlusions, as the comparison of occlusion predictions and human labels in Table 3.2 showed. From a human perspective, a sample was marked as unknown whenever the sample quality was not sufficient for an assessment. For example, this was the case if a bird was blurred or partly occluded. A bird with completely intact plumage with minor occlusions could receive an "unknown" label, even though the visible, major parts of the plumage indicate a good condition. Due to these visible parts, the model-based prediction for this sample still has a quite low uncertainty. Even though the predictions differ from the human assessments, experiment 5 showed that a rejection of samples based on predicted uncertainties increased the overall accuracy above the level that was reached when only the samples marked by the human as known were taken. These results indicate that human assessment cannot be seen as a gold standard for uncertainty estimation. Therefore, it is important to note that high agreement of an uncertainty prediction with the given human labels might be an intuitive confirmation of the prediction, but it does not necessarily mean a good estimation or error identification. One potential approach to mitigate this issue is to conduct multiple human assessments of the same image, thereby assessing the uniformity among the annotators. This technique corresponds to a human Query-by-committee, which could be utilized for the estimation of uncertainty.

3.4.3 Suitability for assessment improvements in real-world scenarios and future applications

Our results showed that predictions for both, aleatoric and epistemic uncertainty can be used for detecting and rejecting false plumage score assessments. In experiment 3, we observed that predictions of aleatoric uncertainty from our model are sensitive to changes in the data quality. This is especially relevant in scenarios with environmental conditions that compromise the image quality, such as in farm environments. There, detections might be occluded, blurred, or contaminated. For epistemic uncertainties, such clear dependencies on image changes were not observed. High epistemic uncertainty indicates that samples are different from the training data. It is important to note that the epistemic uncertainty as calculated by our model refers only to the assessment prediction and not to the detection of the chicken. Therefore, epistemic uncertainty predictions of our model are sensitive to unusual appearance of the plumage or plumage damages. Moreover, the incorporation of Monte-Carlo dropout into our model established a dependency of the uncertainty scores on the dropout rate. We selected this parameter value

based on prior research, which indicated that a dropout probability of 0.2 was effective in estimating epistemic uncertainty. Nonetheless, the influence of varying the dropout probability on the uncertainty scores for our specific application could be the focus of further investigations.

In our approach, we also developed a method to estimate the occlusion of detected chickens. While we modeled occlusion prediction as a regression in our approach to provide a feature-based approach for sample rejection in order to exclude occluded instances from the assessment, it's worth noting that there are alternatives to this method, such as amodal instance segmentation. This approach estimates two masks (an amodal mask and a visible mask) and uses the pixel difference between them to predict occlusions (Blok et al., 2021). Although we chose regression due to its simplicity and less labeling effort, other alternatives might be equally effective for predicting occlusion. Despite demonstrating the ability to estimate occlusions using our trained model, sample rejection based on these estimations did not yield any benefits, sample rejection based on these estimations was not beneficial. It was shown that assessment errors were not exclusively due to occlusion and thus could not be eliminated by increasing the rejection threshold based on occlusion. However, depending on the use case, it could be set as a requirement to only assess fully visible birds or to consider the level of occlusion as additional information. Then, an occlusion-based rejection or the combination of occlusion estimations and other uncertainty measures would be relevant.

In this research, we evaluated the sample rejection over individual images in the test dataset. However, in combination with tracking approaches, the developed methods could also be used to find the most reliable sample out of a sequence of detections in a video stream. In that case, the most certain plumage-condition score would be assigned to the specific chicken, or a probabilistic combination of the multiple observation could be applied.

As indicated in experiment 6, the presented approach is not limited to the plumage-condition assessment of chickens. It was designed to predict uncertainties for regression-based predictions in general object-detection tasks. Thus, it could in principle be applied to all types of use cases that focus the assessment or scoring of any detected object and goes beyond the here addressed tasks.

3.5 Conclusions

In this study, we presented an approach that integrates three different uncertainty measures into an end-to-end trainable instance-segmentation and regression model. Epistemic and aleatoric uncertainty of each regression output were estimated directly using Monte-Carlo dropout and a modified loss function, respectively. Additionally, occlusion levels of detections were predicted to indirectly estimate whether a prediction is certain or not.

In our experiments, we first evaluated the three approaches regarding their individual capability of identifying unreliable predictions and we then investigated whether the prediction performance of the model can be improved by the consideration of the estimated uncertainties.

Results showed that our method for estimation of aleatoric uncertainty corresponds to human uncertainty assessment and picks up on the deterioration of the image quality. We also observed a strong positive correlation between the estimated uncertainty and the predictive error of the model for both aleatoric and epistemic uncertainty. For occlusion predictions, an accuracy of 78.3% was obtained from our method and it was found that occlusions strongly overlapped with human uncertainty assessments. Both aleatoric and epistemic uncertainties correlated well with the predictive error in the plumage condition However, no correlation between occlusion and the predictive error was found. Accordingly, rejection based on the aleatoric and epistemic uncertainties increased the model accuracy with an increasing rejection rate, while this was not observed for rejection based on occlusion.

While for use-cases such as plumage-condition assessment in farm environments, a sufficient quality of the images and therefore the estimation of aleatoric uncertainty might be of high importance, other use cases could prioritize estimation of occlusions and epistemic uncertainty. We conclude that the presented approach provides a flexible framework that allows the simultaneous consideration of multiple uncertainty measures and can be used to extend any application that combines detection and assessment tasks.

Chapter 4

FUSE: A framework for uncertainty-aware object assessment from image sequences in uncontrolled environments

This chapter is based on:

Lamping, C., Derks, M., and Kootstra, G. (2024a). Fuse: A framework for uncertainty-aware object assessment from image sequences in uncontrolled environments. *Submitted to Computers and Electronics in Agriculture*

Abstract

Computer vision and deep neural networks offer a great potential for the automation of labor-intensive and repetitive monitoring tasks, including the assessment of animals in livestock farming. However, in such uncontrolled environments, the application of vision-based methods faces several challenges. This includes environmental conditions such as illumination that affect the image quality, but also animal poses that hinder precise assessment. These challenges contribute to an inherent uncertainty associated with predictions made by neural networks. To enhance robustness of visual assessment systems, particularly in uncontrolled settings, this study proposes an approach that utilizes information from entire image sequences rather than single images. Considering the estimated uncertainty of individual predictions made on each image within the sequence, our method selectively aggregates these predictions into a final output. In our experiments, we evaluated the assessment performance of the proposed approach against conventional approaches on image level using a dataset focused on plumage condition assessment in chickens. To demonstrate the method's general applicability, we additionally utilized the MARS-Attributes dataset for person age estimation. Further, we investigated the impact of limited image numbers on our method and explored the use of different uncertainty estimators. The results demonstrated that our aggregation approach outperformed the conventional image-level model in terms of accuracy across both datasets by up to 7.15%. It also surpassed conventional methods even when confronted with limited data and when utilizing alternative uncertainty metrics. This method will therefore substantially contribute to enhancing the robustness of visual monitoring systems, especially in uncontrolled environments.

4.1 Introduction

In recent years, rapid advancements in computer vision and deep learning technologies have increased their significance in the agriculture and livestock domain. Particularly for labor-intensive and repetitive monitoring tasks like the condition assessment of animals, there is a large potential for automation. While traditionally, farmers have relied on manual inspections of individual animals' condition to ensure their health and well-being, emerging approaches aim to automate these assessments using cameras and advanced deep learning algorithms (Lamping et al., 2022).

Still, however, vision-based applications face various challenges in uncontrolled environments such as farms. Unpredictable factors such as varying illumination, occlusions, and the dynamic motion of animals can significantly impact the quality of captured images. Thus, the reliability of assessments made by deep learning algorithms is influenced by these environmental factors which results in an increased uncertainty of the prediction. Next to this uncertain nature of the data, caused by environmental influences, uncertainty can also arise from the presence of unknown input that the model has not been trained on. This is particularly relevant when considering out-of-distribution data, where the algorithm encounters samples that differ significantly from the training data distribution.

While relevant in livestock farming, the issue of dealing with uncertain predictions and low-quality input is not unique to this domain. It extends to other agricultural applications, such as weed detection (Jeon et al., 2011), and even finds relevance in non-agricultural fields like automated driving (Arnez et al., 2020). Currently, the majority of deep learning models operate at the single-image level, which poses a problem when the input image itself is of low quality, causing the predictions to be highly unreliable. This issue becomes particularly critical as many models lack the capability to provide an indicator or measure of the level of uncertainty in their predictions, leaving users unaware of the reliability of the provided results. Even if multiple observations of an object or a scenario are available, for instance through a video sequence, it is not possible to select the most reliable one without knowledge of the individual prediction uncertainties. To address this issue, this work focuses on the development of an uncertainty-aware approach for reliable object assessment from image sequences. Instead of providing an end-to-end trained solution for the assessment of sequences, our method leverages the capabilities of deep learning models operating at the image level. It selectively incorporates the information derived from multiple images within a sequence to enhance the accuracy of assessments. By adopting this approach, we aim to create a framework that is able to utilize the strength of task-specific standard models while simultaneously exploiting the additional context provided by multiple images. To achieve this, we integrate measures of uncertainty into the image-level predictions, enabling us to carefully select and combine the most reliable predictions for a comprehensive assessment.

To summarize, our main contributions are as follows:

- We propose a novel method that selectively incorporates predictions from multiple images within a sequence, considering the uncertainty of individual predictions. This method is designed to extend the capabilities of pre-trained convolutional neural networks operating at the image level.
- We propose an appearance-based clustering method for image sequences to identify and group detections providing relevant information for visual assessment tasks.
- We demonstrate the general applicability of our method by evaluating it on a dataset from the agricultural domain for the task of plumage condition assessment in chickens, as well as on the MARS-Attributes dataset for person age estimation.
- We evaluate the impact of limited data and alternative uncertainty estimators for use in our method, ensuring robust and reliable performance under varying conditions.

4.1.1 Related work

Our approach for robust object assessment utilizes multiple predictions of a standard neural network made on the individual images of a sequence and integrates them into a final assessment prediction. This methodology is grounded on two essential concepts: Firstly, the estimation of uncertainty for each individual prediction to determine the particular relevance for the final assessment, and secondly, the integration of those predictions obtained from multiple views within the sequence. In both domains, namely, the uncertainty estimation in deep learning and the field of multi-view assessment, considerable research efforts have been made over the past years.

Uncertainty estimation in deep learning
Deep learning approaches have shown great success for various computer vision task such as image classification, object detection, or segmentation. However, these models can provide unreliable predictions due to inherent randomness in the data, noisy inputs or uncertainty in the model parameters. Especially in safety-critical applications, the costs of false predictions are high. Therefore, quantifying the uncertainty of a model's prediction has become a crucial aspect of deep learning. Moreover, uncertainty can arise from various sources, which makes it essential to distinguish between different types. Two types of uncertainty are commonly distinguished; aleatoric and epistemic uncertainty (Kiureghian and Ditlevsen, 2009). Aleatoric uncertainty captures the uncertainty caused by the intrinsic randomness of an observation, such as sensor noise or ambiguities in the input data. As it is a property of the data, this type of uncertainty cannot be reduced even with more training data. Aleatoric uncertainty can further be categorized as homoscedastic uncertainty, which is constant for all inputs, or heteroscedastic uncertainty, with the latter being particularly relevant for computer vision applications (Kendall and Gal, 2017). Epistemic uncertainty, also known as model uncertainty, refers to uncertainty caused by

insufficient capabilities of the deep learning model (Lyzhov et al., 2020). The extent of this uncertainty can be mitigated by enhancing the quality of the model, increasing training data or refining data analysis techniques. Understanding the presence and magnitude of epistemic uncertainty is crucial in determining the model's limitations, especially when presented with inputs that are dissimilar to the training data. Several approaches for the estimation of both aleatoric and epistemic uncertainty have been developed. For example, Kendall and Gal (2017) proposed a Bayesian deep learning framework for quantification of uncertainty. Heteroscedastic aleatoric uncertainty was modeled as the variance of the Gaussian likelihood model and learned directly from the data through maximum likelihood training. By using a modified loss function, the neural network was encouraged to predict a higher variance for erroneous predictions. For estimation of epistemic uncertainty, Monte-Carlo dropout was utilized during inference as a variational Bayesian approximation. In general, Bayesian neural networks (BNNs) are a popular approach for the estimation of uncertainty. They treat weight parameters of a neural network as random variables with a prior distribution instead of assuming deterministic parameters. Bayesian inference then allows quantifying the uncertainty, which is associated to the model predictions by computing a posterior distribution over these variables (Gal and Ghahramani, 2016; Postels et al., 2019). Other methods for estimating uncertainty include ensemble methods (Lakshminarayanan et al., 2017; Gawlikowski et al., 2023), evidential approaches (Charpentier et al., 2020; Sensoy et al., 2018; Amini et al., 2020) and test-time augmentation methods (Lyzhov et al., 2020). Ensemble methods refer to the training of multiple models and combining their outputs, while evidential approaches aim to provide a full probability distribution over the outputs. Test-time augmentation involves applying transformations to the input data to obtain multiple predictions and estimate uncertainty. Overall, these techniques aim to quantify both aleatoric and epistemic uncertainty and have been applied on a variety of computer vision task. As uncertainty quantification allows the numerical comparison of neural network predictions, it provides a useful basis for the aggregation of multiple predictions on a sequence of images.

Multi-view assessment

Deep learning methods for vision-based classification or regression typically rely on single-image inputs and may not capture the complexity of real-world scenes that often have multiple perspectives or views. To address this limitation, several approaches have been developed, which can integrate information from different views to make predictions. It is worth noticing that the term "view" in this context does not necessarily imply different perspectives of looking at a scene or object. Rather, it can refer to different modalities, angles, or representations that are unique and informative. Regarding the assessment of an object based on a sequence of images, different options to incorporate information from multiple views can be distinguished: One option involves the selection of a single, representative image from the sequence, commonly referred to as key frame extraction. Such a key frame usually corresponds to a frame which has a high visual quality but also sum-

marizes the content of the given images. In traditional approaches, key frames were often determined through boundary-based techniques, which simply select the first or middle frame of a sequence (Boreczky, 1996), or through quality estimation methods applied to each image (Lu et al., 2015). Alternatively, frames with least differences from other frames were selected using a variety of similarity measures (Zhuang et al., 1998; Sadiq et al., 2020). Recent approaches mostly used content-based strategies, in which visual features of each frame were extracted and analyzed to determine most relevant frames. For example, deep convolutional neural networks were utilized to learn those features and to estimate the importance of a frame within a sequence (Al Nahian et al., 2017; Ren et al., 2020). Another option involves the aggregation of information from multiple views or images instead of selecting a single view or image. One popular technique is multi-view learning, which trains a neural network using distinct viewpoints of the same data to learn a combined representation that encompasses the information from those viewpoints. A wide range of supervised and unsupervised approaches, such as multi-view clustering (Chen et al., 2022a), multi-view representation learning (Tian et al., 2020b; Bachman et al., 2019; Wang et al., 2021), and multi-view classification (Kendall and Gal, 2017; Seeland and Mäder, 2021; Kiela et al., 2018) have been proposed in the field of multi-view learning. Recent studies further incorporated the estimation of uncertainty for each view into multi-view learning approaches. For example, Han et al. (2021, 2022) dynamically integrated multiple modalities at an evidence level to ensure the reliability and robustness of a classification task in the presence of noisy and out-of-distribution data. These methods were designed as an end-to-end trainable framework and aimed for decision explainability by providing the uncertainty learned for each view. Instead of developing a model that is capable to process multi-modal inputs, other studies utilized late fusion, which involves the combination of multiple predictions of a deep learning model on different representations of the same scene or object into a single prediction. Alternatively, multiple models can be trained on each view to then combine their predictions using the late fusion technique. In Wang et al. (2022b), the authors presented fusion-based approaches for anomaly detection, including fusion-based multi-view solutions that merge data embeddings obtained from various modalities into a joint embedding which is then used for anomaly detection. Here, it was shown that simple averaging could serve as a robust baseline for the fusion of multiple views. Other approaches adopted more sophisticated late fusion strategies that considered certainty of the different views for fusion. For example, Liong et al. (2020), introduced a method for LiDAR semantic segmentation that fuses information from multiple projection-based networks through late fusion. In this approach, the disagreements between class predictions were considered as a measure of uncertainty. Then, fusion of multiple individual network predictions was performed using an extra network to refine the results. Similarly in Morvant et al. (2014), diversity of different classifier predictions was taken into account for late fusion. Various uncertainty measures were considered in Tian et al. (2020a). This work proposed an uncertainty-aware fusion approach for effectively fusing inputs from an arbitrary set of modalities or networks. With each measure

capturing a different aspect of uncertainty, uncertain outputs of the different modalities were integrated into a final prediction for semantic segmentation. These late fusion methods combined multiple predictions and partially integrated uncertainty measures which provides decision explainability for the final prediction. However, multi-view fusion in most approaches referred to multi-modal representations of a static image and did not take into account the temporal component of the views. This introduces additional complexities, including variations in the number of images to be considered for predictions or shifts in perspectives across individual views. Another approach for the integration of multiple views are models using attention mechanisms which selectively focus on specific views of the input that are deemed to be most relevant for a given task. These models are popular for their effectiveness in handling sequential data. Consequently, despite their application on multi-modal data (Tian et al., 2020c; Wei et al., 2020; He et al., 2021b; Chen et al., 2020), they are frequently utilized for data including a temporal component such as video sequences to prioritize individual frames of the sequence (Li et al., 2020a; Chen et al., 2019; Pei et al., 2017; Peng et al., 2018). For instance, attention mechanisms have been incorporated into CNNs in order to recognize facial expressions from image sequences (Li et al., 2020a) or for classification of pedestrian attributes from surveillance camera videos (Chen et al., 2019). Pei et al. (2017) combine the concepts of attention models and gated recurrent networks for the classification of noisy image sequences. This approach encouraged the interpretability of predictions as it utilized temporal attention weights to indicate the significance of each time step in a given sequence. In Heo et al. (2018), aleatoric uncertainty was introduced to the attention mechanism so that attention was predicted with a lower variance if the model was confident about the contribution of a certain feature. In case of uncertain contribution, the variance of the prediction was higher. However, this was applied on classification on time-series data of medical records rather than on images or image sequences.

In summary, while multi-modal approaches have encompassed a variety of methods for multi-view assessment, the existing work on image sequences reveals two severe limitations:

1. End-to-end trained models as they are frequently used in multi-view learning often suffer from a poor explainability. For most of these models, it is hard to understand why they make a particular prediction for a sequence, or why they prioritize a certain view within the sequence.

2. Attention-based and other multi-view models developed for the purpose of image sequence assessment require training on sequential data. Consequently, a substantial volume of annotated training data in the form of image sequences is essential for each assessment task to be trained. These datasets are relatively scarce in comparison to datasets composed of single images. For instance, widely-used datasets like ImageNet (Deng et al., 2009), often leveraged for pre-training primarily consist of

single images. Similarly, the majority of task-specific convolutional neural networks are trained on single images, posing challenges when adapting them for complete sequences.

This study addresses these issues by presenting an approach that integrates multiple predictions of standard, image-level-neural networks into a final assessment taking into account the uncertainty of each individual prediction. Thus, we aim to enhance decision interpretability and establish a method applicable across a wide range of tasks, as detailed in the subsequent sections.

4.2 Material and methods

As most neural networks traditionally operate on image-level, their predictions are based on the information provided from a single view. To enhance the robustness of pre-trained convolutional neural networks for object assessment, our method extends the assessment process to encompass entire sequences of images, rather than individual frames. Notably, our approach is not limited to objects, but also refers to the assessment of animals or persons, collectively denoted as 'identities' hereafter. For each identity, the method selectively incorporates predictions from multiple images within a video sequence, while considering the uncertainty associated with each individual prediction. This uncertainty-aware multi-view assessment leads to a final assessment prediction for the identity of interest. Applying a detection-and-assessment model that operates on image level to a sequence of images initially leads to a list of unrelated predictions. First, these predictions must be matched to their corresponding identities. In this work, this alignment was accomplished by using ground-truth identity information. The detections of one identity within consecutive frames are visually very similar, therefore containing similar information. Nevertheless, some detections may be dissimilar to others, for example if the object of interest moves or the viewpoint changes exposing a different part of the identity. As a result, multiple views of an identity might emerge from a sequence, where each view adds new information, but where some views could be more relevant than others. An intuitive method for obtaining an optimal assessment from a sequence could be to select the most certain assessment. However, the most certain assessment is not necessarily the best assessment. For instance, assessments made from different viewpoints can be contradictory to each other if a certain view reveals relevant features of an identity affecting the assessment, while those features are not visible in another view. An example is the assessment of a chicken's plumage. If a damage remains hidden from a particular perspective, assessments made from that viewpoint may be certain about the plumage's intactness. However, if the chicken changes its position, thereby revealing the previously concealed damage, the initial assessment is found to be incorrect. Therefore, some views could be more important than others in facilitating a holistic assessment of the target, as they provide essential information necessary for the final classification. To consider this

for the assessment of an identity and to distinguish between different views, it is required to know which detections are similar and which provide new information before utilizing them for a final assessment. While predefined features such as the pose of a person or an animal might serve as a valid metric for distinguishing between views in certain use cases, this approach is limited to those features and not capable to dynamically consider other factors that influence the information content of a detection, such as occlusions. Instead, we propose the clustering of detections by their appearance to identify distinct views within the sequence. Figure 4.1 provides a visual representation of the intended clustering procedure for this method when applied to a sequence capturing the movement of a chicken.

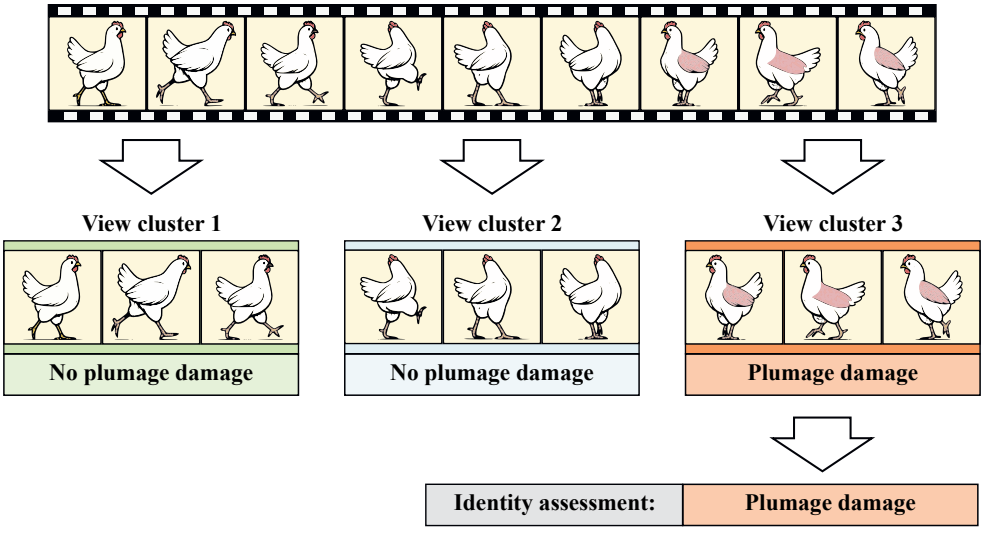


Figure 4.1: Intended procedure for identifying distinct viewpoints from a sequence of detection through clustering. For the given example of a moving chicken, the three cluster represent views from the right, rear, and left sides of the animal. While view clusters 1 and 2 do not exhibit any plumage damages in the chicken, such damages are revealed in the third cluster, impacting the overall assessment of the chicken (identity).

Our approach first processes each image from the sequence and generates detections and assessment predictions together with uncertainty estimates for each identity as presented in Section 4.2.1. Subsequently, an appearance-based clustering method is used to group all visually similar detections together and separate dissimilar ones (Section 4.2.2). Finally, the predictions per cluster are aggregated to derive an assessment prediction and associated uncertainty for each cluster, which is then used to generate a final prediction for each identity (Section 4.2.3). An overview of the complete method is provided in Figure 4.2.

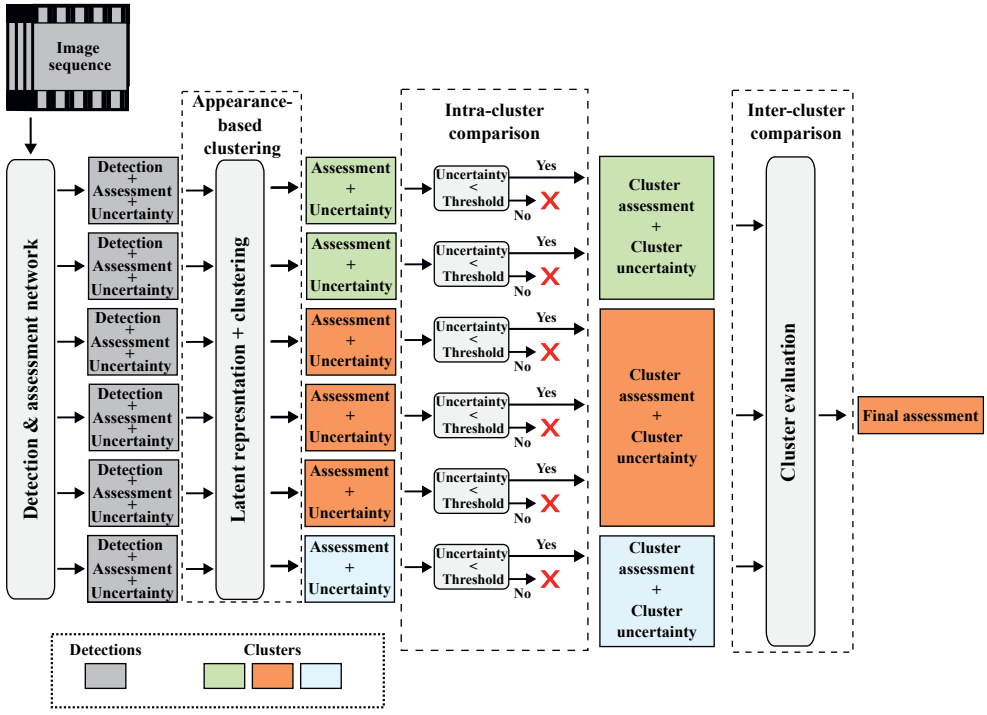


Figure 4.2: Generic pipeline of the developed method for image sequence assessment, consisting of a detection and assessment network for the assessment on instance-level, followed by appearance-based clustering of the individual detections. Subsequently, assessments within a cluster are evaluated to obtain a single assessment for each cluster. Those are then compared with each other, leading to a final assessment.

4.2.1 Detection and assessment network

For the uncertainty-aware aggregation of multiple image predictions, first an uncertainty assessment method is required. As outlined in Section 4.1, there are various approaches to estimate uncertainty in neural network predictions such as end-to-end solutions (Kendall and Gal, 2017; Postels et al., 2019) and inference sampling approaches (Kendall and Gal, 2017; Lyzhov et al., 2020; Gal and Ghahramani, 2016), which allow the Bayesian interpretation of standard architectures without the need to retrain the model. In this study, we employed ChickenNet (Lamping et al., 2022), a convolutional neural network for object detection, segmentation and quality assessment, which included the prediction of multiple types of uncertainties of a regression output without requiring ground-truth uncertainty labels during training (Lamping et al., 2023). ChickenNet was developed by extending the Mask R-CNN architecture with an additional regression output for the purpose of plumage condition assessment in chickens. It detects and segments object

instances from single images, while predicting an assessment score for each instance. To estimate both data- and model-related uncertainty of the regression output, the model integrates estimators for aleatoric and epistemic uncertainty into its architecture. While primarily developed for plumage condition assessment, the model was designed to predict uncertainties for regression-based predictions in general object-detection tasks. For the prediction of aleatoric uncertainty together with the regression score, a modified loss function was implemented following the approach of (Kendall and Gal, 2017). Instead of only predicting a single regression output $\hat{y}_i^{score}$, the presented model simultaneously predicts a measure of aleatoric uncertainty, given by the variance σ_i^2. With y_i^{score} denoting the ground-truth regression score and N denoting the number of samples, the loss function is defined as:

$$L_{score} = \frac{1}{N}\sum_{i=0}^{N}\frac{(y_i^{score} - \hat{y}_i^{score})^2}{2\sigma_i^2} + \frac{1}{2}\sigma_i^2 \tag{4.1}$$

With this, aleatoric uncertainty was learned directly from the data during model training, aiming to give a sense of the model's predictive error. The first term of the function encourages the model to minimize the predictive error, while predicting a high variance also reduces the contribution of this term to the overall loss. As the second term penalizes large variances, the present loss function instructs the network to predict higher variance for uncertain predictions and lower variance for correct ones. Calibrated uncertainty predictions are needed for the comparison of uncertainties among multiple assessments as well as the thresholding of uncertainty values using a fixed threshold. Intuitively, the predicted uncertainty of a regression output should match the difference between the prediction and the ground-truth value. As there is no ground-truth uncertainty for training the aleatoric uncertainty of a prediction, calibration of the uncertainty estimation cannot be guaranteed by solely using the loss function shown in Equation 4.1. Therefore, following the approach of Feng et al. (2019), in the present study, we additionally devised a simple calibration term which was incorporated into the total loss of ChickenNet by adding it to L_{score} . This term forces σ_i^2 to align with the predictive error, resulting in a calibrated score loss, defined as:

$$L_{score} = \frac{1}{N}\sum_{i=0}^{N}\frac{(y_i^{score} - \hat{y}_i^{score})^2}{2\sigma_i^2} + \frac{1}{2}\sigma_i^2 + |\sigma_i^2 - (y_i^{score} - \hat{y}_i^{score})^2| \tag{4.2}$$

Aligning the aleatoric uncertainty prediction to the predictive error allows setting an interpretable threshold to filter uncertain predictions. In addition to aleatoric uncertainty, ChickenNet provides an estimation of epistemic uncertainty for the regression output by applying the Monte-Carlo dropout method (Gal and Ghahramani, 2016). During inference, multiple forward passes with varying dropout patterns are performed to approximate

the distribution of the output predictions and estimate the epistemic uncertainty of the model. Previous experiments showed that both estimation methods, the adapted loss function as well as the Monte-Carlo dropout method, were able to capture the uncertainty in plumage condition assessments with a strong positive correlation between the predicted uncertainty and the predictive error of the model's regression output (Lamping et al., 2023). In the present work, we primarily focused on aleatoric uncertainty, which relates to uncertainty in the image data, making it more intuitive for human interpretation of the results. Nevertheless, we also conducted an experiment that explored the utilization of epistemic uncertainty as an alternative metric for assessing individual instance predictions. Our approach utilizes the ChickenNet architecture to process each image within the sequence and facilitate the shift from the image level to the detection level. Applied on a sequence, it outputs all individual detections from that sequence together with their associated assessment scores and uncertainties. Assigned to their corresponding identity, these individual detections serve as the basis for the subsequent stages of our approach.

4.2.2 Appearance-based detection clustering

To group detections that provide similar information and distinguish them from dissimilar ones, we employed an appearance-based clustering approach. This allows considering the perspective or level of information each detection offers before integrating them into a final assessment. The clustering first requires a latent representation of the different detections, described in the following, which is then used to cluster observations in that latent space.

4.2.2.1 Appearance representation

To form meaningful clusters of detections from an image sequence, detections within a cluster should be more similar than detections between clusters. Representing the visual appearance of the detections as embeddings in a lower-dimensional feature space allows to efficiently measure the similarity between the detections using a distance metric. Thus, the quality of the clusters heavily depends on the representation used for clustering. To identify detections that provide new information for assessment, we propose clustering based on their appearance to capture similarities or dissimilarities. To this end, we computed an appearance descriptor of each detection. The descriptor was obtained from a shallow CNN, as presented in Wojke and Bewley (2018), that had been trained to construct feature embeddings from detections. It provides a method for learning embeddings from images such that they maximize inter-class cosine similarity and minimize intra-class cosine similarity, meaning that the cosine similarity between two embeddings corresponding to images of the same class are likely to be closer than two embeddings corresponding to different classes. This has been shown to be very effective for representation learning, e.g. in the context of person re-identification (Wojke et al., 2017). In this approach, we

utilized the embedding model as proposed in Wojke and Bewley (2018), pre-trained on a large-scale person re-identification dataset (Zheng et al., 2016). This embedding model was then applied on each of the bounding box predictions given by our detection and assessment network to obtain an appearance descriptor for each detection. This resulted in an appearance vector of length 128 for each detection.

4.2.2.2 Clustering Algorithm

To cluster the different samples of an individual, we applied the mean-shift algorithm (Fukunaga and Hostetler, 1975) on the computed vectors of all detections belonging to a single individual. Mean shift is a non-parametric algorithm that can be used to group data points based on their similarity in a feature space. Contrary to the popular K-Means cluster algorithm, it does not require specifying the number of clusters in advance. Instead, the number of clusters is determined by the algorithm with respect to the data. This was essential for our approach, as the ideal number of clusters in our scenario is dependent on the diversity in the appearance of the detections. The higher the number of different perspectives or appearances, the higher the ideal number of clusters. As input, the algorithm received all appearance vectors of an individual together with the radius of the local window used to compute the mean-shift updates. The radius of the local window, was obtained by computing the distances between each pair of appearance vectors from the input. The radius was then set as the median of those, introducing a distance measure that adapts to the data rather than relying on a fixed distance. Initially, Mean Shift clustering treats each data point as the center of its own cluster.

As our approach utilized the appearance-based clustering for the uncertainty-aware assessment of an identity, the prediction $\hat{y}_i$ and prediction uncertainty σ_i^2 of the respective detection were also assigned to the clusters. Thus, for each cluster $c_j = \{a_1, ..., a_{n_j}\}$ with $a_i = \{\hat{y}_i, \sigma_i^2\}$, we obtained a set of assessments clustered by the similarity of their visual appearance. Examples of clustered detections for two identities from different datasets are visualized in Figure 4.3.

4.2.3 Cluster aggregation

The appearance-based clustering resulted in groups of detections and their uncertainty-aware assessments. We aggregated the assessments using a two-step approach. First, we combined the assessments within a cluster to obtain one prediction and its corresponding uncertainty per cluster. Subsequently, in the second step, we determined the most representative cluster while considering the uncertainty associated with each cluster.

4.2.3.1 Intra-cluster comparison

Due to the shared visual characteristics among all detections in a cluster, the assessments in one cluster rely on comparable information, which makes the corresponding assess-

Figure 4.3: Clusters resulting from our appearance-based clustering approach applied on an identity from the MARS-Attributes dataset (left) and the chicken dataset (right). For each of the four clusters, exemplary detections are visualized.

ments more likely to be also similar. To compute a single assessment output for each cluster, we aimed to combine all assessments within this cluster while considering their individual uncertainties. The inverse of this uncertainty value can serve as a measure of the assessments relevance in determining the final output of the cluster. However, simply choosing the assessment with the lowest uncertainty from each cluster may result in a high sensitivity to (false) outliers among the uncertainty predictions. To be robust to noise, we propose a certainty-weighted mean for each cluster, where certainty is defined as the inverse of the associated uncertainty. Weighting individual predictions by their certainty results in predictions with high certainty to contribute more to the output than ones with low certainty and is expected to reduce the impact of erroneous predictions. Given a multi-sample cluster c_j, the weighted mean of a cluster was defined as:

$$\hat{Y}_{c_j} = \frac{\sum_{i=1}^{n_j} w_i \hat{y}_i}{\sum_{i=1}^{n_j} w_i} \quad \text{and} \quad w_i = \frac{1}{\sigma_i^2} \tag{4.3}$$

The uncertainty of each cluster was estimated using the variance of the weighted mean. Considering the inverse-variance weighing, which minimizes the variance of the mean as shown in Meier (1953), this is defined as:

$$\Sigma_{c_j}^2 = \frac{1}{\sum_{i=1}^{n_j} \frac{1}{\sigma_i^2}} \tag{4.4}$$

The estimation of uncertainty for each weighted mean allows a comparison of all clusters of an identity, as described in the next section. However, this comparison can be negatively affected by clusters with either single samples or only samples with high uncertainty. This challenge arises, for instance, if certain detections significantly differ in appearance from the rest, such as when an object is in motion, resulting in blurred detections and uncertain assessments. In such cases, these assessments may be allocated to a distinct cluster characterized by its limited number of samples and high uncertainty. To avoid those clusters, we discarded highly uncertain assessments by setting a threshold τ to the uncertainty metric before computing the weighted average of a cluster. The weight w_i of a sample i was defined as $w_i = 0$ if $\sigma_i^2 > \tau$ so that a sample was ignored in the weighted average if its uncertainty exceeded the given threshold. By rejecting those samples, our algorithm can classify a sequence as not assessable if $\sum_{i=0}^{n} w_i = 0$. Considering an integer-based labeling of ground truth assessments, as it was given in the evaluated datasets, the maximum error, which can still result in a prediction considered as correct is 0.5. Since the uncertainty prediction for an assessment was trained to match the squared expected error between the score prediction and its ground truth, the uncertainty threshold was accordingly set to $\tau = 0.25$. This procedure results in an assessment prediction, $\hat{Y}_{c_j}$, and uncertainty $\Sigma_{c_j}^2$, per cluster.

4.2.3.2 Inter-cluster comparison

After computing the assessment prediction and the corresponding uncertainty for each cluster, these clusters need to be evaluated and compared to each other to obtain a final assessment for each identity. To select the optimal cluster for the final assessment, we distinguished two cases. The first case refers to assessments that are an unavoidable outcome of the existence of specific indicators. An example is the presence of plumage damages in chickens. As soon as damages are visible, the plumage cannot be assessed as completely intact anymore. Other examples would be rotten spots for the assessment of apples or cracks in the surface of a metal component. If an indicator is present once, the assessment of the whole identity cannot improve with the consideration of additional assessments. However, such a dependency on certain indicators can lead to contradictory assessments, depending on the particular perspective on an object. Suppose J different view clusters, each having an assessment prediction $\hat{Y}_{c_j}$ and an associated uncertainty $\Sigma_{c_j}^2$. While all those cluster predictions might be correct, considering the given information, clusters containing detections in which the relevant indicators are visible, are more important than clusters without those indicators. For instance, different viewpoints of a single object, represented by the clusters, can reveal different visual information of the object, leading to different assessments of the object's condition. Clusters with low uncertainty, in which defects are visible, should therefore be preferred for the final assessment. Thus, assuming a higher assessment score indicates a worse condition, the overall assessment score can increase but not decrease with an increasing number of detections. This prioriti-

zation of predictions affected the final assessment of an identity, so that we formulated the cluster selection as the maximum of the cluster predictions, weighted by their particular uncertainty. This ensures the prioritization of higher cluster predictions if predictions are equally certain but also ensures that high but uncertain predictions are neglected:

$$Y = \hat{Y}_{c_k} \quad \text{and} \quad k = \underset{0<j<J}{\arg\max} \frac{\hat{Y}_{c_j}}{\sigma^2_{c_j}} \tag{4.5}$$

The second case refers to applications in which the assessment is not dependent on the presence of indicators for or against a particular assessment. Example application, in which this approach might be chosen, are the age estimation of humans or weight estimation from images. In this case, additional assessments from multiple perspectives might change the outcome in both directions. Therefore, we based the prioritization of the individual clusters only on their associated uncertainty. For the final assessment of an identity, the cluster with the lowest uncertainty was chosen. In this case the final output is defined as:

$$Y = \hat{Y}_{c_k} \quad \text{and} \quad k = \underset{0<j<J}{\arg\max} \frac{1}{\sigma^2_{c_j}} \tag{4.6}$$

4.2.4 Experiments

Experiments were conducted with the objective to compare the performance of our proposed method with standard instance-level approaches and to investigate the strength and weaknesses of our approach. To this end, experiment 1 focused the direct comparison with the standard implementation of ChickenNet. Following that, experiment 2 evaluated the effect of different input quantities on our approach. While the first two experiments considered the aleatoric uncertainty prediction of the assessment network for weighting the predictions, the third investigated the alternative use of epistemic uncertainty. This aimed to determine the effectiveness of our method across various uncertainty metrics that may differ depending on the specific use cases. Approaches were compared on two different datasets for visual assessment tasks, one in the domain of plumage condition assessment in laying hens and one for human age estimation.

4.2.4.1 Data and annotations

Our approach addresses a general method for robust multi-view assessment from image sequences. The chicken dataset on which the present work was focused, includes image sequences of one or multiple chickens, labeled with bounding boxes, segmentation masks and scores for the condition of the plumage (Lamping et al., 2022). In order to investigate the general applicability on visual assessment tasks, our experiments were not limited to the small-scale chicken dataset, but also extended to the MARS-Attributes dataset (Chen

et al., 2019), a dataset, which can be utilized for human age estimation from surveillance camera sequences. While both datasets were from different domains, they share a similar structure. Ground-truth labels for plumage condition scores and ages were given per identity, meaning each label corresponds to either a chicken or a person. For chickens, scores from 0-2 were annotated, with a score of 0 indicating perfect plumage condition, plumages with minor damages were given a score of 1 and heavily damaged plumages received a score of two. In the MARS-Attributes dataset, age attributes ranged from 0-3, indicating children, teenager, adults and elderly people. Further, both datasets comprise an id label for each identity. These ids were needed to assign individual detections to the corresponding person or chicken, respectively. It's worth noting that the chicken dataset contains one or more identities per image, whereas the MARS-Attributes dataset contains only one identity per image. For each identity, both datasets include one or more tracklets, which represent a sequence of instances, as shown in Figure 4.4. An instance denotes a detection of an identity at a certain timestep of the sequence.

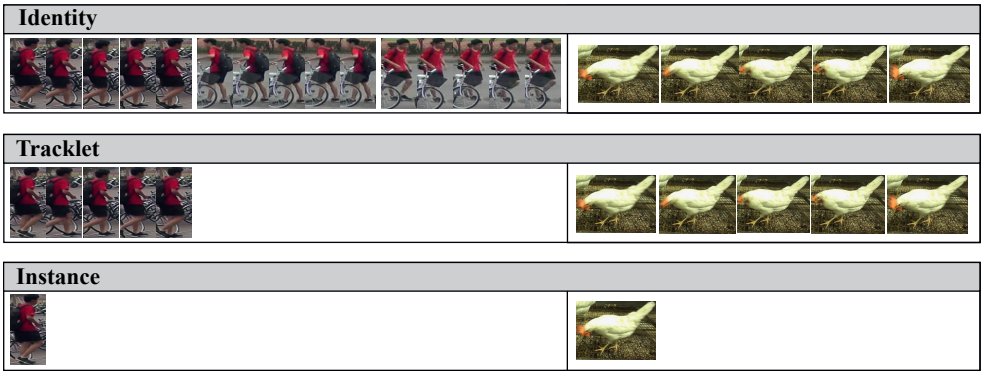


Figure 4.4: Structure of the MARS-Attributes dataset (left) and chicken dataset (right). Per identity, the MARS-Attributes contains multiple tracklets, while the chicken dataset consists of a single tracklet per identity.

The detection-and-assessment model was trained on image level, separately for each dataset. For the chicken dataset, the training data consists of 1888 images with 5057 chicken instances, obtained from video sequences recorded in a commercial laying hen farm following the procedure described in Lamping et al. (2022). For the MARS-Attributes dataset, the training data includes 509,914 images with one instance each. Using the respective network weights of each dataset, our method was tested utilizing the image sequences from the test data of both datasets. The chicken dataset comprised 35 identities and tracklets, totaling 5133 instances. The test data of the MARS-Attributes dataset consists of 634 identities, captured in 8058 tracklets with 509,990 instances in total. Here, images without a ground truth label were ignored.

4.2.4.2 Experiment 1 - Comparison to standard ChickenNet

In the first experiment, we compared the performance of our proposed approach to the predictions generated by the conventional ChickenNet model. While ChickenNet originally predicts a score on instance-level, our method leverages the aggregation of multiple individual assessments of a sequence to obtain a final prediction as described in Section 4.2.3. However, the level on which these assessments are aggregated for a final assessment can be varied. The sequential structure of the present datasets allows a prediction at each timestep of a tracklet, considering all previous assessments upon this timestep, but also enables a single prediction for each tracklet or for each identity by aggregating all assessments from the respective tracklet or identity. To compare our method with assessment on instance level, we distinguished between these alternative aggregation levels. This resulted in a comparison of four different evaluation approaches for our method:

Instance level, aggregation per tracklet
Predictions of our method were evaluated on instance level. Each prediction $Y_{id,t,k}$ for an identity id at a time step k of a tracklet t considered all previous predictions on this identity from tracklet t , starting from $k = 0$.

Instance level, aggregation per identity
Predictions of our method were evaluated on instance level. Each prediction $Y_{id,t,k}$ for an identity id at a time step k of a tracklet t considered all previous predictions on this identity from all previous tracklets.

Tracklet level
Predictions were evaluated on tracklet level. Our method was applied on all instances of a tracklet, so that per tracklet and identity, a single prediction $Y_{id,t}$ was given.

Identity level
Predictions were evaluated on identity level. Our method was applied on all instances and all tracklets of an identity, so that per identity a single prediction Y_{id} was given. As the chicken datasets contained a single tracklet per identity, the identity level was equal to the tracklet level for this dataset. Figure 4.5 visualizes the different evaluation approaches using an identity from the MARS-Attributes dataset as example.

Evaluating approaches for object detection and assessments at the instance level means to verify whether the predicted values of each instance match the corresponding ground truth. As we additionally evaluated our method on tracklet and on identity level, we also obtained single predictions per tracklet and per identity, which were then compared to their corresponding ground truth values. For the present datasets, ages and plumage condition scores were represented by discrete numerical labels. Therefore, a prediction was considered correct, if the predicted value fell within the range associated with the corresponding class. Thus, the accuracy denoted the proportion of correct predictions among the total number of samples. Additionally, we analyzed the mean squared error

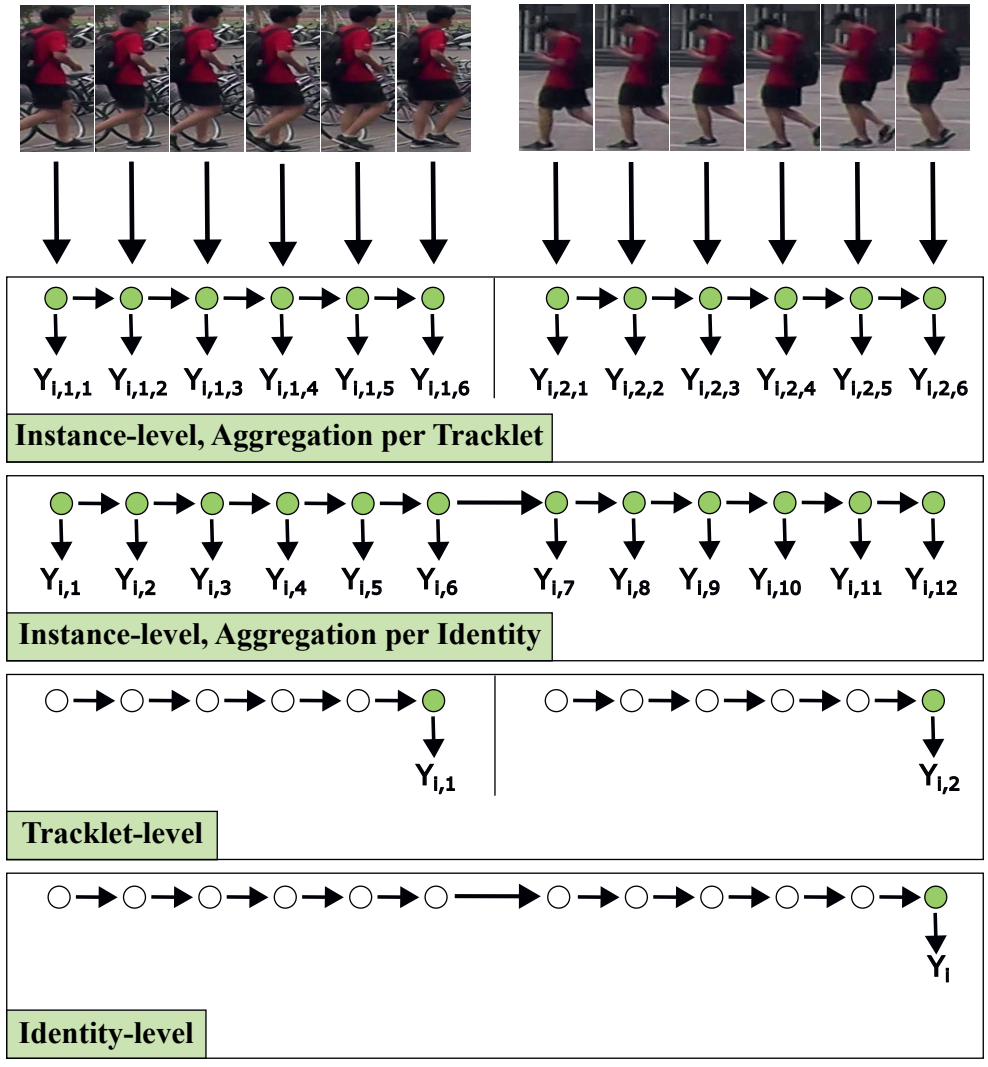


Figure 4.5: Evaluation approaches on instance-, tracklet-, and identity-level. Instance-level approaches result in a prediction for each instance and can be obtained by either aggregating all instances of a tracklet or all instances of an identity. Tracklet-level approaches result in one prediction per tracklet considering all instances of a tracklet. Identity-level approaches result in one prediction per identity, considering all instances of an identity. The given example illustrates and identity consisting of two tracklets and six instances per tracklet.

(MSE) for each prediction. Again, it is worth noting that in this experiment, predictions were obtained per instance, per tracklet and per identity as shown in Figure 4.5.

4.2.4.3 Experiment 2 - Effects of data quantity

Conventional instance-level approaches do not harness the advantages of image sequences, as they treat each frame within a sequence independently. However, given a sufficient number of images per identity and recordings captured from multiple perspectives, it could be expected that a simple average of all available predictions from an instance-level model would also result in an accurate assessment of an identity - without the need for a selective approach as we presented it in this study. Therefore, this experiment compared our method to a simple averaging approach on both datasets. For the chicken dataset, which consists of a single tracklet per identity and includes instances of chickens captured in different poses, we expected that averaging the assessments across all instances would result in an increased assessment accuracy compared to the standard ChickenNet as the number of considered instances increases. The MARS dataset consists of multiple tracklets where instances within each tracklet show a high similarity in terms of perspective and pose of the person while the perspective differs between the tracklets. Therefore, our expectation was that the accuracy resulting from averaging would increase with the inclusion of a greater number of tracklets, while the number of instances per tracklet would have a relatively minor impact. In contrast to the averaging approach, our method presented in this study considers the uncertainty of individual assessments to prioritize the most relevant predictions for a final assessment. This aims to enable precise assessments from sequences, even in cases where multiple predictions within a sequence may be incorrect. Thus, we expected higher accuracy levels when confronted with limited data compared to conventional averaging techniques. To evaluate this hypothesis, we investigated the advantages of our method on limited data. We manipulated the number of instances per tracklet and the number of tracklets per identity in both datasets to compare the performance of our method in different scenarios. We varied the range of instances per tracklet between 3 and 20. Additionally, for the tracklets per identity in the MARS-Attributes dataset, we considered a range of values, including 1, 2, 5, 10, 30, 50, 80, and 100. The chicken dataset remained limited to a single tracklet per identity. The obtained predictions were then compared to simple averages of all predictions per tracklet and to averages of all predictions per identity.

4.2.4.4 Experiment 3 – Alternative uncertainty quantification

The preceding two experiments were performed using aleatoric uncertainty to weight the individual predictions. Aleatoric uncertainty corresponds to uncertainty in the data and is therefore especially relevant in scenarios with environmental conditions that compromise the image quality. However, our method was designed as a general framework for uncertainty-based multi-view assessment from image sequences which allows the incorpo-

ration of various uncertainty indicators. In this experiment we evaluated the assessment performance of our approach while using epistemic uncertainty to weight the individual samples. We hypothesized that also with this alternative measure, our method outperforms traditional approaches. Epistemic uncertainty as implemented in the original ChickenNet architecture (Section 4.2.1) is particularly relevant since, unlike aleatoric uncertainty, it does not require the training of an uncertainty estimator specific to the particular dataset. Instead, epistemic uncertainty is estimated using Monte-Carlo dropout during inference, which allows to utilize pretrained models for conventional assessment on instance-level within our method. To test the hypothesis, we substituted the aleatoric uncertainty estimation with the epistemic uncertainty estimation derived from ChickenNet and evaluated it on both datasets, analogous to experiment 1.

4.3 Results

The results are presented in the order of the experiments. First, the comparison of our method with instance-level assessments is demonstrated. Subsequently, the impact of data quantities on our method, as well as the outcomes derived from our method utilizing epistemic uncertainty, are presented.

4.3.1 Comparison to instance-level assessment

The first experiment aimed to compare our method to a standard approach for visual assessments on instance level. Four alternative aggregation approaches were evaluated and compared to instance level assessment, which does not aggregate any information. Table 4.1 presents the accuracies obtained from the different aggregations for both, the chicken and the MARS-Attributes dataset.

Results showed that, on both datasets, all four aggregation approaches increased the assessment accuracy and decreased the mean squared error compared to the baseline model. For both, the chicken and MARS-Attributes dataset, best performance was obtained when predictions were aggregated on identity-level, resulting in a single assessment per identity. For the chicken dataset, this approach yielded an accuracy of 88.57% and a mean squared error (MSE) of 0.1. In comparison, the baseline model achieved an accuracy of 85.40% and an MSE of 0.18 for the chicken dataset. Similarly, for the MARS-Attributes dataset, identity-level aggregation resulted in an accuracy of 83.57% and an MSE of 0.14, while the baseline achieved an accuracy of 76.42% and an MSE of 0.20. Furthermore, results indicated that employing our method for instance-wise prediction on the chicken dataset increased accuracy to 87.44%, with an MSE of 0.17. For the MARS-Attributes dataset, the performance at the instance level, particularly when aggregated per tracklet, was almost on par with the tracklet-level performance. The difference in accuracy between instance-level with an aggregation per tracklet and tracklet level was only 0.27% and 0.73% between instance-level with an aggregation per identity and identity level.

Table 4.1: Accuracies and MSE for the assessment predictions obtained from our method as well as the baseline model on the chicken and MARS-Attribute dataset. Metrics were evaluated using five different evaluation approaches on instance, tracklet or identity level.

Aggregation Method	Chicken dataset		MARS-Attributes dataset	
	Accuracy (%)	MSE	Accuracy (%)	MSE
Instance level, Baseline (No aggregation)	85.40	0.18	76.42	0.20
Instance level, Aggregation per Tracklet	87.44	0.17	82.38	0.17
Instance level, Aggregation per Identity	87.44	0.17	81.84	0.16
Tracklet level	**88.57**	**0.10**	82.09	0.17
Identity level	**88.57**	**0.10**	**83.57**	**0.14**

Figure 4.6 illustrates examples showcasing the underlying principle of our method using three tracklets from the chicken dataset. The figure provides an instance-wise comparison between the predicted plumage scores of the baseline model and the predictions obtained from our approach, using aggregated information of the entire tracklet. The key observation is that our method was able to select correct predictions from a sequence of predictions, even though false predictions were made by the baseline model on several instances of the tracklet. This was particularly observable for the second tracklet as our method successfully maintained accurate predictions for all instances, despite the baseline model producing three false predictions among the tracklet. Conversely, in the example of tracklet 3, our method ignored those false predictions that were based on blurred instances, even though these were constituting the majority of the tracklet with only two out of seven correct predictions from the baseline model. Our method employs a selective approach, meaning it does not necessarily consider all available predictions of a sequence. Instead, it selects predictions based on their individual predictive uncertainty. If this uncertainty associated with a particular instance exceeds the given threshold, this assessment is rejected and not considered for further processing. In case that all instances of a tracklet or identity surpass the uncertainty threshold, the entire entity is rejected and not assessed. Therefore, the number of assessed tracklets and identities might differ from the overall numbers in the dataset. Table 4.2 presents the number of tracklets, and identities rejected by our method compared to the original numbers for both datasets.

In the given table, the number of original tracklets and original identities pertains to those that consist of at least one detection from the baseline model. It is worth noting that four tracklets within the MARS datasets did not contain any detections, resulting in a discrepancy of 8058 tracklets compared to 8062 tracklets in the original ground-

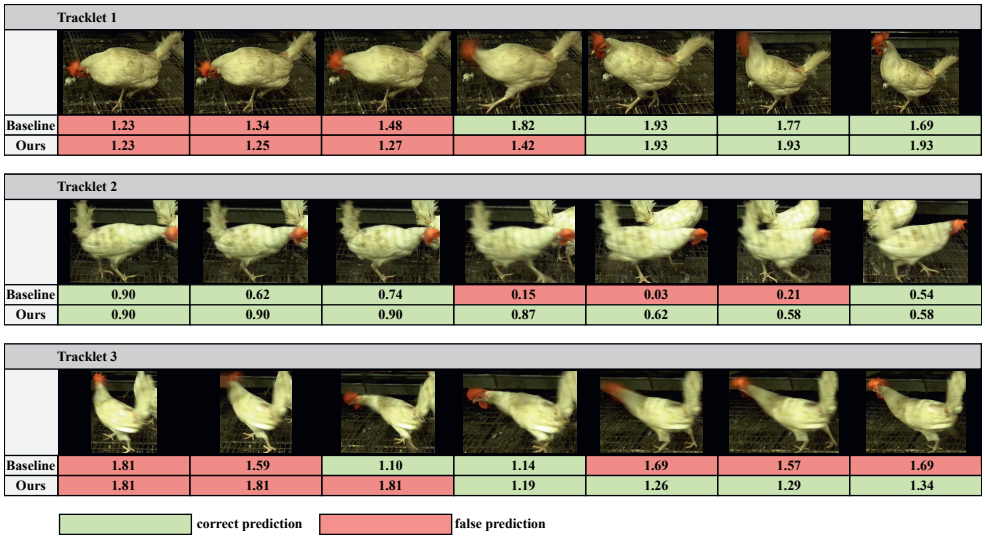


Figure 4.6: Instance-wise assessment score predictions of the standard ChickenNet (baseline) model compared to assessments provided by our method for consecutive instances from three tracklets of the chicken dataset. Colours indicate whether the predicted score was correct or not. A correct prediction in the final frame of a tracklet implies a correct assessment of the entire tracklet.

truth dataset. Results showed that with the defined uncertainty threshold of 0.25, our method provided assessments for all identities and tracklets assessed by the baseline model within the chicken dataset. In the MARS-Attributes dataset, 13.84% of the tracklets were rejected, while 99.84% of the identities were still assessed. Overall, it was shown that our method was able to increase the accuracy for tracklet- and identity assessment while proving valid assessments for almost all identities of the dataset. However, it was also demonstrated that the approach did not increase instance-level accuracies for all tested data. Results of further analyses, exploring the impacts of diverse data structures are presented in the following.

Table 4.2: Number of original tracklets and identities for the chicken and MARS-Attributes dataset, compared to the number of tracklets and identities rejected by our method.

	Chicken dataset	**MARS-Attributes dataset**
Original Tracklets	35	8058
Rejected Tracklets	0 (0%)	1116 (13.84%)
Original Identities	35	634
Rejected Identities	0 (0%)	1 (0.16%)

4.3.2 3.2 Effects of data quantity

In this experiment, we evaluated the performance of our method on a limited amount of data and compared it to the performance of the baseline model as well as simple averaging methods. Figure 4.7 visualizes the accuracies for the different evaluation approaches across a range of instances from 3 to 20 for the chicken dataset. Additionally, it shows the accuracies obtained by averaging all instance predictions of the baseline model per identity or per tracklet, considering the specific number of instances. Figure 4.8 illustrates those accuracies obtained for the MARS-Attributes dataset. It presents the accuracies for instances ranging from 3 to 20 per tracklet, but also for 1-100 tracklets per identity.

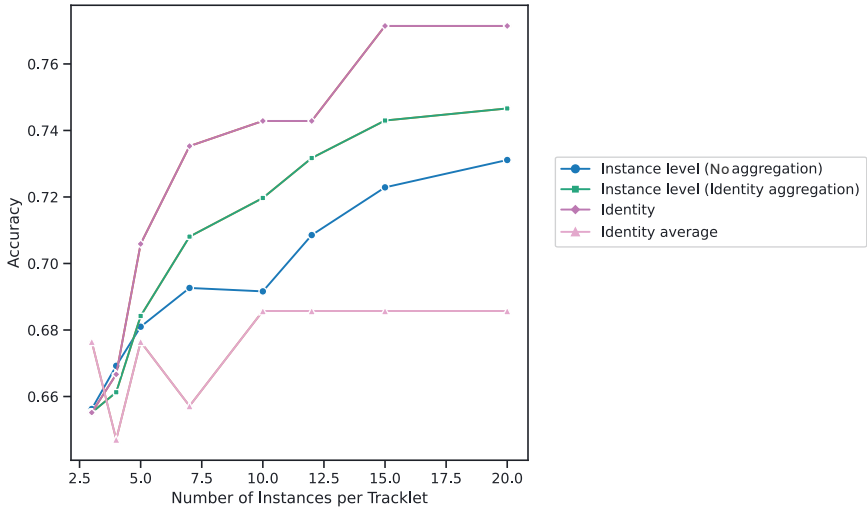


Figure 4.7: Assessment accuracies on the chicken dataset for varying numbers of instances per tracklet considered by our method. As the chicken dataset consists of a single tracklet per identity, corrections per tracklet and per identity are equal.

Results showed that all aggregation approaches based on our method outperformed the baseline model and averaging approaches for both datasets, even with a limited number of considered instances per identity. Solely for the case in which less than five instances were available for an entire identity of the chicken dataset, averaging of all instance predictions resulted in a higher accuracy. Further, experiments on the chicken dataset revealed an increase in the accuracies of our method with an increasing baseline accuracy, while the accuracy of the averaging approach remained constant beyond 10 considered instances. This indicates the correct selection of relevant instances from the total available instances. In contrast to the MARS-dataset, the chicken dataset includes a single tracklet per identity, thus number of instances per tracklet is equivalent to the total num-

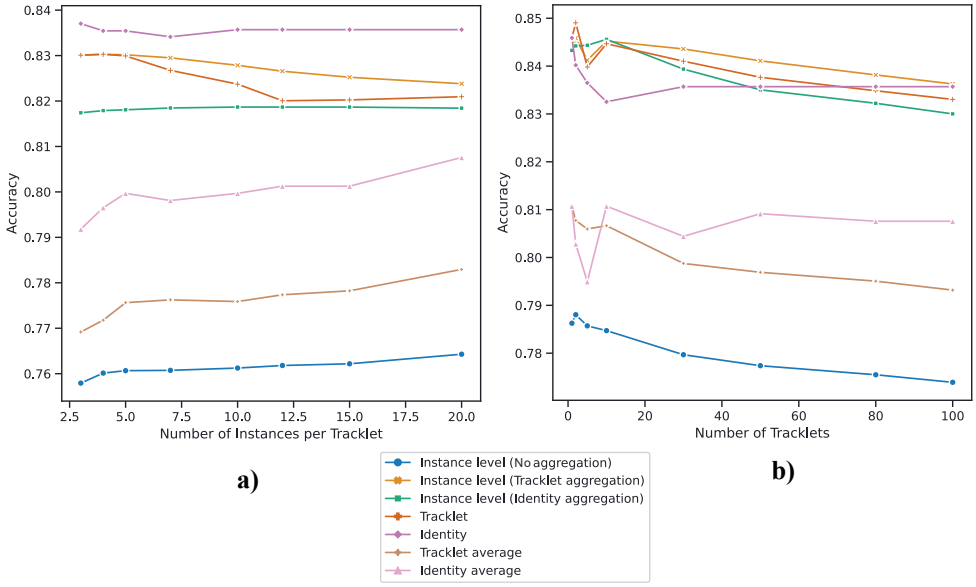


Figure 4.8: Assessment accuracies on the MARS-Attributes dataset for varying numbers of instances per tracklet (a) and tracklets per identity (b) considered by our method.

ber of instances per identity. This might explain the initial increase in accuracy of our method with and increasing number of instances per tracklet which was not observed for the MARS-Attributes dataset. For the MARS-Attributes dataset, the accuracies of the different aggregation approaches did not increase with an increasing number of instances per tracklet, instead tracklet-level accuracy and instance-level accuracy based on tracklet information slightly decreased while accuracies of identity-based aggregations did not significantly change. However, independent of the number of considered instances per tracklet, all accuracies obtained from our method were consistently 5-7% higher, compared to the baseline. Similar observations were made when comparing our method to traditional averaging. While averaging per identity and identity-level aggregation both yield a single prediction value per identity, the accuracies obtained from our method were 3-5% higher. For tracklets, the difference between averaging and tracklet-level aggregation ranged between 4% and 6%. This demonstrates the advantage of our uncertainty-based weighting and clustering approach compared to traditional averaging, also for a limited amount of data. While averaging approaches performed best for higher numbers of instances per tracklet, this dependency was not observed for our method. Increasing the number of considered tracklets per identity resulted in a decrease of the baseline accuracy for the MARS-Attributes dataset. This implies an increasing number of false predictions among the additionally considered tracklets. Thus, accuracies of tracklet-averages and

tracklet-based aggregation approaches also decreased. Averaging all predictions per identity as well as employing our method for a single prediction per identity led to an initial drop in accuracy but then, followed by a relatively stable accuracy throughout the analysis period. This observation deviated from our expectation that an increasing number of considered tracklets would increase the accuracy obtained by averaging all predictions of an identity. However, our expectation of an increased accuracy through our method was confirmed. Similar to the experiment on the number of instances per tracklet, accuracies based on our method were 5-7% higher than the baseline accuracy and about 2-4% higher than those obtained from averaging approaches. Further, it was shown that the difference in accuracy between identity-level aggregation and averaging per identity increased while the identity-level accuracy remained constant, and the averaging accuracy decreased. This implies that our method was able to prioritize the correct instance predictions and downgrade the false instance predictions among an identity. Moreover, while the influence of the baseline predictions on our method was evident, we found no clear difference in performance impact between limited instances per tracklet and limited tracklets per identity.

4.3.3 Alternative uncertainty quantification

Using epistemic uncertainty to weight individual instance predictions yielded similar results to using aleatoric uncertainty. Table 4.3 and 4.4 present the results of the experiments on the chicken dataset and the MARS-Attributes dataset.

Table 4.3: Accuracies and MSE for the assessment predictions obtained from our method using epistemic uncertainty to weight the individual predictions.

Aggregation Method	Chicken dataset		MARS-Attributes dataset	
	Accuracy (%)	**MSE**	**Accuracy (%)**	**MSE**
Instance level, Baseline (No aggregation)	85.40	0.18	76.42	0.20
Instance level, Aggregation per Tracklet	83.83	0.21	80.46	0.18
Instance level, Aggregation per Identity	83.83	0.21	82.05	0.16
Tracklet level	**88.57**	**0.13**	79.78	0.18
Identity level	**88.57**	**0.13**	**84.02**	**0.14**

In line with the results obtained using aleatoric uncertainty, we observed that our method was able to surpass the baseline model in terms of accuracy, also when utilizing epistemic uncertainty as a metric for weighting instance predictions. However, it was shown that

Table 4.4: Number of original tracklets and identities for the chicken and MARS-Attributes dataset, compared to the number of tracklets and identities rejected by our method based on epistemic uncertainty.

	Chicken dataset	MARS-Attributes dataset
Original Tracklets	35	8058
Rejected Tracklets	0 (0%)	347 (4.31%)
Original Identities	35	634
Rejected Identities	0 (0%)	2 (0.32%)

corrections on instance-level based on the estimated epistemic uncertainty led to a decreased accuracy for the chicken dataset. In combination with an increased accuracy on tracklet level, this implies that accurate assessments of a tracklet were primarily achieved in the later instances of that tracklet when using epistemic uncertainty. The accuracies obtained at the tracklet and identity levels were 88.57%, which was equivalent to those achieved using aleatoric uncertainty. However, the mean squared error was 0.13, slightly higher than the MSE of 0.10 obtained in the aleatoric approach. Furthermore, similarly to the experiments with aleatoric uncertainty, our method successfully assessed all 35 identities/tracklets in the chicken dataset without any rejections. Experiments on the MARS-Attributes dataset revealed a slightly higher accuracy at the identity level and a decrease in accuracy at the tracklet level when compared to the assessment based on aleatoric uncertainty. However, simultaneously, the number of rejected tracklets decreased from 1116 to 342 and the number of rejected identities increased from one to two, using an uncertainty threshold of 0.25. Utilizing epistemic uncertainty resulted in increased accuracy across all types of aggregation compared to the baseline model. The highest accuracy achieved was 84.02%, obtained at the identity level, surpassing the accuracy observed in the aleatoric uncertainty experiments.

4.4 Discussion

This study tackled the issue of obtaining reliable assessments from image sequences, originally intended for the assessment of chickens in challenging farm environments. However, it was shown that our approach is also applicable on alternative use cases focusing image sequences assessment. One addressed limitation, which most previously developed approaches faced, was the requirement for complete sequences during training of the model. Instead of developing an end-to-end trainable model, such as Chen et al. (2019) or Pei et al. (2017), our approach was designed to leverage standard models that operate on image level. Experiments demonstrated that the method was able to increase the assessment accuracy on sequences compared to such standard models. This improvement was observed not only for entire sequences but also for a limited number of instances within

a sequence and for a restricted number of sequences per identity. The second limitation that this study addressed was the lack of explainability in the predictions of models for sequence assessment. By considering the uncertainty of predictions on instance-level for the subsequent aggregation, we not only aimed to improve the assessment, but also focused the transparency of decisions. Similar strategies have been pursued by other approaches, such as (Morvant et al., 2014) and Tian et al. (2020a), which utilized uncertainty measures on multiple modalities for refining neural network predictions. However, our approach deviates in two key aspects. Firstly, instead of using multiple modalities, we applied this methodology specifically to image sequences and aggregated assessments of individual instances over time. Secondly, before fusing the individual, weighted assessments, we applied an appearance-based clustering approach. This enabled the consideration of different viewpoints for the assessment and thus allowed a prioritization of specific views.

4.4.1 Impact of chosen model components

The presented framework includes an assessment model, a feature encoder for appearance-based clustering, and an uncertainty metric to weigh individual predictions. These components are modular and can be replaced depending on the specific task, enabling the applicability of our method across multiple use cases and facilitating the extension of existing pre-trained assessment models. Thus, the choice of these individual modules significantly affects the performance of the overall method. In our experiments, we primarily focused on the application on chicken assessment which justified the utilization of the ChickenNet model. While this implementation was shown to be effective on other data such as the MARS-Attributes dataset, it is important to note that the assessment performance on image level could be further improved for this dataset by replacing ChickenNet with an alternative baseline model specifically tailored for the age estimation use case. Similarly, for the appearance-based clustering we employed an appearance descriptor obtained from a shallow CNN originally designed for representation learning in the context of person re-identification (Wojke and Bewley, 2018). However, depending on the data at hand, our method allows to replace it by an alternative feature descriptor, customized for distinguishing between different views, tailored for the particular application. Here, it is worth recognizing that the structure of the present data affects the appearance-based clustering. While for tracklets in which the individual detections differ a lot in terms of perspective or appearance, such as in the chicken dataset, our method resulted in a higher number of clusters. In contrast, a high similarity between the detections of a tracklet, as we observed it in the MARS-Attributes dataset often led to a single cluster per tracklet. In the latter case, our method comes down to uncertainty-weighted averaging. When examining the application of clustering at the identity level, it became apparent that the resulting clusters often align with the individual tracklets present in the MARS-Attributes dataset, as illustrated in Figure 4.3. However, although this correspondence may seem intuitive, it is not a necessary outcome. In our method, clustering serves the purpose

of differentiating instances that offer additional informative value. Despite tracklets typically being captured from different perspectives, it does not automatically imply that they provide complementary information that is relevant for the age estimation of the detected persons. For the quantification of uncertainty, we initially employed an estimation of aleatoric uncertainty given by ChickenNet to weight individual predictions. However, our experiments demonstrated a successful use of epistemic uncertainty as an alternative metric. Epistemic uncertainty estimation through Monte-Carlo dropout, as we modeled it in this study, further offers the opportunity to obtain an uncertainty estimation during inference. This allows the estimation of uncertainty on pre-trained models without the need for retraining the assessment model and makes it convenient to integrate existing standard models into our approach and leverage them for sequence assessment.

4.4.2 Aggregation methods and evaluation

Our method aggregates multiple detections obtained from a standard neural network for object detection aiming for reliable sequence assessment. However, it allows to vary the level on which predictions are fused into a final prediction, as explained in Section 4.2.4. In our experiments, we compared aggregations on tracklet and identity level resulting in a single prediction, but also instance-wise predictions obtained from aggregated information at each timestep within a sequence. While instance-level predictions offered a direct comparison to the conventional ChickenNet model, it is worth noting that in this case, the number and order of considered detections influences the assessment. For example, if relevant features crucial for the assessment are observed in the last frame of a tracklet, leading to a correct final assessment of that tracklet, the instance-level accuracy would be one divided by the number of instances, while the tracklet-level accuracy would be one. On the other hand, if those relevant features are revealed in an early frame, resulting in an early correct assessment, instance-level accuracy would be increased while maintaining the same tracklet-level accuracy. This effect became apparent when evaluating our method's performance on a varying number of instances on the chicken dataset and accounts for the differences in accuracy between tracklet level and corresponding instance level evaluations. The accuracy at the tracklet level was consistently higher, primarily due to tracklets for which the final prediction becomes correct after observing more than one instance. As more instances are considered, the number of false instance predictions increases. If all tracklets were to have their final predictions made after the first instance, tracklet-level and instance-level accuracy would be equal. Conversely, if the instance-level accuracy surpasses the tracklet-level accuracy, it indicates that the final tracklet prediction is incorrect while the individual instances of the tracklet are correctly assessed. For both datasets, as well as both tested uncertainty metrics, results showed that best predictions were obtained when evaluating on identity level. Identity level aggregation combines and clusters all available detections for an identity to obtain one final prediction, thereby eliminating the dependency on the detection order. This characteristic also applies to

evaluation on tracklet level and makes both evaluation approaches more meaningful for assessing the performance of our method even though they do not allow an instance-wise comparison to the baseline model.

4.4.3 Future research

One aspect for further investigations relates to the determination of thresholds for the instance-level prediction uncertainty. In this study, we established a static threshold to filter out assessments with an expected error exceeding 0.5. This choice was made due to our integer-labeled datasets, as this value corresponds to the maximum error that can still lead to a correct class-prediction. Nevertheless, employing a fixed threshold introduces an additional parameter that requires prior specification. This provides an opportunity for optimization, such as the integration of dynamic or learning-based approaches that adapt the threshold based on contextual information. Further work could also be dedicated to enhancing the efficiency of our method. Currently, all instances of a sequence are clustered each time a new instance is added, resulting in increased computational requirements as the sequence length grows. To address this issue, an alternative approach would involve limiting the number of considered instances. Finally, a fundamental aspect to address is the aggregation of individual predictions in real-life applications, where ground-truth information is unavailable. This requires the association of individual predictions within a sequence. While for single-instance recordings this might be accomplished through the detection model itself, scenarios involving multiple instances necessitate the incorporation of an additional tracking method to assign predictions to specific identities. Consequently, the selection of a robust association technique is crucial for the overall performance of the application.

4.5 Conclusions

In this study, we presented a novel approach for robust assessment from image sequences, specifically addressing animal monitoring under challenging environmental conditions. Our method focused the selective incorporation of information derived from multiple detections within an image sequence. To this end, it clusters the individual detections based on their appearance and accounts for uncertainty associated to the assessment of each detection. In our experiments, we primarily analyzed the assessment performance of our approach in comparison to the assessments made by conventional models operating on instance-level. Additionally, we explored the impact of limited data on our method's performance and evaluated alternative metrics for uncertainty estimation. Here, we distinguished between two dataset and three alternative aggregation levels to evaluate the assessment accuracy. Results showed that our method outperformed the baseline instance-level approaches on both datasets when aggregating information per tracklet or per identity. For the chicken dataset, it was able to increase the accuracy from 85.40% to

88.57% and for the MARS-Attributes dataset, an improvement from 76.42% to 83.57% was observed. Moreover, we demonstrated that the advantage against the instance-level approaches persists when considering a limited number of tracklets per identity and instances per tracklet. Similarly, the utilization of epistemic uncertainty as an alternative uncertainty metric also showed increased accuracies on both datasets. We conclude that the presented approach provides an effective method that enables the utilization of standard neural networks for the purpose of animal assessment from image sequences. In combination with an appropriate tracking approach, it becomes a versatile tool to be used in a wide range of real-world monitoring applications requiring robust assessments.

Chapter 5

Transformer-based similarity learning for re-identification of chickens

This chapter is based on:

Lamping, C., Kootstra, G., and Derks, M. (2024b). Transformer-based similarity learning for re-identification of chickens. *Submitted to Computers and Electronics in Agriculture*

Abstract

Continuous animal monitoring relies heavily on the ability to re-identify individuals over time, essential for both short-term tracking, such as video analysis, and long-term monitoring of animal conditions. Traditionally, livestock re-identification is approached using tags or sensors, which require additional handling effort and may potentially impact animal welfare. In response to these limitations, non-invasive vision-based approaches have emerged recently, with existing research primarily focusing on the re-identification of pigs and cows. Re-identification of chickens, which exhibit high uniformity and are housed in larger groups, remains challenging and has received less research attention. This study addresses this gap by exploring the feasibility of re-identifying individual laying hens within uncontrolled farm environments using images of their heads. It proposes the first similarity-learning approach based on a VisionTransformer architecture to re-identify chickens without requiring training images for each individual bird. In our experiments, we compared the transformer-based approach to traditional CNN architectures while assessing the impact of different model sizes and triplet mining strategies during training. Moreover, we evaluated practical applicability by analyzing the effects of the number of images per chicken and overall population size on re-identification accuracy. Finally, we examined which visual features of the chicken head were most relevant for re-identification. Results show Top-1 accuracies exceeding 80% for small groups and maintaining over 40% accuracy for a population of 100 chickens. Moreover, it was shown that the transformer-based architecture outperformed CNN models, with the use of semi-hard negative samples during training yielding the best results. Furthermore, it was revealed that the evaluated models learned to prioritize features such as the comb, wattles, and ear lobes, often aligning with human perception. These results demonstrate promising potential for re-identifying chickens even when recorded in an uncontrolled farm environment, providing a foundation for future applications in animal tracking and monitoring.

5.1 Introduction

Regularly monitoring health and welfare parameters is crucial for efficient and animal-friendly management of laying hens. Keeping track of the exterior, feed and water intake, and stress-related behavior, for example, provides important information about (the susceptibility to) disease of both the flock and the individual birds (Bryden et al., 2021; Tilbrook and Fisher, 2021; Michel et al., 2022). Currently, assessing individual animals is typically a manual and, therefore, labor-intensive task. Also, there is evidence that differences in farm management and farming systems can impact the quality of the assessment (Edwards and Hemsworth, 2021). Given the large number of animals on modern poultry farms, this can lead to a lack of individual care. Therefore, there is significant potential for automation of welfare assessments. Recent studies have proposed various approaches for the automated monitoring of welfare and behavior (Li et al., 2020b), such as the identification of activities (Yang et al., 2023) or assessment of plumage conditions (Lamping et al., 2022). To link multiple assessment measurements over an extended period with a particular hen, accurately re-identifying and tracking an individual is essential (Li et al., 2020b). This is relevant for welfare monitoring in livestock farming, but also within scientific research to investigate individual chicken characteristics and traits. Alongside individual assessment, re-identification further enables animal traceability throughout the entire value chain, which is of great interest for food production to provide certification of the quality and safety of the product.

Traditional methods for re-identification of individual chickens often include the attachment of external tags or sensors to the animal's body, especially for the purpose of behavior monitoring. For instance, Zhang Feiyang et al. (2016) utilized radio-frequency identification (RFID) tags to track individual chickens within a flock and collect long-term behavioral data. Another common identification system for chicken, but also for birds in general, is the use of leg- or wing-bands (Carroll et al., 2017). By using varying colors or numbers printed on the bands, they allow to re-identify individuals, also over a long-term period. However, it has been shown that body-worn markers may negatively affect the chicken's behavior or physical development. Consequences such as increased stress levels, feather pecking, or lighter animal weights were observed for the marked chickens (Dennis et al., 2008). In addition to the effects on animal welfare, tags and sensors require the attachment to the chicken's body a priori. This additional, often manual effort makes these solutions less relevant for large-scale applications, such as in commercial livestock farming.

In contrast to traditional methods, vision-based approaches allow non-intrusive identification of individuals without requiring additional on-body equipment or handling steps. This has led to the development of diverse vision-based re-identification methods for various animals. Research studies on this subject can be found in wildlife ecology, utilizing images captured by camera traps, as well as in large-scale animal farming. Early

vision-based methods usually focused on developing algorithms to detect and analyze predetermined biometric features on an animal, enabling its identification and discrimination from others. For example, unique coat patterns were shown to be discriminative markers in cattle (Andrew et al., 2016), but also in tigers, cheetahs, and whales (Schneider et al., 2019). Another biometric feature investigated in multiple studies, e.g. on cows and sheep, is the retinal pattern in the eye of an animal (Allen et al., 2008; Gonzales Barron et al., 2008). For cows, it has also been shown that the tailhead (Schilling et al., 2018) and muzzle (Gaber et al., 2016) provide unique prints that enable the identification of an individual animal. While those feature-based approaches require awareness of relevant markers and their specific characteristics, the rise of deep learning methods allowed to train algorithms to learn these features through labeled samples. In recent studies, this became the dominant method for identification of individual animals.

For deep learning based (re-)identification of animals, two main concepts can be distinguished. The first concept refers to a classification approach used to distinguish a known set of individuals. After training on multiple images per individual, the model is supposed to be able to recognize each of these individuals and assign the corresponding ID. For example, Freytag et al. (2016) utilized the AlexNet architecture (Krizhevsky et al., 2017) to identify individual chimpanzees from facial images. Similarly, Brust et al. (2017) trained a YOLO object detection model (Redmon et al., 2016) for the purpose of identifying Gorillas on camera trap images. In livestock farming, comparable classification approaches were developed for pigs (Hansen et al., 2018; Sihalath et al., 2021; Marsot et al., 2020) and for cows (Yang et al., 2019; Bhole et al., 2019). These studies indicated promising accuracies for the recognition of previously known individuals. However, this classification approach requires a fixed number of individuals in advance, making it impossible to add any without retraining the model. Furthermore, it is essential to acquire a substantial number of images for each individual to train the algorithm. These aspects make the classification approach impractical for large-scale applications in livestock farming, where collecting a sufficient number of training images for every animal is often infeasible. Instead, a model should be able to recognize an individual without being trained on that specific individual.

This is where the alternative concept of deep learning-based re-identification approaches, similarity learning, becomes relevant. Rather than assigning a specific ID to an image, similarity learning aims to predict whether two input images are similar or dissimilar (Bromley et al., 1993). Here, the key is to train a model to learn feature representations from image data that can be used to distinguish individuals from each other. The aim is to learn these representations in such a way that those belonging to the same individual are closer than those of different individuals, while being robust to external effects such as changes in poses or illumination. For animal re-identification, learning representations enables the recognition of individuals without training the model specifically on them. Consequently, new individuals can be added whenever needed. After training, similarity

networks only require one known image of an animal to predict whether another image depicts the same individual. Similarity learning networks are a proven tool, utilized in various recent studies on animal re-identification. For example, as one of the first, Deb et al. Deb et al. (2018) implemented a face recognition approach for primates, which was evaluated on datasets of golden monkeys, lemurs, and chimpanzees. Later studies utilized similarity learning for the re-identification of other species in wildlife ecology, such as dolphins (Bouma et al., 2018), manta rays (Moskvyak et al., 2021), and ring seals (Nepovinnykh et al., 2020, 2024).

In livestock farming, approaches have been developed on face-recognition of pigs, but also on re-identification of cattle (Andrew et al., 2021; Wang et al., 2023). Despite the substantial progress made in deep learning-based similarity learning, the poultry sector has received limited attention in studies on re-identification. Challenges such as the uniformity of animals resulting from breeding practices, as well as the high number of animals within a farm, make this task particularly complex. To our knowledge, the only existing study addressing poultry re-identification is presented in Corkery et al. (2009). Their approach presented a preliminary investigation of the avian comb as a potential marker for re-identification using images recorded in a controlled environment. As it exclusively focused on the avian comb, the study employed an analysis of predefined features rather than a learning-based methodology. In uncontrolled environments, lacking a standardized orientation of the comb, this strategy will be susceptible to errors. To this end, the approach developed in our study addresses the question of whether it is possible to re-identify individual laying hens within a realistic, uncontrolled farm environment. To tackle the significant challenges associated with this task, our method makes use of state-of-the-art deep learning techniques.

Current approaches in animal re-identification using deep learning typically employ traditional convolutional network architectures to learn feature representations to distinguish individuals. Although CNNs performed well among various re-identification applications, their reliance on local receptive fields limits them to capturing information within small spatial regions without properly considering the context (Luo et al., 2016). Thus, the convolutional concept struggles to capture relationships between spatially distant features. This characteristic could limit the task of chicken re-identification, which may benefit from extracting multiple discriminative parts beyond local regions, such as the comb and beak of the animal. Similar findings have also been reported in person re-identification (He et al., 2021a). Furthermore, due to the presence of down-sampling operators, leading to a reduction in the spatial resolution of feature maps, CNNs have shown weaknesses in distinguishing between similar objects that only differentiate in fine-grained details (He et al., 2021a). Such fine-grained details are typically relevant for the task of chicken re-identification. To better deal with this, transformer-based models have recently been proposed as an alternative to CNNs in general object re-identification. Traditionally developed to process sequential data in natural language processing (Vaswani et al., 2017),

transformer architectures are mainly based on the self-attention mechanism, which allows for weighting the importance of different elements within an input sequence. By dividing images into non-overlapping patches and processing the resulting sequence, the work introduced in Dosovitskiy et al. (2020) extended the application of transformer architectures to vision tasks. Approaches, such as those presented in He et al. (2021a) employ transformers on human data to address the previously mentioned issues of CNNs in object re-identification. Rather than solely capturing information from small local regions, the transformer architecture enables a model to learn global concepts. Moreover, the exclusion of down-sampling operations contributes to the preservation of fine-grained features.

This motivates the selection of a transformer-based model for representation learning in the present study, aiming to re-identify individual chickens within a farm environment. In conclusion, our contributions are as follows:

- We present a fine-grained dataset for laying hen re-identification including 3644 recorded video sequences of hens with 18,819 images.
- We present the first learning-based approach for chicken re-identification from images and evaluate it under the conditions of an uncontrolled farm environment.
- We assess a transformer-based approach for similarity learning in comparison to traditional CNN architectures and analyze the influence of model size, triplet mining strategy, and different environmental aspects.
- We investigate the relevance of different parts of the chicken's head for visual re-identification.

5.2 Material and methods

The following subsections present the materials and methods used in this study. Data collection and dataset generation are described in Section 5.2.1 and Section 5.2.2, respectively. Subsequently, Section 5.2.3 delves into the architectural details of the re-identification model and its training procedure. Finally, Section 5.2.4 details the experimental setup.

5.2.1 Data collection

Data collection was conducted with the aim to obtain images of individual chickens in a realistic, uncontrolled environment to address the real-world conditions of potential re-identification applications. To this end, images were recorded in a free-range barn with a flock of 18,000 white laying hens in Garrel, Germany. This housing type allows chickens to freely move around within and outside of the barn. Thus, the recorded animals were not manually selected but instead randomly appeared in front of the recording system.

For large-scale image retrieval, a recording setup was installed to automatically detect and capture passing chickens. This setup comprised an Intel RealSense D405 camera connected to an NVIDIA Jetson Nano board for image processing and storage. The camera and computing unit were housed in a protective case, and two additional LED strips were attached for proper illumination. The recording setup was positioned at a height of 1.8 meters in the aviary, with a distance of 90 cm to a perch. From this perch, chickens were able to reach an adjacent drinker. With this setup, we aimed to ensure comparable poses and orientations of the captured chickens (Figure 5.1). The recording process was initiated when a chicken appeared within a range of 50-100 cm in front of the camera, as measured by a LiDAR distance sensor. Once started, images were captured for 20 seconds at a frame rate of one frame per second. This low frame rated was intended to prevent a high number of very similar images within a recording. Each recorded sequence was stored separately and annotated with a specific timestamp indicating its capture time to enable later differentiation of individual chickens. Images were saved in a resolution of 1280 x 720 pixels.

To increase the variety of images in the dataset and minimize duplicate captures of individual chickens, the data collection was conducted in three different stages. The first round of recordings took place in March and April 2023 for three consecutive weeks, from 2 pm to 7 pm every day. The second round of images was collected in July 2023 for three days, spanning from 10 am to 7 pm. The final recordings were made in October 2023 over two days, also from 10 am to 7 pm. This data collection resulted in a total of 4959 recorded sequences, comprising 99,180 images.

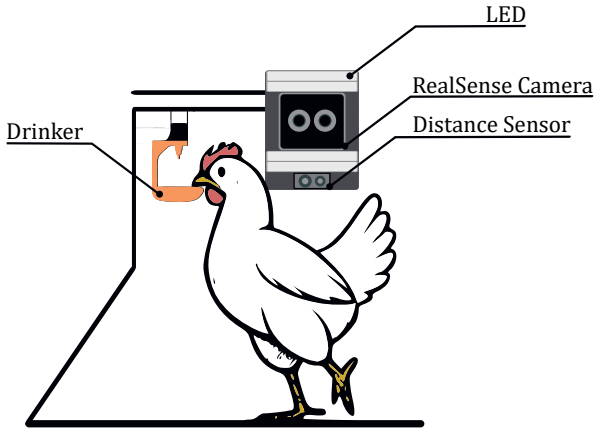


Figure 5.1: Schematic diagram of the image acquisition setup with an Intel RealSense D405 camera and additional LED illumination. Recordings were started when a chicken was detected by the distance sensor.

5.2.2 Dataset generation

In preparation for the creation of a chicken re-identification dataset, the raw image data underwent preprocessing to exclude irrelevant images. This involved several preprocessing steps (Figure 5.2). In the first step, we focused on selecting images in which the heads of the chickens were fully visible. Compared to the rest of the body, the head region of a chicken provides several visual features, such as the avian comb, which was indicated as potential biometric marker in a previous study (Corkery et al., 2009). This motivated our focus on the head region in the present dataset. To ensure the consistent inclusion of this area in all images, we employed a feature detection approach for the detection of the heads. Our approach involved annotating a subset of 500 images with bounding boxes for the head, eye, and beak of each chicken. This annotated data was then used to train a YOLOv5 object detection model (Jocher et al., 2022) to identify these body parts in all images in the dataset. Considering the high number of available images, we prioritized high precision over a high recall rate for the detection algorithm, which led to the implementation of a minimum confidence threshold of 0.95 for each detection. By setting this high threshold, we ensured that only images where the model exhibits high confidence in the presence of specific body parts are included. As a result, this reduces the number of images passing this selective filter.

Prior studies have shown that re-identification systems can be significantly improved by ensuring consistent image features between two images of the same individual (Ghosh et al., 2023). This prevents a model from being forced to assign two highly different images to a certain individual even though they do not share any visual features. In the context of chickens, an example of the latter scenario would be capturing two images of the same chicken's head, each taken from opposite sides. Even though both images capture the same animal, they do not share any visual features due to their differing viewpoints. Thus, training an algorithm for similarity learning using these diverse images offers no benefit. Instead of applying a feature matching approach during training as proposed in Ghosh et al. (2023), we used the YOLOv5 body-part detection model to reject irrelevant images. These were discarded from the dataset according to the following criteria:

1. An image was discarded if either the number of detected heads, eyes, or beaks per image was not equal to one. This was intended to avoid images featuring multiple chickens, which would require matching chickens across different images in a sequence. Moreover, this step excluded chickens with occluded faces.

2. An image was discarded if the x-position of the beak was smaller than the x-position of the eye. This ensured that all chicken heads within the dataset were captured from the right side.

If both criteria were met, the union area of all three predicted bounding boxes for the head, eye, and beak was computed. The image was then cropped to the outer coordinates

of this area to only include the head region of each chicken. Figure 5.2 illustrates the preprocessing steps for an accepted image and a rejected image from the dataset.

Following this procedure for all recorded sequences, the final dataset comprised 3,644 valid sequences, including 18,819 images of cropped chicken heads. Within the sequences, the number of images per sequence varied between 1 and 20. From the dataset, we selected 90% of the sequences for training and 10% for testing.

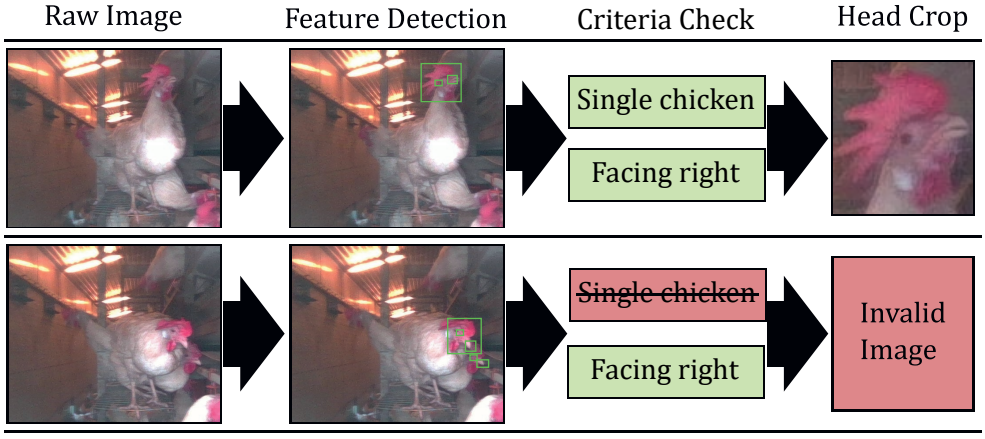


Figure 5.2: Image preprocessing steps. The bottom example illustrates an invalid image that was rejected due to the presence of more than one chicken in the image.

5.2.3 VisionTransformer for animal re-identification

Following the earlier introduced animal re-identification methods from other domains, in this study we propose a neural network for similarity learning to be trained on the generated dataset. Considering the use case in poultry, we selected similarity learning over a classification approach, as similarity learning does not require multiple recordings for each individual chicken during training. The following sections present the model architecture as well as the details of the training process.

5.2.3.1 Model architecture

For similarity learning, we employed a model architecture designed to extract representations from chicken face images by mapping each input image into a 128-dimensional feature space, as illustrated in Figure 5.3. This feature representation aims to encode discriminative information of an individual while being invariant to changes within the same individual. Additionally, a substantial dimensionality reduction is achieved. Within the feature space, the distances between representations are intended to function as a direct indicator of input similarity, bringing representations corresponding to the same

chicken closer together, and those of different chickens further apart. By employing this approach, re-identification becomes a task of finding the closest representation for each image of a particular chicken.

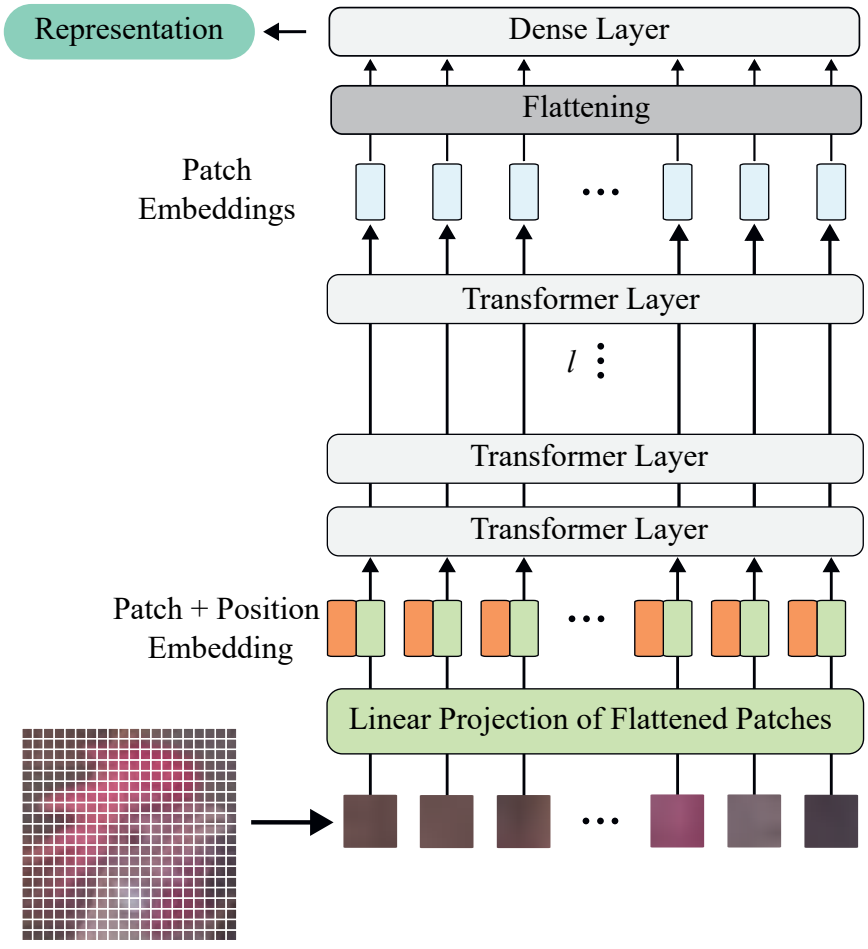


Figure 5.3: Architecture of the similarity learning network. The input image is processed in multiple patches, which are mapped into an embeddings space and augmented with positional embeddings. All embeddings are then encoded by the transformer encoder, finally resulting in a 128-dimensional feature representation.

To obtain a discriminative representation from each image, our model architecture follows the principles of the VisionTransformer architecture, as presented in Dosovitskiy et al. (2020). As input, the model receives an image, defined by its height and width of 64 x 64 pixels and three color channels. First, the input is preprocessed by dividing it into a sequence of non-overlapping patches. Considering the small image size, we chose a patch

size of 3. Each of the resulting patches is flattened, leading to a size of $3 * 3 * 3 = 27$ for each, considering the three color channels. The flattened patches are then mapped into a v-dimensional embedding space, using a single dense layer to train a linear projection for each patch. Additionally, each patch embedding is augmented with a learnable position embedding to retain its positional information (Dosovitskiy et al., 2020). The resulting sequence of embeddings serves as the input to the transformer encoder. This encoder consists of l transformer layers containing multi-head self-attention (MSA) and multi-layer perceptron (MLP) blocks as introduced in Vaswani et al. (2017). Each multi-head self-attention block includes multiple heads, which enable attending to different parts of the image. The outputs of the transformer encoder are subsequently passed to the final prediction layer. In contrast to the original implementation of the VisionTransformer, this approach does not employ an extra embedding for image representation and instead utilizes all image-patch embeddings. The additional embedding was originally introduced to align with the architecture of the transformer for text, but has previously been shown to not improve the performance of a VisionTransformer, while introducing additional parameters (Dosovitskiy et al., 2020). In our model, the encoder outputs are flattened and then passed to a final dense layer to obtain the 128-dimensional feature representation. To make those representations comparable for a subsequent nearest neighbor search, the outputs are L2-normalized, making them unit vectors.

With the same basic architecture of the similarity-learning network, we varied the number of attention heads h and transformer layers l as well as the embedding dimension v during the experiments to compare networks of different complexity and size. Specifically, we created three variations of the described network architecture: small, medium, and large, as outlined in Table 5.1.

Table 5.1: Configurations of the three different VisionTransformer models

Model	**VisionTransformer S**	**VisionTransformer M**	**VisionTransformer L**
Transformer layers l	2	4	4
Attention heads h	2	4	8
Embedding dimensions v	64	128	256
Trainable parameters	2,217,024	5,554,688	27,407,232

The proposed models employ the so-called triplet loss function (Schroff et al., 2015) to learn meaningful representations that can discriminate between individual chickens while

being robust to variations such as pose or lighting changes. A triplet comprises an anchor image a, a positive example p that shares a common ID with the anchor, and a negative example n with an ID different from the anchor. For each of the samples, representations are obtained by passing them through the network, while the loss of the network is determined by the triplet loss function. Intuitively, the triplet loss function encourages the model to learn representations where the distance $d_{a,p}$ between the representation of an anchor image y_a and a representation of a positive sample y_p is smaller than the distance $d_{a,n}$ between y_a and the representation of a negative sample y_n by at least a certain margin. In this study, we adapted the formulation of the triplet loss as defined in Hermans et al. (2017), in which this margin is not explicitly given as a fixed value. Instead, it dynamically adapts based on the similarity between the positive and negative samples, as outlined in Hermans et al. (2017). This removes the additional hyperparameter m and results in the following loss function:

$$L_{Triplet} = \sum_{\{a,p,n\} \in T} \log 1 + e^{D_{a,p} - D_{a,n}} \quad \text{and} \quad T = \{\{a_1, p_1, n_1\}, ..., \{a_t, p_t, n_t\}\} \tag{5.1}$$

The distance d between two r-dimensional representations was computed using the Euclidean distance, as in Hermans et al. (2017), it was demonstrated to be more stable during optimization compared to the squared Euclidean distance. Consequently, it was consistently applied in all experiments in this paper and is defined as $d_{y_1,y_2} = ||y_{1,i} - y_{2,i}||$. Later during the evaluation, this distance was also employed to perform nearest-neighbor search to find the most similar representations for a particular sample.

5.2.3.2 Triplet mining strategies

Triplet loss for similarity learning requires the creation of image triplets to train the model. The selection of triplets during training significantly impacts both the training efficiency and the re-identification model's performance (Schroff et al., 2015). To form a set of triplets T, various strategies can be employed. One straightforward approach is to generate all possible triplets by combining all available images from the entire training set. However, this results in an excessively large number of triplets, growing cubically with a larger dataset. This leads to significant memory requirements, increased processing time, and unequal distribution between positive and negative samples. Additionally, most of the created training samples are trivial, providing minimal contribution to the learning process. This includes negative pairs with low similarity and positive pairs with high similarity.

Therefore, the alternative, which was also used in this paper, is the online generation of triplets during training, where triplets are created from the samples within a training batch. This aims to only utilize samples that provide relevant insights while reducing the

computational effort. Following the approach of Hermans et al. (2017), in each training step, s random identities, corresponding to sequences in our case, were sampled from the dataset. Then, for each of these identities, k images were randomly chosen from the sequence, resulting in a batch of sk images. To construct the most valuable triplets from this batch, we compared three different mining strategies:

Hard Negative Mining
This strategy utilizes a hard negative sample for each anchor, which is defined by the image with the most similar representation from the batch that has a different label. Within the batch, triplets are then generated by combining each possible anchor-positive pair with the hard negative sample of the respective anchor. This approach has been shown to significantly improve performance and speed up the convergence of the training process compared to the combination of all possible triplets (Schroff et al., 2015; Hermans et al., 2017).

Semi-Hard Negative Mining
Semi-Hard Negative Mining is a variation of the Hard Negative Mining strategy, proposed by Schroff et al. (2015). Instead of selecting the most similar negative sample for each anchor, this approach utilizes the most similar negative sample, which is less similar than the corresponding positive sample. These negatives are referred to as semi-hard negatives.

Semi-Hard Negative + Easy Positive Mining
In previous research it was shown that the exclusive use of hard positives, meaning the least similar positive samples, did not increase the model performance (Xuan et al., 2020). However, it was also demonstrated that using easy positives instead enhanced performance. Easy positives refer to the positive samples that have the highest similarity to the corresponding anchor image. We employed this approach in line with Semi-Hard Negative Mining, but instead of using possible anchor-positive pairs, we only selected the easiest positive sample for each anchor.

5.2.3.3 Implementation details

In this study, batches for training were created by randomly sampling $s = 16$ sequences from the training dataset. This was done while considering the timestamps of each folder, ensuring a minimum time difference of at least 15 minutes between each selected sequence to prevent potential false negative samples, which occur when sequences are labeled differently despite showing the same chicken. Such false negatives could theoretically occur if a chicken was sitting on the perch for an extended period, resulting in more than one recording of a certain chicken. By implementing a minimum time difference, the risk of sampling such consecutive recordings from the dataset was excluded. For each sampled sequence, $k = 4$ images were randomly selected. This value was chosen to be smaller than the dataset's mean image count of approximately five because we aimed to sample differ-

ent sets of images from a sequence in each sampling iteration. In case that a sequence from the training set contained fewer than four images, the remaining images were duplicated until the required number was reached. Combined with the data augmentation methods detailed below, this duplication generated additional positive training samples that were different from the anchor images. This approach allowed us to fully exploit the collected data and ensured that recordings with less than four images were not ignored during training. Consequently, the size of each training batch was 64, resulting from 16 sequences with four images per sequences. Labels were assigned to the images online during training, using integer numbers from 1 to 16 to indicate the identity of each image in the batch.

The configurations for both model architecture and triplet mining were implemented using the TensorFlow 2 framework. During training, images were first rescaled to a fixed sized of 64 x 64 pixels before being passed to the network. To augment the diversity of the training dataset and enhance the model's invariance to image perturbation, a series of random image augmentations were applied during training. Specifically, these augmentations included random rotations within the range of -45 to 45 degrees, zooming with a variable factor between 0.8 and 1.2, and adjusting brightness with a factor ranging from 0.9 to 1.2. These augmentations were selected to prevent the model from learning to identify animal poses or image-related characteristics as potential markers for chicken re-identification. Optimization of the model weights was performed over 15,000 training iterations using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001 which was reduced by a factor of 0.95 every 1000 iterations.

5.2.4 Experiments

The four experiments, which are detailed in the following subsections, were primarily conducted to evaluate whether individual chickens can be re-identified in an uncontrolled environment through the application of deep learning. To this end, experiment 1 evaluated the re-identification performance of the proposed transformer network architectures in comparison to two convolutional neural network architectures. Additionally, in experiment 2, we analyzed the effects of the different triplet mining strategies for each architecture. Experiment 3 then shifted the focus towards the applicability of the presented method and evaluated the impact of varying the number of samples per chicken and the overall population size on the re-identification performance. This aimed to investigate how well the method handles situations with different numbers of chickens and images per chicken, a crucial aspect for practical applications. Finally, experiment 4 analyzed the relevance of certain characteristics of the chicken's head for re-identification within our approach.

All experiments were evaluated on the test dataset using the Top-1 and Top-5 accuracy metrics. For this, we processed the entire test data in multiple batches, each consisting

of $s = 32$ sequences with $k = 4$ images per sequence. This configuration balanced the challenge of the re-identification task by associating each chicken image with three positive samples and 124 negative samples. The impact of these numbers on performance was explored in a dedicated investigation within experiment 3. To ensure comparability, all evaluations used the same sampled sequences and images. Sequences containing fewer than k images were excluded. Notably, when selecting the positive samples, we aimed to ensure that the chicken's identity was the only consistent element among the images to avoid consideration of image-, pose-, or background-related characteristics. Therefore, despite closely cropping the chicken's head as discussed in Section 5.2.2, we also excluded consecutive frames within a sequence when choosing s samples per chicken for testing. This aims to avoid selecting positive samples that are highly similar to the anchor. Exemplary anchor images and their corresponding positive samples are shown in Figure 5.4.

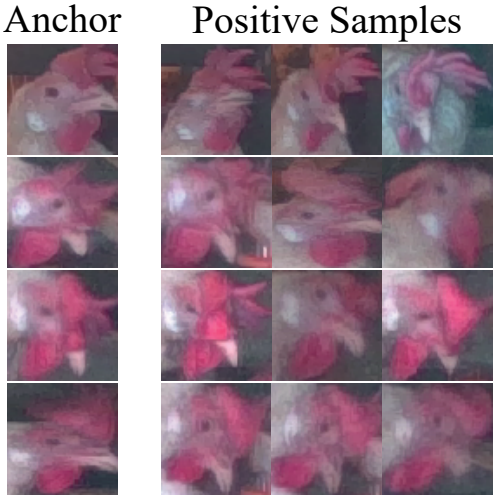


Figure 5.4: Exemplary anchor images and their corresponding positive samples from the test dataset.

After sampling the batches from the test dataset, they were processed by the different models, generating 128-dimensional representations for each image. Subsequently, the pairwise distance between each pair of representations was computed to obtain a distance matrix containing sk^2 values. Each representation was once utilized as an anchor, and the closest representation from the remaining test batch was determined. For the Top-1 accuracy, a prediction was considered correct if the label corresponding to the closest representation matched the anchor label. Similarly, the five closest representations were determined for the Top-5 accuracy, where a prediction was considered correct if the anchor label was among these representations.

5.2.5 Experiments 1 and 2 - Effects of network architecture and triplet mining strategies

Although most of the existing work on animal re-identification utilizes convolutional neural networks to learn similarity between images, we hypothesized that transformer-based architectures are more suitable for the task of chicken re-identification compared to the traditional architectures.

Therefore, in experiment 1, we evaluated the three variations of our transformer-based model in comparison to two variations from the EfficientNet family as presented in Tan and Le (2019). Specifically, we used the smallest available model, EfficientNetB0, with approximately four million trainable parameters, and the larger model EfficientNetB5. The latter encompasses around 28 million parameters, making it comparable to the largest version of our transformer model in terms of the number of parameters. The architectures of these models largely remained unchanged. However, analogous to the transformer-based approach presented in 5.2.3.1, the CNNs needed to be adapted for representation learning. That required replacing the final model layer with a single fully connected layer to extract a 128-dimensional representation vector from each input image. Each of the compared models was then optimized using the training parameters outlined in 5.2.3.3.

In line with the architecture comparison in the first experiment, experiment 2 assessed the effectiveness of different triplet mining strategies employed during training for each model. These strategies included Hard-Negative Mining, Semi-Hard Negative Mining, and Semi-Hard Negative + Easy Positive Mining, as detailed in Section 5.2.3.2.

5.2.6 Experiment 3 - Effects of sample quantity

Considering the application of re-identification systems in poultry, it is especially relevant to evaluate whether a certain chicken can be found within a population of multiple chickens. This matching challenge is basically affected by two variables. First, the number of distinct chickens within the population being searched, and second, the number of images available for each chicken. Intuitively, increasing the number of chickens in the population complicates the challenge, while increasing the number of images per chicken simplifies it. To this end, we evaluated our approach using both numerical values. The number of distinct chickens s, corresponding to individual sequences in our case, was varied between 2 and 100, while the number of images per chicken k was varied between two and five during the batch sampling from the test dataset. In this experiment, the maximum number of images per chicken was limited to five to ensure a substantial count of sequences with at least that number of images. Then, for each chicken and each image, the five most similar images were determined. In this experiment, we focused on our proposed VisionTransformer L architecture in combination with the Semi-Hard Negative Mining strategy. Top-1 and Top-5 accuracy metrics of the different configurations were evaluated following the approach explained in Section 5.2.5.

5.2.7 Experiment 4 - Analysis of discriminative features (Grad-CAM)

Previous studies investigated various animal body parts as potential biometric markers for the purpose of re-identification. An example from the poultry sector is the avian comb, for which a preliminary trial indicated the feasibility of re-identification in a controlled experimental setting. Our approach, which concentrates on the entire head region of a chicken, allows us to consider multiple features within this area. To analyze which parts of the image are most relevant for the tested approaches, we applied gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017), which visualizes the parts of an image that are important for the prediction. In the context of our method, this prediction corresponds to an embedding trained to identify individual chickens, highlighting features that indicate characteristics contributing to the uniqueness of each chicken. This provides a qualitative evaluation of the re-identification, complementing the previously conducted quantitative evaluation.

Following the approach of the first two experiments, we compared the different model architectures and triplet mining strategies and applied Grad-CAM to each of them. This aimed to understand differences and similarities in the learned features across the approaches but also aimed to compare the features identified as relevant by the models with those considered relevant from the perspective of a human observer. The original Grad-CAM approach was implemented on CNN architectures and utilizes the gradient of the networks output with respect to the feature maps of a particular convolutional layer. Based on this gradient, the importance of each feature map for the final output is determined to generate a heat map indicating important regions within the input image. In this experiment, we chose the last convolutional layer for the EfficientNet architectures. As pure transformer-based architectures do not have any convolutional layers, in these models, we used the last layer of the transformer encoder to obtain the gradients.

5.3 Results

The results are presented in the order of the experiments. First, the evaluation of the re-identification performance among the different network architectures and triplet mining strategies is shown. Following that, the effects of sample numbers and population size, as well as the results of the Grad-CAM analyses, are the focus.

5.3.1 Network architectures and triplet mining strategies

The comparison of the different network architectures and triplet mining strategies revealed some significant differences among the evaluated approaches. As shown in Table 5.2, the best performance was observed for the VisionTransformer L architecture in combination with semi-hard negative triplet mining. This configuration yielded a Top-1 accuracy of 0.76 and a Top-5 accuracy of 0.92. Intuitively, these metrics denote that

among all test batches, each consisting of 32 chickens with four images, 76% of the images were correctly matched while the correct match was within the Top-5 results in 92% of cases. Figure 5.5 illustrates exemplary sample images and their corresponding closest matches for this configuration. Conversely, the lowest accuracy was obtained for the EfficientNetB5 model in combination with hard negative mining, resulting in a Top-1 accuracy of 0.25 and a Top-5 accuracy of 0.46.

Table 5.2: Top-1 and Top-5 accuracies for the evaluated models and triplet-mining configurations

Model	# Parameters	Triplet-Mining strategy	Top-1 accuracy	Top-5 accuracy
VisionTransformer S	2,217,024	HardNegative	0.37	0.57
VisionTransformer S	2,217,024	SemiHardNegative	0.57	0.78
VisionTransformer S	2,217,024	SemiHardNegative + Easy Positive	0.58	0.76
VisionTransformer M	5,554,688	HardNegative	0.39	0.61
VisionTransformer M	5,554,688	SemiHardNegative	0.70	0.86
VisionTransformer M	5,554,688	SemiHardNegative + Easy Positive	0.69	0.87
VisionTransformer L	27,407,232	HardNegative	0.39	0.61
VisionTransformer L	**27,407,232**	**SemiHardNegative**	**0.76**	**0.92**
VisionTransformer L	27,407,232	SemiHardNegative + Easy Positive	0.76	0.89
EfficientNetB0	4,171,516	HardNegative	0.47	0.72
EfficientNetB0	4,171,516	SemiHardNegative	0.58	0.80
EfficientNetB0	4,171,516	SemiHardNegative + Easy Positive	0.55	0.80
EfficientNetB5	28,603,056	HardNegative	0.25	0.46
EfficientNetB5	28,603,056	SemiHardNegative	0.60	0.81
EfficientNetB5	28,603,056	SemiHardNegative + Easy Positive	0.45	0.77

The experiment further showed that the triplet mining strategy affected the re-identification performance of the models. Across the different model architectures and sizes, hard negative triplet mining consistently yielded the lowest accuracies, with none of the approaches surpassing a Top-1 accuracy of 0.47. In contrast, the best results were achieved through the implementation of semi-hard negative mining, either using all pos-

Figure 5.5: Exemplary results of the VisionTransformer L model in combination with Semi-Hard Negative Mining. For each anchor, the five samples with the most similar representation to the anchor are ranked in the columns. Samples with the same ID as the anchors are framed in green (correct match), samples with a different ID are framed in red (false match).

itive samples or only easy positives. Between those two alternatives, no clear advantage was observable. Among the five assessed models, the use of easy positives yielded better Top-1 results for two models, whereas employing all positives achieved better performance in two other scenarios. In one case, equal performance between the two strategies was observed.

Regarding the different model architectures, it was noticeable that the transformer-based approaches outperformed the CNN models, irrespective of the model size. While the number of parameters was comparable between EfficientNetB5 and the largest VisionTransformer, as well as between EfficientNetB0 and VisionTransformer M, the VisionTransformers demonstrated higher accuracies compared to their respective CNN counterparts.

In general, this experiment indicated that the performance of the transformer-based models improved with increasing model size, while this trend was not evident for the tested CNN-based approaches, Accuracies for both CNN models were comparable, with one model occasionally outperforming the other, depending on the chosen triplet mining strategy. EfficientNetB5, with about 28.6 Mio trainable parameters, achieved the highest performance among the CNNs with a Top-1 accuracy of 0.60 while even the smallest transformer with less than a tenth of the parameter count of EfficientNetB5 was able to achieve a Top-1 accuracy of 0.58.

5.3.2 Varying sample numbers and population sizes

Figure 5.6 illustrates the Top-1 re-identification accuracies obtained by varying the number of chickens within the population being searched and the number of images per chicken. Similarly, Figure 5.7 shows the Top-5 accuracies for the different configurations.

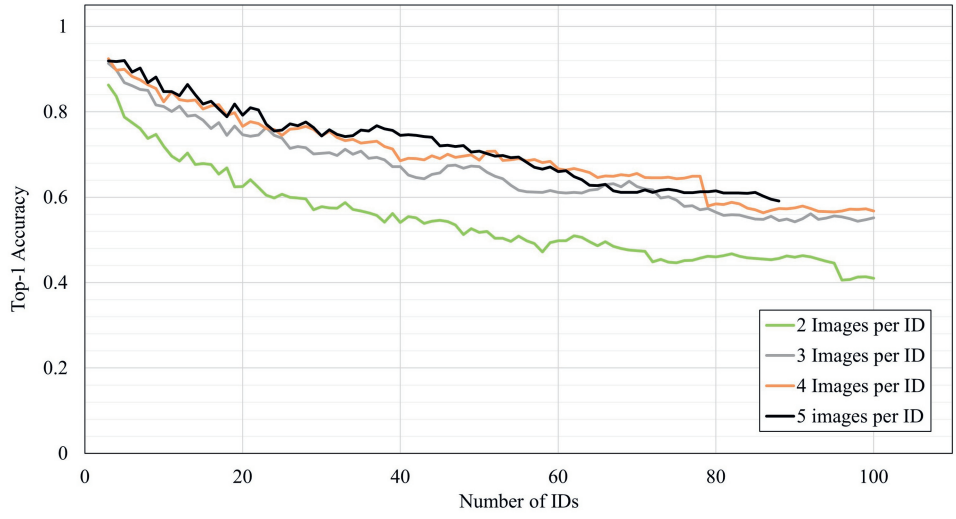


Figure 5.6: Top-1 re-identification accuracies for varying numbers of distinct chickens and images per chicken. Results were obtained by averaging over multiple sets of 2-100 IDs.

The experiment yielded the highest accuracies for a population of two chickens, each represented by five images. In this scenario, 98% of the images were correctly matched (Top-1 accuracy) with a Top-5 accuracy of 1. As expected, accuracy decreased with both an increasing number of chickens and a decreasing number of images per chicken. This can be attributed to the growing challenge of a re-identification task when there are more chickens but fewer images available per chicken.

The most challenging configuration involved 100 chickens and only two images per chicken ($k = 100$, $s = 2$). In this setting, where only one matching image existed among 199 searched images, a Top-1 accuracy of 0.41 was achieved. This is significantly higher than the probability of correct matching by chance, which is approximately 0.005. In this setting, a Top-5 accuracy of 0.64 was obtained, which also marks the lower limit of all tested configurations. When increasing the number of samples per chicken to three, a Top-1 accuracy of 0.55 and Top-5 accuracy of 0.78 were reached for a setting with 100 chickens. Similarly, four images per chicken increased the Top-1 accuracy to 0.57 and the Top-5 accuracy to 0.81. Here it is noteworthy that only 88 chickens in the entire test set had five or more images. Therefore, this configuration was limited to a searched

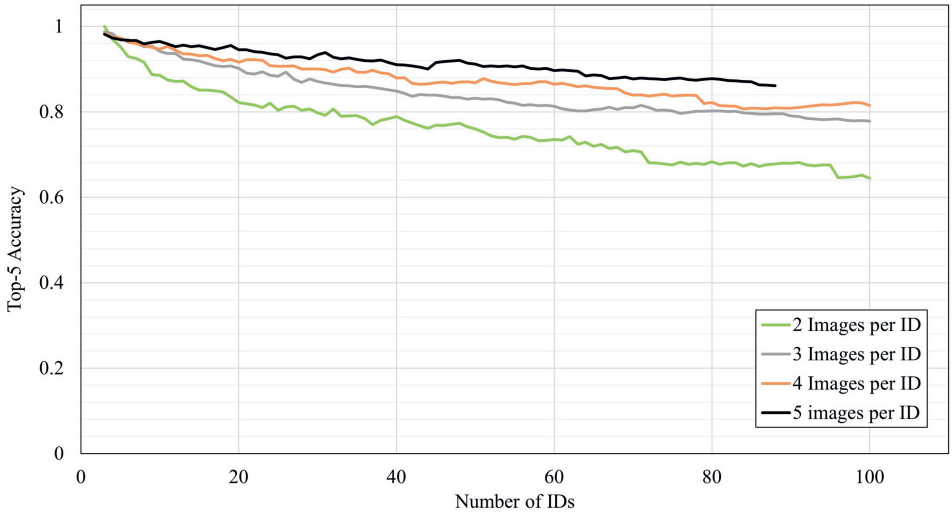


Figure 5.7: Top-5 re-identification accuracies for varying numbers of distinct chickens and images per chicken. Results were obtained by averaging over multiple sets of 2-100 IDs.

population of 88 chickens, yielding accuracies of 0.59 (Top-1) and 0.86 (Top-5).

5.3.3 Discriminative features

In this experiment, we utilized Grad-CAM to analyze the relevance of image regions among the evaluated approaches. Figure 5.8 shows the visualizations obtained for five sample images from the test dataset for each of the assessed architectures and triplet mining strategies. While the results cannot always be interpreted as a deterministic assessment of the discriminative importance for each feature, they do provide an intuition about which parts of the image are relevant for the purpose of re-identification. The experiment indicated that the avian comb, wattles, and occasionally, the chicken's earlobes were the features which were most frequently emphasized across the different approaches. Furthermore, it was noticeable that the CNN-based models captured information from smaller spatial regions and often focused on a single feature, such as either the comb or the wattles, while the transformer-based model attended to multiple, more diversified parts of the head. Instead of focusing on a specific local area, transformers captured global dependencies within the image, such as the comb, wattles, and various parts of the chicken's beak simultaneously. In addition, it was shown that all approaches, except those based on Hard Negative Triplet Mining, were able to learn discriminative features within the chicken's head region. In contrast, utilizing hard negative samples caused the models to sometimes focus on background features beyond the chicken, which may not be advantageous for animal re-identification. This observation was made for both

transformer and CNN-based approaches and aligns with the lower accuracies which were observed for the Hard Negative Mining strategy in experiment 1.

The frequently noted focus on the avian comb and wattles as discriminative features of a chicken aligns with human intuition and is supported by the approach presented in Corkery et al. (2009), which investigated the avian comb as a biomarker. However, the evaluations revealed that not only the outer comb profile, as used in the referenced paper, served as a relevant feature. Depending on the approach, the entire comb (e.g., VisionTransformer L with Semi-Hard Negative Mining) and the edge between the comb and the head (VisionTransformer S with Semi-Hard Negative + Easy Positive Mining) were also considered relevant. Moreover, particularly in the case of transformer-based approaches, attention was also given to diverse, small features within the chicken's face that were challenging to recognize as distinctive features from a human perspective. Overall, the experiment confirmed that the approaches based on Semi-Hard Negative mining were able to learn meaningful features for re-identification, going beyond random chance.

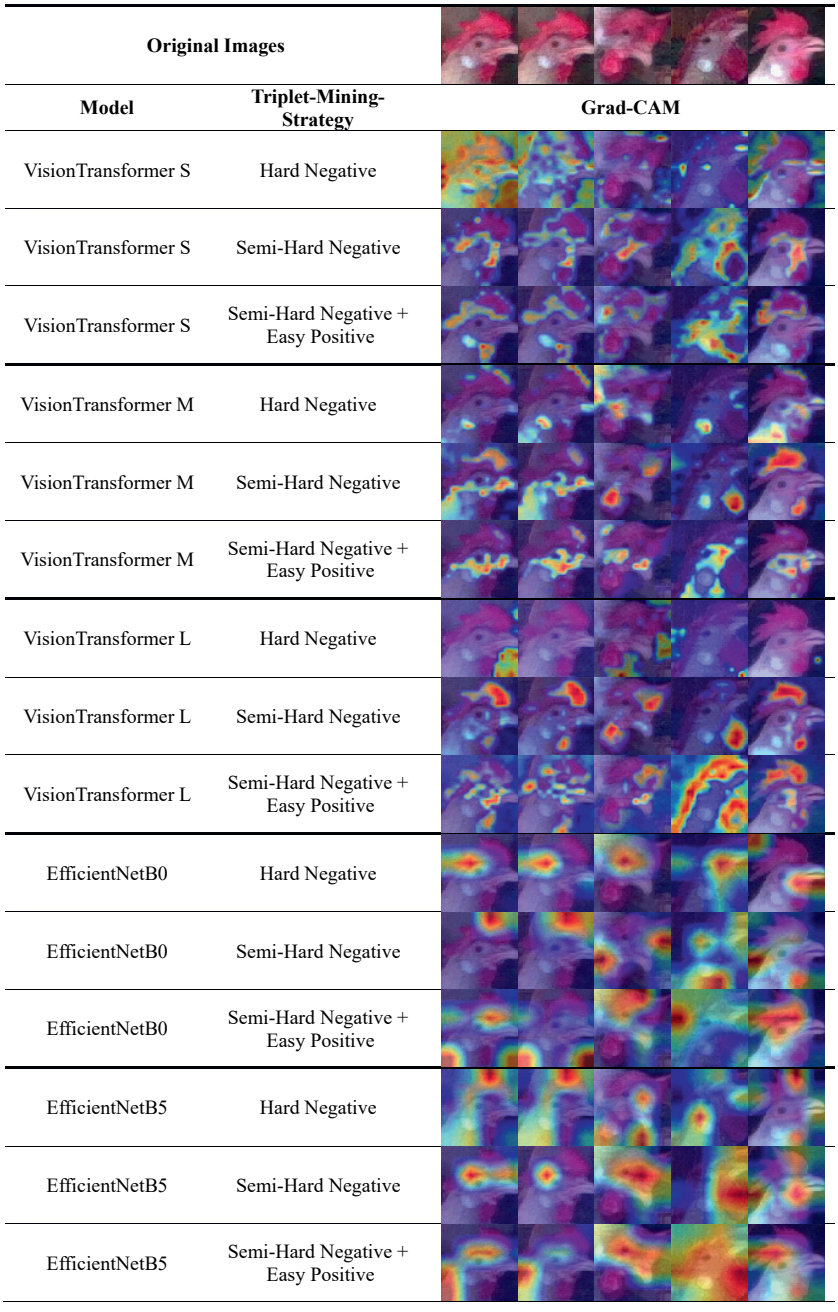


Figure 5.8: Exemplary Grad-CAM visualizations for the evaluated model architectures and triplet mining strategies.

5.4 Discussion

5.4.1 Comparison to existing approaches for chicken re-identification

The existing literature on re-identification approaches within the domain of poultry is quite limited. The method presented in this study is the first to employ deep learning for the purpose of re-identifying chickens from images. In a previous study (Corkery et al., 2009), the avian comb, was identified as a potential biomarker, and re-identification based on its shape was evaluated in a test setup involving 40 chickens with four images per chicken. For that experiment, a Top-1 accuracy of up to 0.84 was reported. While these results were promising for an initial investigation, the authors outlined several limitations of their work that needed to be addressed. Their work focused solely on the avian comb, requiring controlled conditions with chickens held in front of a clean background and in a standardized pose to ensure clear visibility of the comb in all images. Moreover, each recorded chicken had a relatively small comb, which prevented it from flopping to one side – a condition that cannot be guaranteed in real farm environments. Instead, combs might be damaged or distorted, which, combined with their natural growth, hinders the exclusive reliance on combs for practical re-identification applications. In contrast to that study, our approach utilized the entire head region and employed a deep learning approach to learn which characteristics of the head region are beneficial for re-identification without predefining specific features. The work in Corkery et al. (2009) focused on general feasibility, while our research takes a step towards a more robust solution applicable in livestock applications. Using the same setting as Corkery et al. (2009) with 40 chickens and four images per chicken, our approach achieved a Top-1 accuracy of 0.70. Despite being lower than the pilot study results, it's crucial to note that our images were captured under real farm conditions in a commercial chicken house, encompassing all external factors such as varying poses of the chickens, damages or distortion of the comb, and varying illumination conditions. While our study was not limited to the avian comb, the analysis of important image regions in experiment 4 suggested it to be a relevant feature for re-identification. However, most of the evaluated configurations in the experiment did not solely attend to the comb but rather considered it in conjunction with other features such as the wattles and earlobes. This multi-feature approach enhances the robustness of the method by ensuring functionality even when the comb is obscured.

5.4.2 Ground-truth data limitations

In this study, we based our approach for re-identification on images capturing the head area of the chickens. At first sight, this might seem counterintuitive, considering the potential additional visual features on the chicken's body that are ignored if the focus is limited to the head region. However, the analysis of the body primarily focuses on the plumage, which is highly subject to change over time and appears very similar for each chicken, lacking distinct features at first glance. While deep learning algorithms

may be able to learn discriminative features from the plumage or other body parts, there is a significant risk of unintentionally learning indicators for chicken re-identification. Examples of such features could include damages or dirt in the plumage of a chicken, or the orientation of the feet. These features are consistent within a short period of time, for example, between consecutive frames, but they lack permanence and are not useful as biometric markers for re-identification. Especially in a data collection setup that is based on single sequences per chicken, as was used in this study, the risk of learning non-permanent features is significant. Focusing solely on the head region can still result in similarities between frames within a sequence; however, head orientation typically varies throughout the recording. This variation in the training data reduces the risk of the model learning head orientation as an unintended feature for re-identification.

While experiment 4 demonstrated that the head region provides features relevant for re-identification, some of these features, such as the comb morphology changing as chickens mature, may potentially limit their long-term effectiveness. Therefore, it is recommended to conduct long-term data collection, including labeling chickens, to evaluate this potential issue.

Another relevant aspect of our data mining strategy is the occurrence of false negative samples. Although the automated image recording in real laying hen farms, as employed here, allows for the simplified collection of many images, it can also result in a single hen being recorded multiple times and misidentified as distinct animals. To mitigate this issue, we implemented several measures for both training and validation. These measures included collecting data over multiple weeks and enforcing a minimum time distance between negative pairs when forming training triplets. Nevertheless, completely eliminating false negatives may not be entirely feasible. In the context of validation, this means that a correctly identified chicken could be evaluated as a false prediction, potentially underestimating the true re-identification accuracy.

5.4.3 Practical applicability

As the present study aimed to establish a foundation for re-identification approaches in practical settings, its real-world applicability is crucial. Given our method's focus on matching chicken head images, its application in a practical farm setting requires integration with a head detection and cropping method. The current study utilized a standard YOLOv5 model for preprocessing image datasets, which could also be a promising option for future applications. This would also enable automatic image filtering based on the presence of key chicken features, mirroring our data collection methodology that ensured images of sufficient quality for accurate matching.

Experiments showed that matching accuracies beyond 90% could be achieved for small animal populations when considering multiple images per chicken. These accuracies decreased to around 0.4 (Top-1) and 0.6 (Top-5) when only one matching image was avail-

able within a population of 100 chickens. Determining whether these accuracies are "good enough" for real applications depends on the specific application. For tracking methods that aim to identify individual chickens within a video sequence, the number of chickens considered is usually small, and consecutive frames often show high similarity. Therefore, our re-identification method can be expected to deliver high accuracy in such tracking scenarios. Similarly, monitoring small test groups of chickens in a farm environment, for example, for scientific purposes, where providing multiple images of a certain chicken is feasible, is expected to be successful. On the other hand, the most extreme application would be a one-shot recognition of chickens within crowded commercial farms. In such a case, only one matching image would be available, even though the population size is large. While our results indicate accuracies significantly exceeding chance levels, they might not be reliable enough for such applications. Addressing this challenge requires either simplifying the re-identification task itself or further enhancing the current accuracy of our method. To enhance the performance of our method, one option could be to further standardize recordings to ensure clear visibility of relevant features. Although the camera in our recording setup was positioned near a drinker to capture each animal in a similar pose, the actual poses varied significantly. An effective strategy to address this issue could be to record only when chickens are actively using the drinker, ensuring a more standardized pose during capture. Recording entire sequences of each chicken would further allow for acquiring multiple images per chicken, thereby enhancing matching accuracy. Additionally, future data collection efforts should consider incorporating long-term data encompassing various stages of chicken development, such as comb and wattle growth. Given the increased complexity of the data collection, this issue was not addressed in the current study.

5.5 Conclusions

In this work, we presented an approach for the re-identification of individual laying hens within an uncontrolled farm environment. We gathered a detailed image dataset to train a neural network using the transformer architecture. The goal was to develop a transformer-based neural network capable of learning distinct representations for individual chickens. Using various triplet-mining strategies and model sizes, the model architectures were then assessed in comparison to traditional CNN-based approaches. Moreover, the analysis examined whether the number of available images per chicken and the population size affect the re-identification accuracy, and which visual features of a chicken's head are relevant for the specific models. Results demonstrated that the transformer-based architectures outperformed CNN models, while the use of semi-hard negatives during triplet mining yielded the best results among all evaluated configurations. The variation in population size and sample number per chicken resulted in re-identification accuracies exceeding 0.95 for simple configurations with small populations and multiple images per chicken. In the

most challenging scenario, with only one matching image and a population size of 100 chickens, accuracies remained around 0.4 (Top-1) and 0.6 (Top-5). Furthermore, it was revealed that the evaluated models learned to prioritize features such as the comb, wattles, and earlobes, aligning with human perception. However, they also attended to the entirety of individual features and characteristics of a chicken's head.

We conclude that the proposed approach shows promise for re-identifying individual hens even when recorded in an uncontrolled farm environment, laying the groundwork for future applications in animal tracking and monitoring.

Chapter 6

General discussion

This thesis explored the use of computer vision for individual livestock monitoring, aiming to advance the field by developing robust, practical, and effective methods specifically designed for application in commercial farm environments. This involved the development and evaluation of both novel techniques and advancements to existing solutions, with a particular focus on addressing limitations that impede their practical application. These limitations included robustness to environmental effects and variations, as well as factors such as decision transparency and the adaptability of methods to multiple use cases.

In this context, it was hypothesized that the utilization of state-of-the-art deep learning methods would enable individual animal monitoring solutions in uncontrolled commercial farm environments. In Chapters 2 through 5, four deep learning-based methods were presented, focusing on the use case of assessing plumage condition in laying hens. However, these methods were developed with broader applicability in mind, aiming to establish a monitoring framework that is modular and adaptable beyond the specific use case. To achieve this, the four approaches were designed as modular components, allowing them to be combined or used independently with other methods.

In Chapter 2, ChickenNet, a deep learning model for detection and segmentation with an additional regression output for assessment tasks, was developed. The model was trained on image data from a commercial laying hen farm, with images artificially augmented during training to introduce additional variance. It was shown that the model was able to detect chickens with an accuracy of up to 98%, while plumage conditions were assessed with an accuracy of 92% compared to human annotators. Moreover, it was revealed that the performance of neither hen detection nor plumage condition assessment was generally enhanced by the use of additional depth information.

Based on these results, Chapter 3 utilized the developed ChickenNet architecture and investigated the implementation of different uncertainty estimators into the assessment model. In this work, it was revealed that the uncertainty estimation derived from the predicted occlusion level did not correlate with the predictive error of the model. However, the proposed estimation of epistemic and aleatoric uncertainties did demonstrate such a correlation. Consequently, rejecting assessments identified as uncertain by these two estimators enhanced the overall assessment performance of the model. In addition, the transferability of both the uncertainty estimation methods and the ChickenNet architecture to a different use case, such as human age estimation, was demonstrated.

Chapter 4 then focused on the assessment based on information from entire image sequences rather than single images to enhance the quality of assessment. The core principle of this approach was to fuse multiple individual assessment predictions made within one or more sequences to generate a final output. For the weighting of the individual predictions, the uncertainty estimators developed in Chapter 3 were utilized. The fusion-based assessment outperformed the traditional image-based approach by up to 3% for plumage condition assessment and up to 7% when evaluated on the human age estimation dataset.

Such improvements were confirmed for both aleatoric and epistemic uncertainty estimators, as well as for image sequences of varying lengths.

Finally, in Chapter 5, a similarity learning approach was developed for the re-identification of individual hens to enhance visual monitoring. This study addressed both the association of detections within sequences, building upon the previously developed method, and the re-identification of laying hens in general. While the evaluated transformer-based architectures outperformed the CNN-based approaches in all tested configurations, the best results were obtained when combining the transformer architecture with a training strategy that utilizes semi-hard negative samples. Within a small group of fewer than five chickens and given multiple image samples per chicken, the Top-1 matching accuracies surpassed 90%. Accuracy decreased as the number of chickens increased and the number of images per chicken decreased. This resulted in matching accuracies between 40% and 60% for a group of 100 chickens, depending on the available number of images per chicken.

In the following sections, a more detailed discussion and methodical reflection is provided on each of the developed approaches, emphasizing their contributions to the overarching objectives of this thesis. Given the high importance placed on the practical application of the developed methods, considerations for deployment are discussed in a dedicated section. Finally, the societal relevance and potential implications for current practices in livestock farming are discussed, concluding with a look towards future research directions.

6.1 Contributions and methodical reflection

This section discusses the scientific contributions of each approach developed in this thesis. The discussion centers on how these approaches align with the overall thesis objectives and their potential impact on the practical application of individual animal monitoring. Additionally, a methodological reflection, addressing both the strengths and limitations of the techniques employed.

6.1.1 End-to-end deep learning for individual animal assessment

The method developed in Chapter 2 aimed to provide an approach for assessing individual animals within a flock while considering the challenges posed by commercial farm environments. Looking at the overall scope of this thesis, the outcome of this chapter — a neural network for simultaneous detection, segmentation, and assessment — can be seen as the foundation of the methods developed in subsequent chapters, providing individual animal assessments for further processing.

The first contribution to the overarching objective of advancing individual animal monitoring in livestock farming was the transition from the group level to the individual animal. This was addressed by utilizing the Mask R-CNN architecture as a backbone,

offering multi-object detection and segmentation for each input image. This allows to extract multiple animals from an image and to provide assessments on an individual level. Reflecting on this methodology, it could be argued that a simple detection without an additional segmentation might have been sufficient, shifting the focus towards the individual more efficiently than a complete segmentation. While the argument is valid, the architectural choice was mainly made for flexibility reasons. During this early stage of development, it was not yet clear whether subsequent methods could benefit from the precise segmentation of an animal without background information. To maintain this option, both bounding boxes and segmentation masks were obtained. The additional regression output was introduced to enable the model to learn numerical assessments for each individual detection, in extension to the usual classification. While most assessment tasks in animal monitoring could also be formulated as classifications, the regression output becomes valuable for assessments whose outputs are based on a continuous spectrum. Other examples beyond the evaluated use case of plumage condition assessment could include estimating animal weight or age. In such cases, regression can produce smoother transitions, whereas classification might lead to abrupt changes in the output.

The second contribution of Chapter 2 to the overall objective was to ensure robustness against environmental conditions in a commercial farm. This motivated the choice of a learning-based approach that was trained on image data acquired in a practical farm environment. In contrast to traditional approaches that rely on manually designed features, relevant features can be learned directly from the images. This allows to achieve invariance against variations typically encountered in farm settings, such as image noise, illumination changes, and occlusions. Additionally, to accommodate changes in environmental conditions and the animals themselves, data collection was conducted in three stages, with several weeks apart. To further increase the variation, artificial augmentations were incorporated into the images during the training process. Nevertheless, even with the provided image data and additional augmentations, it remains impossible to encompass all potential variations that may arise in a farm environment. It is undeniable that conditions can be more extreme than what the data collection in the aviary of a single farm could capture. For instance, factors such as camera pollution or the effect of different chicken breeds were not taken into consideration. These can only be addressed through extended and long-term data collection across multiple farms.

While being evaluated for the use case of assessing plumage condition in laying hens, the selected learning-based approach offers the flexibility to be extended to other animal monitoring tasks by training the model on different data. Potentially, this includes any task that requires assessing individual animals. When adapting the method to other tasks, it is important to note that the quality of the results is significantly affected by the training data. For animal monitoring tasks traditionally conducted manually, this presents two key challenges. The first challenge concerns the quantity of the collected training data. Larger datasets capture a wider range of visual variances compared to

datasets with a limited number of images, which is beneficial for the generalization ability of the trained model. Acquiring and manually labeling large datasets, however, requires significant effort, storage capacity, and eventually, higher costs. The second challenge relates to the quality of data, particularly the ground-truth labels. Human assessments of animal monitoring tasks are often subjective, potentially leading to ambiguity in the labels used to train the model, as observed in our experiments on plumage condition assessment. This issue could be approached by collecting multiple independent assessments for each image and averaging the resulting labels to create a more reliable ground-truth value. While this may enhance the quality of the labels, it also increases the efforts required for data collection and the associated costs.

Theoretically, the application of the presented method is not limited to the domain of animal monitoring and could be employed for any type of assessment task that requires the prediction of a numerical value for a detected object. This potential for transferability was indicated by our successful retraining of the model for age estimation, achieving good performance. Nevertheless, the promising results do not guarantee universal effectiveness in all use cases. Moreover, while the method can be adapted for various purposes, it might not always be the most efficient choice. This can be illustrated by the example of human age estimation from pedestrian images. While our experiments on these images aimed to showcase transferability on a publicly available dataset, this specific task did not necessarily require object segmentation or detection, potentially making the method over-engineered for this application.

6.1.2 Uncertainty estimation for reliable assessments

After Chapter 2, the methods developed for animal monitoring involved a neural network capable of providing assessments for individual animals housed in larger groups. Still, at that stage, the network was giving assessment predictions whenever an animal was detected in an image. These predictions were made entirely without considering aspects like the complete visibility of the animal or the overall image quality. Additionally, no indication of the assessments' quality or reliability was provided. In regards of practical application, this approach was suboptimal for understanding the decisions made by the model and also hindered the prioritization of specific predictions for subsequent processing tasks.

Therefore, in Chapter 3, this issue was addressed by developing methods to estimate the uncertainty of each prediction made by the assessment model. This estimation allowed us to identify and reject erroneous predictions, as demonstrated in the experiments of this chapter. Regarding the implementation of automated animal monitoring solutions in real-world scenarios, the approach tackled two key challenges.

First, it allowed the handling of unexpected inputs and inputs of low quality. As discussed in Section 6.1.1, it is nearly impossible to collect training data that covers all potential

variations that could arise from the farm environments and the monitored animals. This means, in an uncontrolled environment, the model will inevitably encounter input that it has not been trained on. The estimation of epistemic uncertainty enables the indication of such unforeseen situations and allows the model to say that it's uncertain due to a lack of knowledge. Aleatoric uncertainty enables the same capability but for input images that are of insufficient quality for the assessment task. While in a farm environment, it cannot be avoided that factors such as reflections, dirt, or motion blur may affect the image quality, the indication of uncertainty at least allows for the exclusion of such compromised images.

The second challenge addressed by this approach was the transparency of decisions. This becomes particularly relevant for a monitoring system when conveying assessments to users. Instead of being a black box that simply provides a prediction, the assessment model can now indicate the level of uncertainty associated with a prediction and whether the uncertainty stems from weaknesses in the model itself or from poor input quality. This would, for instance, allow a farmer to determine whether indicated health issues in a farm are based on reliable model predictions or on predictions compromised by external factors. In general, by offering a clearer understanding of decisions, uncertainty estimation helps prevent overinterpretation or excessive reliance on specific model predictions.

The motivation behind implementing uncertainty estimation in the presented manner was to avoid the need for manual definition of every source of uncertainty beforehand. Especially for the application in uncontrolled environments with numerous influencing factors, this characteristic is crucial. The method's ability to either learn uncertainty implicitly during training (aleatoric uncertainty) or to estimate it during inference (epistemic uncertainty) also facilitates its adaptation for other monitoring tasks under varied conditions. However, this enhanced flexibility simultaneously poses a limitation in terms of the practical application of the approach, as it only offers a quantitative estimation of uncertainty. Even though the method distinguishes between epistemic and aleatoric uncertainty, it currently lacks the capability to precisely explain the reasons behind high or low uncertainty estimates. For instance, high aleatoric uncertainty could arise from a dirty camera lens or from a non-ideal pose of the animal hindering the assessment. From a user perspective, such additional differentiation would enhance assessment transparency and simplify the initiation of corrective measures, such as cleaning the camera. Future research could, therefore, focus on providing more comprehensive information on this aspect. An initial step might involve distinguishing technical quality metrics of an image, such as exposure or sharpness, from other content-related quality issues, such as the pose of an animal. While the technical quality metrics apply uniformly to various visual assessment tasks, content-related issues strongly depend on the specific use case. Checking the technical image quality in an initial preprocessing step before estimating the uncertainty of the prediction could enhance the traceability of decisions and still maintain flexibility for the specific application.

While our experiments on human age estimation made an initial attempt to demonstrate the transferability of uncertainty estimation to other domains, future research could delve deeper into exploring uncertainty-aware predictions in animal monitoring beyond regression tasks. Specifically, extending uncertainty estimation to other aspects of the assessment method, such as the animal detection, could be promising. This would enable the filtering of unreliable detections, thereby ensuring that only the most confident predictions are used for further analysis. This extension would introduce an additional layer of detail and transparency by indicating whether the animal itself is well-recognized before conducting any assessment.

6.1.3 Uncertainty-aware fusion of sequence information

While Chapters 2 and 3 focused on enhancing assessments derived from individual images, Chapter 4 expanded this scope by investigating the fusion of information from entire image sequences. The central idea in this part of the research was that assessments derived from multiple images within a sequence of an animal can vary in their reliability and informativeness. By aggregating these assessments and selectively incorporating those with low uncertainty into the final assessment, the aim was to improve the overall quality and robustness of the assessment process.

This work made use of the previously developed methods to obtain assessments on individual animal level and weight them by their estimated uncertainty. Considering the goal of this thesis is to develop robust deep learning approaches for individual animal monitoring, it may initially seem contradictory not to choose an end-to-end learning approach for sequence assessment. Such an end-to-end approach appears as a valid choice and could have been realized by adjusting ChickenNet to process entire image sequences. However, the decision to utilize multiple independent predictions on image level and fuse them was made having the practical application and the adaptability of the method in mind. The fusion of multiple individual assessments enables the exchange of the underlying assessment model. Thus, existing models trained on single images can be extended to process entire sequences, which is crucial when integrating this method with existing tools. In contrast, an end-to-end model would require training the model on entire image sequences, which might not always be available for each animal monitoring use case. Looking at similar assessment approaches in other domains, end-to-end models typically require defining the sequence length beforehand (Wang et al., 2022b). This, in turn, would restrict the system to sequences of a specific length while excluding any recordings of animals that are shorter than the required length. Both characteristics hinder the practical applications of the approach.

In contrast, our method presented in Chapter 4 does not require training on entire sequences and provides a greater flexibility regarding sequence length. In the experiments, it was shown that the approach outperformed image-level assessment even with a limited

number of images per sequence. Moreover, avoiding simultaneous processing of multiple images, as required by end-to-end solutions, reduces computational complexity. This is particularly relevant for commercial applications where efficiency is crucial.

Similar to Chapters 2 and 3, Chapter 4 also demonstrated the transferability of the developed approach to other use cases, as exemplified by its application in human age estimation. Here, it is noteworthy that applying this method across various use cases, within animal monitoring or beyond, might require the adjustments of model parameters. One parameter is the uncertainty threshold that determines the level of uncertainty beyond which an assessment is disregarded for the final evaluation. The higher the threshold, the more assessments are ignored. In use cases where obtaining a sufficient quantity of images per animal is not a challenge, this threshold could be set higher compared to scenarios with limited observations per animal. Future research could further focus on replacing this fixed threshold with an automated calibration mechanism or a learning-based approach. Such advancement would simplify deployment for other monitoring tasks.

Likewise, the prioritization of specific viewpoint clusters, as explained in Chapter 4, is only beneficial when the presence of specific characteristics definitively determines the final assessment. For assessment tasks where this is not the case, such prioritization may be irrelevant and could be disregarded, as was done for the evaluated task of human age estimation.

6.1.4 Similarity learning for chicken re-identification

A limitation of the method presented in Chapter 4 was the use of ground-truth identity labels to associate different assessments to a certain individual. Real-world applications, however, lack this pre-existing identity information, which becomes critical when dealing with multiple animals recorded simultaneously. Consequently, an online mechanism for assigning measurements to specific animals would be necessary in such situations.

Integrating an existing, fully functional tracking method in the final chapter could have addressed this limitation and provided a comprehensive synthesis of the methods developed throughout this thesis. In this regard, established methods like Simple Online and Real-time Tracking (SORT) (Bewley et al., 2016) or Deep SORT (Wojke et al., 2017) appeared readily integrable. However, these approaches have limitations for their application in animal monitoring. Basic methods such as SORT, which depend only on motion models, exhibit shortcomings in the presence of occlusions (Wojke et al., 2017), a common challenge in dense animal groups. While more advanced methods like Deep SORT additionally consider appearance information through learned feature embeddings for identity discrimination, the underlying embedding models are typically trained on person re-identification datasets, limiting their effectiveness for chickens. Given the critical role of data association in the overall monitoring system, it seemed impractical to use a pre-trained feature embedding model.

A straightforward approach could have been to train the feature embedding model of an existing tracking method on chicken data. However, the neural networks in these methods are often simple, focusing on short-term tracking and instance-matching between individual frames of a sequence. In animal monitoring, data association is not only relevant within short sequences, as required by the method in Chapter 4. Instead, continuous monitoring of an animal requires the assignment of measurements taken at different points in time, thus demanding long-term re-identification of the individuals.

Therefore, in Chapter 5, we chose to delve deeper into the re-identification topic, developing an approach that contributes to the automation of animal monitoring by not only serving as a foundation for improved short-term tracking, but also addressing long-term re-identification goals. A key consideration during method development was, once again, the applicability in commercial farm environments. Despite environmental effects on the visual appearance of the animals, this scenario involves large groups where obtaining training images for each individual chicken is infeasible. Thus, we opted for a deep learning approach for similarity learning, which aims to predict whether two input images belong to the same identity. This approach allowed for the recognition of an individual without prior training on images of that specific individual. Furthermore, it also facilitated future training on images of other livestock species.

Despite the development of a data association method suitable for commercial farm environments, Chapter 5 also addressed the fundamental question of the feasibility of re-identifying chickens. In this context, this chapter can be considered early-stage research, in contrast to the previous chapters, which primarily focused on addressing weaknesses in existing monitoring approaches and enhancing their robustness. This focus on feasibility was primarily motivated by the significant proportion of chickens in livestock, making them highly relevant for animal monitoring. Although distinguishing individual chickens may seem challenging for humans due to the chicken's similar appearance, successfully implementing re-identification techniques could offer significant benefits for livestock monitoring and was therefore investigated. Despite its importance, image-based re-identification methods have received little attention for chickens compared to other livestock animals. For example, re-identification methods have been developed for cows (Wang et al., 2023; Chen et al., 2022b) and pigs (Wang et al., 2022a; Wang and Liu, 2022), whereas chicken re-identification has only been explored in one preliminary study (Corkery et al., 2009).

The experiments in Chapter 5 indicated the general ability to re-identify individual chickens in different images, even when recorded in uncontrolled farm environments. Here, promising results were obtained for groups of chickens with up to 100 individuals. Nevertheless, the results were not yet at a level that would enable reliable re-identification of individual animals within a commercial flock comprising thousands of birds. However, for small groups such as those present in research environments, or for data association

within image sequences, the shown accuracy can expected to be sufficient. Verifying this expectation by integrating the re-identification approach with the previously developed monitoring methods could be focused by subsequent research. Another question to address could be the application of the re-identification model to other livestock species. Considering the promising results on laying hens and the learning-based method that allows retraining on other data, there is a good chance, but currently no evidence, that the approach would reach similar performance in other animals.

6.2 Practical implications

While the first part of this final chapter discussed the scientific contributions and limitations of the developed approaches in relation to the overall research objectives, the following section addresses considerations for the practical application of the developed methods, which was a central focus of the research. This includes the integration and transferability of the presented methods, enabling them to be utilized as a toolbox for various monitoring tasks. Moreover, broader practical considerations that extend beyond the software focus of this thesis are discussed.

6.2.1 A toolbox for automated animal monitoring

The methods developed in this thesis aimed to enhance monitoring robustness, primarily focusing on their application in uncontrolled farm environments. Additionally, these approaches were designed to be sufficiently flexible to adapt to various animal monitoring use cases. This resulted in the development of a framework comprising several deep learning-based techniques, addressing common challenges encountered in animal monitoring.

To finally evaluate all developed methods in combination, it is still necessary to integrate the re-identification approach from Chapter 5 into the sequence assessment method developed in Chapter 4. However, it is important to recognize that the presented work encompasses a comprehensive range of methods while not all of them may be essential for every use case. Instead, they can be seen as a “toolbox” for individual animal monitoring, from which the methods can be chosen as needed. For instance, in tasks such as visual weighting of animals, it might be beneficial to indicate the uncertainty of predictions while the fusion of measurements taken from different viewpoints might not improve the quality of the weighting results. In contrast, more holistic assessments, such as the evaluated plumage condition assessment benefit more from the fusion of information from multiple images. Similar applies to the re-identification of animals, which, as explained in Chapter 5, is not necessary in every application. The added value of simultaneously employing all methods further depends on whether they are intended for integration into existing systems. For instance, if a monitoring system that only provides single images per

animal is to be expanded, the ability to fuse multiple images from a sequence is logically unnecessary.

A core principle pursued during development was the interchangeability of individual methods within the entire framework. This allows for the replacement of, for example, the detection and assessment model without the requirement to adjust downstream processing methods. Moreover, the framework can be adapted to other monitoring tasks by simply using different training data. This capability was demonstrated throughout Chapters 2 to 4, with human age estimation serving as an example.

Despite the interchangeability, transferability is crucial for practical implementation. It reduces development effort and associated costs, eliminating the need to create entirely new approaches for each specific application. When adapting a new use case, the effectiveness of the entire framework is significantly influenced by the underlying detection and assessment model. Therefore, for the practical implementation of these methods in other applications, it is crucial that the model is well-trained for the specific application case. This requires sufficient high-quality data for the particular task. In this regard, diversity and balance of the data are important to ensure that the model is not biased towards specific patterns or viewpoints. However, data collection for specialized animal monitoring tasks is often time-consuming and expensive, resulting in relatively small datasets. In this regard, transfer learning offers a promising approach to address this limitation. Large-scale public datasets like ImageNet encompass an extensive number of images from various domains, providing valuable training resources for deep learning algorithms. Moreover, the underlying visual characteristics shared among different animal monitoring tasks enable the pre-training of models that capture these common patterns. This significantly reduces the amount of training data required for specialized tasks. An example is health monitoring in laying hens and broiler chickens. While both tasks differ in their specific objectives and focus on different animal species, shared visual characteristics, such as body shape and posture, enable pre-training models on a large dataset of chicken images and fine-tune them on a smaller dataset of images specifically for each species.

6.2.2 Implementation beyond algorithms

While this thesis focused on the question of how deep learning approaches can be utilized to enhance the performance of animal monitoring algorithms in uncontrolled farm environments, it is crucial to acknowledge that the development of robust software alone is not the only determinant of an effective camera-based monitoring system. Even though it is essential to ensure that algorithms are able to handle diverse environments, it is usually unnecessary to build resilience against every imaginable extreme situation. Instead, a more efficient approach involves proactively mitigating external disturbances before they can impact the recorded images. Considering image-compromising factors such as dirt

on the camera lens or sunlight reflections, the developed methods addressed these issues and focused on minimizing their impact on monitoring performance. However, avoiding such environmental disturbances from the beginning would further enhance the overall performance and reliability of the animal monitoring system.

In this regard, one key aspect is the placement of cameras within the farm, which can significantly minimize the influence of environmental factors. For instance, selecting a contrasting background, avoiding direct sunlight, and ensuring that cameras are not placed under porches can prevent dirt on the lens and reduce variability in input images.

Beyond camera placement, hardware selection plays a significant role in mitigating environmental disturbances. Proper housing protects cameras from dust, moisture, and other environmental factors that can degrade image quality. Additionally, cameras equipped with adaptive lighting capabilities and automatic focus mechanisms can effectively handle varying lighting conditions and ensure a sharp focus on the target animal.

Another promising approach, which has been the subject of research in other domains such as the automotive industry (Lee et al., 2017), is the development of self-cleaning devices. Unlike passive protection, these devices actively clean the camera lens to maintain a clear view. Nonetheless, it is obvious that not all external disturbances can be avoided, especially regarding variations among animals themselves, which are beyond the control of the monitoring system. This underscores the relevance of robust image processing methods, as developed in this work.

6.3 Societal relevance

All approaches and methods developed in this thesis aimed to contribute to the establishment of systems for automated individual animal monitoring in practice. In addition to the scientific objectives discussed earlier, this aim also involved addressing societal concerns and influencing current practices in livestock farming. Therefore, the following section explores the societal relevance of the developed methods while highlighting four aspects: public acceptance, economic benefits, animal welfare and ethical considerations. The section is concluded by a brief overview of potential future research directions, extending the present work.

Public acceptance

To actually yield a benefit for society, automated animal monitoring needs to achieve widespread adoption beyond research and prototypes, which means it must be adapted by stakeholders in commercial livestock farming. To achieve this, it is crucial that stakeholders trust technology and are willing to actively use it. In this context, the term "stakeholders" primarily refers to farmers, but also encompasses researchers, animal experts, and the general public. A recent study that investigated the perception of sensor-based continuous monitoring in laying hens identified doubts about the effectiveness and

validity of health and welfare assessment methods as the biggest obstacle for successful implementation in the sector (van Veen et al., 2023).

In this regard, the contributions of this thesis were twofold. First, it addressed the effectiveness of monitoring solutions in uncontrolled farm environments by enhancing their robustness. The deep learning methods showed strong performance for the use case of plumage condition assessment under conditions of a practical farm settings, while also being able to flag potentially inaccurate predictions. This increased transparency was the second contribution to address stakeholders' doubts about effectivity. Even though a model might not provide completely precise assessments at any time, the ability to indicate the reliability of a predictions can significantly foster trust in the overall monitoring system. However, while the estimation of uncertainty in the neural network decisions was a first step in the right direction, it must be acknowledged that there is still a huge potential to improve the transparency of deep learning-based monitoring systems. Ideally, such a system would offer explainable recommendations to users, enabling them to understand and evaluate the system's decisions.

Moreover, according to Hubbard and Scott Hubbard and Scott (2011), the acceptability of monitoring techniques can be improved by incorporating measures that align with those already used by farmers and experts in manual assessments. In this thesis, the principle was applied by using plumage condition scores from established manual scoring systems to train the assessment model. Nevertheless, this aspect should be carefully considered when adapting the methods to other applications.

Finally, to ensure the acceptance of monitoring methods in practice, it is evident, yet crucial, that the application actually provides benefits to the stakeholders. These may include a reduced workload for farmers, the potential for new insights for researchers, or enhanced animal welfare standards. Often, these benefits align with economic advantages, which will be further explored in the next section.

Economic considerations

Economic aspects play a key role in driving the commercial adoption of automated animal monitoring systems. In general, most automated solutions in livestock farming aim to reduce labor, either in response to labor shortages or to make processes cost-efficient, thereby increasing productivity. This typically involves replacing manual labor with machinery. Automated monitoring does not exclusively focus on replacing manual tasks; instead, novel approaches often aim to provide additional value that could not be achieved through manual monitoring. For instance, manual welfare assessments are typically sample-based and subjective, whereas automated solutions offer the opportunity for continuous and objective assessments. However, this does not imply a complete replacement of manual animal checks, nor is it the intended result. Automated monitoring, instead, serves as a supportive tool for farmers to make more informed, data-driven management decisions. From an economic perspective, this is particularly beneficial in improving reaction times

during emergencies, thereby enhancing farm yields. Even if the systems do not replace manual interaction with animals, they reduce the overall workload for farmers in the long term. Repetitive tasks such as documentation and data transfer can be significantly simplified by automated systems, freeing up time for farmers to focus on more valuable aspects of animal care and management.

Despite the monetary benefits of improved decision-making and reduced manual workload, investment costs represent the second critical factor to consider when evaluating the financial viability of automated animal monitoring systems. Key cost drivers in this context are the installation and hardware of the systems. For vision-based systems, as focused on in this thesis, the primary contributors to hardware costs are cameras and processing units. To ensure the practical applicability of the developed methods, device selections were made with financial considerations in mind. For instance, in Chapter 2, both normal color images and color images with additional depth information were evaluated, yielding comparable accuracies for the specific use case. Due to the higher cost of cameras and processing required for depth information, the choice was made for standard color images.

In general, the more devices needed on a farm, the higher the investment and installation costs. Thus, the preferred solution from an economic perspective would be a one-for-all monitoring system. This makes scalability of monitoring methods a crucial feature. Compared to other sensors which are specialized on measuring a single metric, camera-based monitoring offers the advantage of being able to address multiple tasks with a single sensor. For example, scales used for weighing animals could be replaced by cameras, which could also assess the animals' health and welfare. Considering the task of plumage condition assessment in laying hens, an intuitive approach might involve using thermal cameras to detect damage in the plumage. In a preliminary data collection, we tested such a camera and indeed revealed clear visibility of plumage damages in thermal images (Figure 6.1). However, the use of the tested device would be highly specific to this application, while exceeding the costs of standard color cameras by orders of magnitude. Thus, it was not selected for development. Moreover, to reduce the need for multiple cameras, this thesis focused on methods that enable the simultaneous recording of multiple animals while shifting from group-level to individual animals within the software. This approach enhanced scalability, resulting in reduced investment costs per task.

Animal welfare

The developed methods have made efforts to bring automated animal monitoring closer to practical application and to promote its usage in commercial farm environments. As previously discussed, robust and effective monitoring methods can extend the scope of manual animal checks, achieving an assessment quality that surpasses human inspection. This also applies to the assessment of welfare parameters. While the quality of manual assessments often depend on the observer, standardized data acquisition by autonomous

Figure 6.1: Thermal image recording of laying hens on a farm. Plumage damages are clearly visible due to higher temperatures in the affected areas.

systems could establish new standards for measuring animal welfare. Having a continuous stream of assessment data, it becomes possible to precisely track changes over time. In large-scale livestock farming, this is essential for early detection of issues, such as diseases and critical behaviors, which laying hen farmers indicated as the most important advantage of monitoring systems in a recent study (van Veen et al., 2023). Identifying deviations in animal welfare parameters would further allow the immediate evaluation of the effects of changes in feeding, environment, or human interaction on the animals. This capability would be relevant for both research and commercial livestock farming. Assuming a robust method for assigning distinct measurements to a particular animal, permanent tracking of changes could even be possible at individual animal level.

Nevertheless, it must not be neglected that monitoring alone does not guarantee improved animal welfare. Reliable systems are essential for supporting decisions, identifying issues, and provide warnings. However, true welfare improvements require translating the information into action. Therefore, the effective integration of monitoring methods into farm management is essential, either as a support tool for human decision-making or through automated adjustments. In this context, it is crucial to avoid complete dependence on technology in the early stages, as this may lead to animal neglect, reduced husbandry skills, and ultimately a negative impact on animal well-being.

Ethical aspects
Beyond potential impacts on animal welfare, there are other ethical considerations that arise with automated animal monitoring. Especially, the integration of deep learning approaches as a foundational element of the developed methods raises concerns inherent in artificial intelligence and automated decision-making in general. With the increasing focus on ethical implications of artificial intelligence, a multitude of guidelines and principles have emerged (Jobin et al., 2019). Although this expansive field cannot be fully covered here, two principles are worth highlighting, as they are particularly relevant for the application of automated animal monitoring systems: *fairness* and *transparency*.

These two aspects are closely related. Automated animal monitoring is designed to assist in management decisions or even trigger autonomous adjustments in the farm. Therefore, decisions made by the underlying models can have a significant impact on the animals. If such a model is biased and makes unfair decisions regarding certain animals, this can lead to inaccurate assessments, overlooked welfare concerns, or even harm to animals. In order to prevent potential biases and ensure fairness, transparency is crucial. Without understanding how decisions are made, underlying issues cannot be addressed. In Chapters 3 and 4 of this thesis, efforts were made to improve transparency in the model assessments by indicating the uncertainty associated to the predictions. For detecting biases, especially the epistemic uncertainty is relevant, as this model-related uncertainty indicates whether model predictions are uncertain due to a lack of knowledge. For example, if a model trained on data from one chicken breed is used to assess chickens of a different breed, the resulting predictions would likely exhibit significantly higher epistemic uncertainty. As already discussed in the context of public acceptance, this uncertainty estimation can be seen as a first step towards transparency, while there is still a great potential to make decisions more explainable.

Although monitoring systems themselves do not directly interfere with the animals and can rather be seen as silent observers, the actions they trigger may have ethical implications for the animals involved. While assessments can be used to identify and address health issues, improve living conditions, and enhance welfare, they can also be misused to make decisions with negative consequences, such as culling of animals that are "underperforming." Ultimately, ethical considerations must guide the interpretation and application of monitoring systems to ensure they serve as powerful tools for positive change and not as a source of harm.

6.4 Future directions

While the four methods developed in this thesis addressed practical challenges of animal monitoring, further research is essential to ensure the transition of automated monitoring from prototypes to real-world applications in livestock farming. Initially, this involves implementing the previously discussed improvements on the individual methods and con-

ducting extensive testing on other animal species to ensure their effectiveness. One key aspect in this regard is the creation of representative and balanced datasets to train and validate the learning-based approaches.

Looking beyond the methodical improvements, the integration of the methods is a fundamental aspect for the (commercial) success of automated animal monitoring in livestock farming. This integration includes two aspects. Firstly, it involves combining the diverse elements of the "toolbox" into a cohesive system. As argued previously, not every use case requires the utilization of all methods; however, when combined, each method influences the performance of the final system. For example, a perfectly accurate uncertainty estimator that indicates false assessments becomes redundant if all assessments are consistently inaccurate. Integration further involves combining multiple monitoring tasks within one system. Simultaneously monitoring several animal-related parameters is not only economically beneficial but also facilitates the detection of relationships between different metrics. An example of this could be the simultaneous monitoring of health conditions and feeding behavior. Having such an integrated system allows for centralizing data and presenting analysis in user-friendly visual representations to support informed decision-making.

Secondly, integration extends beyond the monitoring system itself, requiring incorporation into the existing farm infrastructure, and potentially even beyond. This includes the connection with other disciplines, such as animal behavior science, to provide better analyses and effectively utilizing the obtained data. Here, the question of the optimal utilization of the generated information arises. At its most basic level, an integrated solution could involve a system that simply presents information, such as trend lines, to be used by the farmer for making informed decisions. Taking a more sophisticated approach, this could be extended to the automated detection of deviations and warning systems that allow earlier adjustment management. Evolving further, monitoring systems could also be connected to other systems, such as farming computers to automatically make adjustments. Ultimately, information could even be used to be shared with other stakeholders, such as veterinarians or feed suppliers, to further optimize animal health and production efficiency across the entire agricultural network.

However, it remains crucial to be aware that establishing automated animal monitoring in practice requires the trust and acceptance of users. Regarding the impact on living animals, human interaction will be needed at least to double-check decisions made by the system. Furthermore, surveys among poultry farmers have shown that completely autonomous operating systems are currently not desired. This attitude might change with the evolving establishment of automated monitoring solutions, but starting with collaborative systems allows to understand and trust the technology before potentially transitioning to more autonomous modes.

References

Abdulla, W. (2017). Mask r-cnn for object detection and instance segmentation on keras and tensorflow. Available online: https://github.com/matterport/Mask_RCNN (accessed on 19.11.2020).

Agethen, K. (2023). Steckbrief zur tierhaltung in deutschland: Ein überblick. *Thünen-Institut für Betriebswirtschaft*. Available online: https://literatur.thuenen.de/digbib_extern/dn065683.pdf (accessed on 20.01.2024).

Al Nahian, M., Iftekhar, A., Islam, M. T., Rahman, S. M., and Hatzinakos, D. (2017). Cnn-based prediction of frame-level shot importance for video summarization. In *2017 International Conference on New Trends in Computing Sciences (ICTCS)*, pages 24–29. IEEE. https://doi.org/10.1109/ICTCS.2017.13.

Allen, A., Golden, B., Taylor, M., Patterson, D., Henriksen, D., and Skuce, R. (2008). Evaluation of retinal imaging technology for the biometric identification of bovine animals in northern ireland. *Livestock Science*, 116(1-3):42–52. https://doi.org/10.1016/j.livsci.2007.08.018.

Alonso, W. J. and Schuck-Paim, C. (2022). Cumulative pain: An evidence-based, easily interpretable and interspecific metric of welfare loss. https://doi.org/10.20944/preprints202208.0247.v1.

Amini, A., Schwarting, W., Soleimany, A., and Rus, D. (2020). Deep evidential regression. *Advances in neural information processing systems*, 33:14927–14937. https://doi.org/10.48550/arXiv.1910.02600.

Andrew, W., Gao, J., Mullan, S., Campbell, N., Dowsey, A. W., and Burghardt, T. (2021). Visual identification of individual holstein-friesian cattle via deep metric learning. *Computers and Electronics in Agriculture*, 185. https://doi.org/10.1016/j.compag.2021.106133.

Andrew, W., Hannuna, S., Campbell, N., and Burghardt, T. (2016). Automatic individual holstein friesian cattle identification via selective local coat pattern matching in rgb-d imagery. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 484–488. IEEE. https://doi.org/10.1109/ICIP.2016.7532404.

Arnez, F., Espinoza, H., Radermacher, A., and Terrier, F. (2020). A comparison of

uncertainty estimation approaches in deep learning components for autonomous vehicle applications. *arXiv preprint*. https://doi.org/10.48550/arXiv.2006.15172.

Astill, J., Dara, R. A., Fraser, E. D., Roberts, B., and Sharif, S. (2020). Smart poultry management: Smart sensors, big data, and the internet of things. *Computers and Electronics in Agriculture*, 170:105291. https://doi.org/10.1016/j.compag.2020.105291.

Avci, M. Y., Li, Z., Fan, Q., Huang, S., Bilgic, B., and Tian, Q. (2021). Quantifying the uncertainty of neural networks using monte carlo dropout for deep learning based quantitative mri. *arXiv preprint arXiv:2112.01587*. https://doi.org/10.48550/arXiv.2112.01587.

Aydin, A. (2017). Using 3d vision camera system to automatically assess the level of inactivity in broiler chickens. *Computers and Electronics in Agriculture*, 135:4–10. https://doi.org/10.1016/j.compag.2017.01.024.

Bachman, P., Hjelm, R. D., and Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32. https://doi.org/10.48550/arXiv.1906.00910.

Banhazi, T. M., Lehr, H., Black, J., Crabtree, H., Schofield, P., Tscharke, M., and Berckmans, D. (2012). Precision livestock farming: an international review of scientific and commercial aspects. *International Journal of Agricultural and Biological Engineering*, 5(3):1–9. https://doi.org/10.3965/j.ijabe.20120503.001.

Baxter, M. and O'Connell, N. E. (2020). Testing ultra-wideband technology as a method of tracking fast-growing broilers under commercial conditions. *Applied Animal Behaviour Science*, 233:105150.

Ben Sassi, N., Averós, X., and Estevez, I. (2016). Technology and poultry welfare. *Animals*, 6(10):62. https://doi.org/10.3390/ani6100062.

Berckmans, D. (2017). General introduction to precision livestock farming. *Animal Frontiers*, 7(1):6–11. https://doi.org/10.2527/af.2017.0102.

Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. (2016). Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE. https://doi.org/10.1109/ICIP.2016.7533003.

Bhole, A., Falzon, O., Biehl, M., and Azzopardi, G. (2019). A computer vision pipeline that uses thermal and rgb images for the recognition of holstein cattle. In Vento, M. and Percannella, G., editors, *Computer Analysis of Images and Patterns*, volume 11679 of *Lecture Notes in Computer Science*, pages 108–119. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-29891-3_10.

Blok, P. M., Kootstra, G., Elghor, H. E., Diallo, B., van Evert, F. K., and van Henten, E. J. (2022). Active learning with maskal reduces annotation effort for training mask r-cnn on a broccoli dataset with visually similar classes. *Computers and Electronics in*

Agriculture, 197. https://doi.org/10.1016/j.compag.2022.106917.

Blok, P. M., van Henten, E. J., van Evert, F. K., and Kootstra, G. (2021). Image-based size estimation of broccoli heads under varying degrees of occlusion. *Biosystems Engineering*, 208:213–233. https://doi.org/10.1016/j.biosystemseng.2021.06.001.

Blokhuis, H. J. (1989). *The development and causation of feather pecking in the domestic fowl: @Wageningen, Landbouwuniversiteit, Dissertation*. Landbouwuniversiteit, Wageningen.

Bo, L., Ren, X., and Fox, D. (092011). Depth kernel descriptors for object recognition. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 821–826. IEEE. https://doi.org/10.1109/IROS.2011.6095119.

Boreczky, J. S. (1996). Comparison of video shot boundary detection techniques. *Journal of Electronic Imaging*, 5(2):122. https://doi.org/10.1117/12.238675.

Bosse, S., Maniry, D., Wiegand, T., and Samek, W. (2016). A deep neural network for image quality assessment. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3773–3777. IEEE. https://doi.org/10.1109/ICIP.2016.7533065.

Bouma, S., Pawley, M. D., Hupman, K., and Gilman, A. (2018). Individual common dolphin identification via metric embedding learning. In *2018 international conference on image and vision computing New Zealand (IVCNZ)*, pages 1–6. IEEE. https://doi.org/10.1109/IVCNZ.2018.8634778.

Brambell, F. W. R. (1965). Report of the technical committee to enquire into the welfare of animals kept under intensive livestock husbandry systems. http://www.archive.org/details/b3217276x.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1993). Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6. https://doi.org/10.1142/S0218001493000339.

Brust, C.-A., Burghardt, T., Groenenberg, M., Kading, C., Kuhl, H. S., Manguette, M. L., and Denzler, J. (2017). Towards automated visual monitoring of individual gorillas in the wild. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2820–2830. IEEE. https://doi.org/10.1109/ICCVW.2017.333.

Bryden, W. L., Li, X., Ruhnke, I., Zhang, D., and Shini, S. (2021). Nutrition, feeding and laying hen welfare. *Animal Production Science*, 61(10):893–914. https://doi.org/10.1071/AN20396.

Campe, A., Hoes, C., Koesters, S., Froemke, C., Bougeard, S., Staack, M., Bessei, W., Manton, A., Scholz, B., Schrader, L., et al. (2018). Analysis of the influences on plumage condition in laying hens: How suitable is a whole body plumage score as an outcome? *Poultry science*, 97(2):358–367. https://doi.org/10.3382/ps/pex321.

Carroll, J. M., Hamm, R. L., Hagen, J. M., Davis, C. A., and Guthery, F. S. (2017). Eval-

uation of leg banding and attachment of radio–transmitters on ring–necked pheasant chicks. *Wildlife Biology*, 2017(1):1–6. https://doi.org/10.2981/wlb.00263.

Charpentier, B., Zügner, D., and Günnemann, S. (2020). Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in neural information processing systems*. https://doi.org/10.48550/arXiv.2006.09239.

Chen, K., Yao, L., Zhang, D., Wang, X., Chang, X., and Nie, F. (2020). A semisupervised recurrent convolutional attention model for human activity recognition. *IEEE transactions on neural networks and learning systems*, 31(5):1747–1756. https://doi.org/10.1109/TNNLS.2019.2927224.

Chen, M.-S., Lin, J.-Q., Li, X.-L., Liu, B.-Y., Wang, C.-D., Huang, D., and Lai, J.-H. (2022a). Representation learning in multi-view clustering: A literature review. *Data Science and Engineering*, 7(3):225–241. https://doi.org/10.1007/s41019-022-00190-8.

Chen, X., Yang, T., Mai, K., Liu, C., Xiong, J., Kuang, Y., and Gao, Y. (2022b). Holstein cattle face re-identification unifying global and part feature deep network with attention mechanism. *Animals*, 12(8).

Chen, Z., Li, A., and Wang, Y. (2019). A temporal attentive approach for video-based pedestrian attribute recognition. In *Pattern Recognition and Computer Vision: Second Chinese Conference, PRCV 2019, Xi'an, China, November 8–11, 2019, Proceedings, Part II 2*, pages 209–220. Springer. https://doi.org/10.48550/arXiv.1901.05742.

Chien, Y.-R. and Chen, Y.-X. (2018). An rfid-based smart nest box: an experimental study of laying performance and behavior of individual hens. *Sensors*, 18(3):859. https://doi.org/10.3390/s18030859.

Cicalese, P. A., Mobiny, A., Shahmoradi, Z., Yi, X., Mohan, C., and van Nguyen, H. (2021). Kidney level lupus nephritis classification using uncertainty guided bayesian convolutional neural networks. *IEEE journal of biomedical and health informatics*, 25(2):315–324. https://doi.org/10.1109/JBHI.2020.3039162.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46. https://doi.org/10.1177/001316446002000104.

Corkery, G. P., Gonzales-Barron, U., Ayalew, G., Ward, S., and McDonnell, K. (2009). A preliminary investigation of avian comb as a potential biometric marker for identification of poultry. *Transactions of the ASABE*, 52(3):991–998. https://doi.org/10.13031/2013.27383.

De Wet, L., Vranken, E., Chedad, A., Aerts, J.-M., Ceunen, J., and Berckmans, D. (2003). Computer-assisted image analysis to quantify daily growth rates of broiler chickens. *British poultry science*, 44(4):524–532. https://doi.org/10.1080/00071660310001616192.

Deb, D., Wiper, S., Gong, S., Shi, Y., Tymoszek, C., Fletcher, A., and Jain, A. K.

(2018). Face recognition: Primates in the wild. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–10. IEEE. https://doi.org/10.1109/BTAS.2018.8698538.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE. https://doi.org/10.1109/CVPR.2009.5206848.

Dennis, R. L., Fahey, A. G., and Cheng, H. W. (2008). Different effects of individual identification systems on chicken well-being. *Poultry science*, 87(6):1052–1057. https://doi.org/10.3382/ps.2007-00240.

Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F., and Udluft, S. (2018). Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, pages 1184–1193. PMLR. https://doi.org/10.48550/arXiv.1710.07283.

Directive, E. (1999). Council directive 99/74/ec of 19 july 1999 laying down minimum standards for the protection of laying hens. *Official journal of the European Communities*, 203:53–57. Available online: http://data.europa.eu/eli/dir/1999/74/oj.

Dixon, L. M. (2008). Feather pecking behaviour and associated welfare issues in laying hens. *Avian Biology Research*, 1(2):73–87. https://doi.org/10.3184/175815508X363251.

Döhring, S., Jung, L., and Andersson, R. (2020). *Gefiederschäden bei Legehennen – automatisierte Erfassung im Praxistest -Technische Mitteilung*. Verlag Eugen Ulmer. https://doi.org/10.1399/eps.2020.317.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. https://doi.org/10.48550/arXiv.2010.11929.

Edwards, L. E. and Hemsworth, P. H. (2021). The impact of management, husbandry and stockperson decisions on the welfare of laying hens in australia. *Animal Production Science*, 61(10):944–967. https://doi.org/10.1071/AN19664.

Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M., and Burgard, W. (2015). Multimodal deep learning for robust rgb-d object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687. IEEE. https://doi.org/10.1109/IROS.2015.7353446.

Ellen, E. D., Van Der Sluis, M., Siegford, J., Guzhva, O., Toscano, M. J., Bennewitz, J., Van Der Zande, L. E., Van Der Eijk, J. A., de Haas, E. N., Norton, T., et al. (2019). Review of sensor technologies in animal breeding: phenotyping behaviors of laying hens to select against feather pecking. *Animals*, 9(3):108. https://doi.org/10.3390/ani9030108.

Everingham, M., Eslami, S. M. A., van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136. https://doi.org/10.1007/s11263-014-0733-5.

Everingham, M., van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338. https://doi.org/10.1007/s11263-009-0275-4.

FAWC (1993). *Report on Priorities for Animal Welfare: Research and Development*. Farm Animal Welfare Council. Available online: https://books.google.de/books?id=W4KDwgEACAAJ.

Feiyang, Z., Yueming, H., Liancheng, C., Lihong, G., Wenjie, D., and Lu, W. (2016). Monitoring behavior of poultry based on rfid radio frequency network. *International Journal of Agricultural and Biological Engineering*, 9(6):139–147. https://doi.org/10.3965/j.ijabe.20160906.1568.

Feng, D., Rosenbaum, L., Glaeser, C., Timm, F., and Dietmayer, K. (2019). Can we trust you? on calibration of a probabilistic object detector for autonomous driving. *arXiv preprint arXiv:1909.12358*. https://doi.org/10.48550/arXiv.1909.12358.

Filos, A., Farquhar, S., Gomez, A. N., Rudner, T. G., Kenton, Z., Smith, L., Alizadeh, M., De Kroon, A., and Gal, Y. (2019). A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv preprint arXiv:1912.10481*. https://doi.org/10.48550/arXiv.1912.10481.

Food and Agriculture Organization of the United Nations (2023). Poultry meat production – FAO. [Dataset]. With major processing by Our World in Data. Available online: https://ourworldindata.org/grapher/global-meat-production-by-livestock-type?tab=table&time=1970..latest (accessed on 15.01.2024).

Freytag, A., Rodner, E., Simon, M., Loos, A., Kühl, H. S., and Denzler, J. (2016). Chimpanzee faces in the wild: Log-euclidean cnns for predicting identities and attributes of primates. In Rosenhahn, B. and Andres, B., editors, *Pattern Recognition*, volume 9796 of *Lecture Notes in Computer Science*, pages 51–63. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-45886-1_5.

Fujii, T., Yokoi, H., Tada, T., Suzuki, K., and Tsukamoto, K. (2009). Poultry tracking system with camera using particle filters. In *2008 IEEE International Conference on Robotics and Biomimetics*, pages 1888–1893. IEEE. https://doi.org/10.1109/ROBIO.2009.4913289.

Fukunaga, K. and Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40. https://doi.org/10.1109/TIT.1975.1055330.

Gaber, T., Tharwat, A., Hassanien, A. E., and Snasel, V. (2016). Bio-

metric cattle identification approach based on weber's local descriptor and adaboost classifier. *Computers and Electronics in Agriculture*, 122:55–66. https://doi.org/10.1016/j.compag.2015.12.022.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML'16: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. https://doi.org/10.48550/arXiv.1506.02142.

Gallardo, R. K. and Sauer, J. (2018). Adoption of labor-saving technologies in agriculture. *Annual Review of Resource Economics*, 10:185–206. https://doi.org/10.1146/annurev-resource-100517-023018.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589. https://doi.org/10.1007/s10462-023-10562-9.

Ge, Z. and Wang, X. (2021). Evaluation of various open-set medical imaging tasks with deep neural networks. *arXiv preprint arXiv:2110.10888*. https://doi.org/10.48550/arXiv.2110.10888.

Geffen, O., Yitzhaky, Y., Barchilon, N., Druyan, S., and Halachmi, I. (2020). A machine vision system to detect and count laying hens in battery cages. *Animal: an international journal of animal bioscience*, 14(12):2628–2634. https://doi.org/10.1017/S1751731120001676.

Ghosh, A., Shanmugalingam, K., and Lin, W.-Y. (2023). Relation preserving triplet mining for stabilising the triplet loss in re-identification systems. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4840–4849. https://doi.org/10.48550/arXiv.2110.07933.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. PMLR. https://proceedings.mlr.press/v9/glorot10a.html.

Gonzales Barron, U., Corkery, G., Barry, B., Butler, F., McDonnell, K., and Ward, S. (2008). Assessment of retinal recognition technology as a biometric method for sheep identification. *Computers and Electronics in Agriculture*, 60(2):156–166. https://doi.org/10.1016/j.compag.2007.07.010.

Gundavarapu, N. B., Srivastava, D., Mitra, R., Sharma, A., and Jain, A. (2019). Structured aleatoric uncertainty in human pose estimation. In *CVPR Workshops*, volume 2, page 2.

Han, Z., Zhang, C., Fu, H., and Zhou, J. T. (2021). Trusted multi-

view classification. In *International Conference on Learning Representations*. https://openreview.net/forum?id=OOsR8BzCnl5.

Han, Z., Zhang, C., Fu, H., and Zhou, J. T. (2022). Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):2551–2566. https://doi.org/10.48550/arXiv.2204.11423.

Hansen, M. F., Smith, M. L., Smith, L. N., Salter, M. G., Baxter, E. M., Farish, M., and Grieve, B. (2018). Towards on-farm pig face recognition using convolutional neural networks. *Computers in Industry*, 98:145–152. https://doi.org/10.1016/j.compind.2018.02.016.

Hartung, J. (2013). A short history of livestock production. In *Livestock housing*, pages 21–34. Wageningen Academic. https://doi.org/10.3920/978-90-8686-771-4_01.

He, K., Gkioxari, G., Dollar, P., and Girshick, R. (2017). Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988. IEEE. https://doi.org/10.1109/ICCV.2017.322.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE. https://doi.org/10.1109/CVPR.2016.90.

He, S., Luo, H., Wang, P., Wang, F., Li, H., and Jiang, W. (2021a). Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022. https://doi.org/10.48550/arXiv.2102.04378.

He, X., Deng, Y., Fang, L., and Peng, Q. (2021b). Multi-modal retinal image classification with modality-specific attention network. *IEEE transactions on medical imaging*, 40(6):1591–1602. https://doi.org/10.1109/TMI.2021.3059956.

Hemsworth, P. H. (2003). Human–animal interactions in livestock production. *Applied Animal Behaviour Science*, 81(3):185–198. https://doi.org/10.1016/S0168-1591(02)00280-0.

Heo, J., Lee, H. B., Kim, S., Lee, J., Kim, K. J., Yang, E., and Hwang, S. J. (2018). Uncertainty-aware attention for reliable interpretation and prediction. *Advances in neural information processing systems*, 31. https://doi.org/10.48550/arXiv.1805.09653.

Hermans, A., Beyer, L., and Leibe, B. (2017). In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*. https://doi.org/10.48550/arXiv.1703.07737.

Hernandez-Ortega, J., Galbally, J., Fierrez, J., Haraksim, R., and Beslay, L. (2019). Faceqnet: Quality assessment for face recognition based on deep learning. In *2019 International Conference on Biometrics*. https://doi.org/10.48550/arXiv.1904.01740.

Hubbard, C. and Scott, K. (2011). Do farmers and scientists differ in their under-

standing and assessment of farm animal welfare? *Animal Welfare*, 20(1):79–87. https://doi.org/10.1017/S0962728600002451.

Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506. https://doi.org/10.1007/s10994-021-05946-3.

Jeon, H. Y., Tian, L. F., and Zhu, H. (2011). Robust crop and weed segmentation under uncontrolled outdoor illumination. *Sensors (Basel, Switzerland)*, 11(6):6270–6283. https://doi.org/10.3390/s110606270.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399. https://doi.org/10.1038/s42256-019-0088-2.

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., Kwon, Y., Michael, K., Fang, J., Yifu, Z., Wong, C., Montes, D., et al. (2022). ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. *Zenodo*. https://doi.org/10.5281/zenodo.3908559.

Jungo, A., Meier, R., Ermis, E., Herrmann, E., and Reyes, M. (2018). Uncertainty-driven sanity check: Application to postoperative brain tumor cavity segmentation. In *Medical Imaging with Deep Learning (MIDL), 2018*. https://doi.org/10.48550/arXiv.1806.03106.

Kashiha, M. A., Green, A. R., Sales, T. G., Bahr, C., Berckmans, D., and Gates, R. S. (2014). Performance of an image analysis processing system for hen tracking in an environmental preference chamber. *Poultry science*, 93(10):2439–2448. https://doi.org/10.3382/ps.2014-04078.

Kawamura, N., Takaya, M., Hayashi, H., and Goto, T. (2023). Housing systems affect eggshell lightness and free amino acid contents of egg albumen in tosa-jidori chickens: A preliminary research. *Animals*, 13(11). https://doi.org/10.3390/ani13111837.

Kendall, A. and Cipolla, R. (2016). Modelling uncertainty in deep learning for camera relocalization. In *ICRA 2016*. https://doi.org/10.48550/arXiv.1509.05909.

Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1703.04977.

Kiela, D., Grave, E., Joulin, A., and Mikolov, T. (2018). Efficient large-scale multimodal classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32. https://doi.org/10.48550/arXiv.1802.02892.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. https://doi.org/10.48550/arXiv.1412.6980.

Kiureghian, A. D. and Ditlevsen, O. (2009). Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112. https://doi.org/10.1016/j.strusafe.2008.06.020.

Knierim, U., Andersson, R., Keppler, C., Petermann, S., Rauch, E., Spindler, B., and

Zapf, R., editors (2016). *Tierschutzindikatoren: Leitfaden für die Praxis - Geflügel: Vorschläge für die Produktionsrichtungen Jung- und Legehenne, Masthuhn, Mastpute*. KTBL, Darmstadt.

Koerkamp, P. G., Keen, A., Van Niekerk, T. G., and Smit, S. (1995). The effect of manure and litter handling and indoor climatic conditions on ammonia emissions from a battery cage and an aviary housing system for laying hens. *Netherlands Journal of Agricultural Science*, 43(4):351–373. https://doi.org/10.18174/njas.v43i4.560.

Kraus, F. and Dietmayer, K. (2019). Uncertainty estimation in one-stage object detection. *IEEE Intelligent Transportation Systems Conference*. https://doi.org/10.48550/arXiv.1905.10296.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90. https://doi.org/10.1145/3065386.

Laakom, F., Raitoharju, J., Iosifidis, A., Nikkanen, J., and Gabbouj, M. (2021). Monte carlo dropout ensembles for robust illumination estimation. In *2021 International Joint Conference on Neural Networks*. https://doi.org/10.48550/arXiv.2007.10114.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30. https://doi.org/10.48550/arXiv.1612.01474.

Lamping, C., Derks, M., Groot Koerkamp, P., and Kootstra, G. (2022). Chickennet - an end-to-end approach for plumage condition assessment of laying hens in commercial farms using computer vision. *Computers and Electronics in Agriculture*, 194. https://doi.org/10.1016/j.compag.2022.106695.

Lamping, C., Derks, M., and Kootstra, G. (2024a). Fuse: A framework for uncertainty-aware object assessment from image sequences in uncontrolled environments. *Submitted to Computers and Electronics in Agriculture*.

Lamping, C., Kootstra, G., and Derks, M. (2023). Uncertainty estimation for deep neural networks to improve the assessment of plumage conditions of chickens. *Smart Agricultural Technology*, 5. https://doi.org/10.1016/j.atech.2023.100308.

Lamping, C., Kootstra, G., and Derks, M. (2024b). Transformer-based similarity learning for re-identification of chickens. *Submitted to Computers and Electronics in Agriculture*.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159. https://doi.org/10.2307/2529310.

Le, M. T., Diehl, F., Brunner, T., and Knol, A. (2018). Uncertainty estimation for deep neural object detectors in safety-critical applications. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3873–3878. IEEE. https://doi.org/10.1109/ITSC.2018.8569637.

Lee, K. Y., Hong, J., and Chung, S. K. (2017). Smart self-cleaning lens cover for miniature cameras of automobiles. *Sensors and Actuators B: Chemical*, 239:754–758. https://doi.org/10.1016/j.snb.2016.08.032.

Lenz, I., Lee, H., and Saxena, A. (2015). Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724. https://doi.org/10.1177/0278364914549607.

Leroy, T., Vranken, E., Struelens, E., Sonck, B., and Berckmans, D. (2005). Computer vision based recognition of behavior phenotypes of laying hens. In *2005 Tampa, FL July 17-20, 2005*, St. Joseph, MI. American Society of Agricultural and Biological Engineers. https://doi.org/10.13031/2013.19471.

Li, G., Hui, X., Lin, F., and Zhao, Y. (2020a). Developing and evaluating poultry preening behavior detectors via mask region-based convolutional neural network. *Animals: an open access journal from MDPI*, 10(10). https://doi.org/10.3390/ani10101762.

Li, N., Ren, Z., Li, D., and Zeng, L. (2020b). Review: Automated techniques for monitoring the behaviour and welfare of broilers and laying hens: towards the goal of precision livestock farming. *Animal: an international journal of animal bioscience*, 14(3):617–625. https://doi.org/10.1017/S1751731119002155.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-10602-1_48.

Liong, V. E., Nguyen, T. N. T., Widjaja, S., Sharma, D., and Chong, Z. J. (2020). Amvnet: Assertion-based multi-view fusion network for lidar semantic segmentation. *arXiv preprint arXiv:2012.04934*. https://doi.org/10.48550/arXiv.2012.04934.

Liu, C., Gu, J., Kim, K., Narasimhan, S., and Kautz, J. (2019). Neural rgb->d sensing: Depth and uncertainty from a video camera. In *Computer Vision and Pattern Recognition 2019*. https://doi.org/10.48550/arXiv.1901.02571.

Lu, X., Lin, Z., Shen, X., Mech, R., and Wang, J. Z. (2015). Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 990–998. IEEE.

Luke, J., Joseph, R., and Balaji, M. (2019). Impact of image size on accuracy and generalization of convolutional neural networks. *International journal of research and analytical reviews (IJRAR)*, 6(1):70–80.

Luo, W., Li, Y., Urtasun, R., and Zemel, R. (2016). Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29. https://doi.org/10.48550/arXiv.1701.04128.

Lyzhov, A., Molchanova, Y., Ashukha, A., Molchanov, D., and Vetrov, D. (2020). Greedy policy search: A simple baseline for learnable test-time augmentation. In *Conference on uncertainty in artificial intelligence*, pages 1308–1317. PMLR. https://doi.org/10.48550/arXiv.2002.09103.

Madzingira, O. (2018). Animal welfare considerations in food-producing animals. *Animal Welfare*, 99:171–179. https://doi.org/10.5772/intechopen.78223.

Marsot, M., Mei, J., Shan, X., Ye, L., Feng, P., Yan, X., Li, C., and Zhao, Y. (2020). An adaptive pig face recognition approach using convolutional neural networks. *Computers and Electronics in Agriculture*, 173:105386. https://doi.org/10.1016/j.compag.2020.105386.

McAdie, T. and Keeling, L. (2000). Effect of manipulating feathers of laying hens on the incidence of feather pecking and cannibalism. *Applied Animal Behaviour Science*, 68(3):215–229. https://doi.org/10.1016/s0168-1591(00)00107-6.

Meier, P. (1953). Variance of a weighted mean. *Biometrics*, 9(1):59. https://doi.org/10.2307/3001633.

Michel, V., Berk, J., Bozakova, N., van der Eijk, J., Estevez, I., Mircheva, T., Relic, R., Rodenburg, T. B., Sossidou, E. N., and Guinebretière, M. (2022). The relationships between damaging behaviours and health in laying hens. *Animals : an open access journal from MDPI*, 12(8). https://doi.org/10.3390/ani12080986.

Milisits, G., Szász, S., Donkó, T., Budai, Z., Almási, A., Pőcze, O., Ujvári, J., Farkas, T. P., Garamvölgyi, E., Horn, P., et al. (2021). Comparison of changes in the plumage and body condition, egg production, and mortality of different non-beak-trimmed pure line laying hens during the egg-laying period. *Animals*, 11(2):500. https://doi.org/10.3390/ani11020500.

Morvant, E., Habrard, A., and Ayache, S. (2014). Majority vote of diverse classifiers for late fusion. In *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, S+ SSPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings*, pages 153–162. Springer. https://doi.org/10.48550/arXiv.1404.7796.

Moskvyak, O., Maire, F., Dayoub, F., Armstrong, A. O., and Baktashmotlagh, M. (2021). Robust re-identification of manta rays from natural markings by learning pose invariant embeddings. In *2021 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE. https://doi.org/10.1109/DICTA52665.2021.9647359.

Neethirajan, S. and Kemp, B. (2021). Digital livestock farming. *Sensing and Bio-Sensing Research*, 32:100408. https://doi.org/10.3390/agriengineering5010032.

Nepovinnykh, E., Eerola, T., and Kalviainen, H. (2020). Siamese network based pelage pattern matching for ringed seal re-identification. In *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 25–34. IEEE. https://doi.org/10.1109/WACVW50321.2020.9096935.

Nepovinnykh, E., Eerola, T., Kälviäinen, H., and Chelak, I. (2024). Norppa: novel ringed seal re-identification by pelage pattern aggregation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1–10. https://doi.org/10.48550/arXiv.2206.02498.

OECD, Food, and of the United Nations, A. O. (2023). *OECD-FAO Agricultural Outlook 2023-2032*. https://doi.org/https://doi.org/10.1787/08801ab7-en.

Okinda, C., Lu, M., Liu, L., Nyalala, I., Muneri, C., Wang, J., Zhang, H., and Shen, M. (2019). A machine vision system for early detection and prediction of sick birds: A broiler chicken model. *Biosystems Engineering*, 188:229–242. https://doi.org/10.1016/j.biosystemseng.2019.09.015.

Okinda, C., Nyalala, I., Korohou, T., Okinda, C., Wang, J., Achieng, T., Wamalwa, P., Mang, T., and Shen, M. (2020). A review on computer vision systems in monitoring of poultry: A welfare perspective. *Artificial Intelligence in Agriculture*, 4:184–208. https://doi.org/10.1016/j.aiia.2020.09.002.

Ophoff, T., van Beeck, K., and Goedemé, T. (2019). Exploring rgb+depth fusion for real-time object detection. *Sensors (Basel, Switzerland)*, 19(4). https://doi.org/10.3390/s19040866.

Otte, J., Roland-Holst, D., Pfeiffer, D., Soares-Magalhaes, R., Rushton, J., Graham, J., and Silbergeld, E. (2007). Industrial livestock production and global health risks. *Food and Agriculture Organization of the United Nations, Pro-Poor Livestock Policy Initiative Research Report*.

Pei, W., Baltrusaitis, T., Tax, D. M., and Morency, L.-P. (2017). Temporal attention-gated model for robust sequence classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6730–6739. https://doi.org/10.48550/arXiv.1612.00385.

Peng, Y., Zhao, Y., and Zhang, J. (2018). Two-stream collaborative learning with spatial-temporal attention for video classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(3):773–786. https://doi.org/10.48550/arXiv.1711.03273.

Pires, C., Barandas, M., Fernandes, L., Folgado, D., and Gamboa, H. (2020). Towards knowledge uncertainty estimation for open set recognition. *Machine Learning and Knowledge Extraction*, 2(4):505–532. https://doi.org/10.3390/make2040028.

Postels, J., Ferroni, F., Coskun, H., Navab, N., and Tombari, F. (2019). Sampling-free epistemic uncertainty estimation using approximated variance propagation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2931–2940. https://doi.org/10.48550/arXiv.1908.00598.

Prado, A., Kausik, R., and Venkataramanan, L. (2019). Dual neural network architecture for determining epistemic and aleatoric uncertainties. *arXiv preprint arXiv:1910.06153*. https://doi.org/10.48550/arXiv.1910.06153.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788. Available online: https://pjreddie.com/darknet/yolo/.

Ren, J., Shen, X., Lin, Z., and Mech, R. (2020). Best frame selection in a short video. In *2020 IEEE Winter Conference*, pages 3201–3210. https://doi.org/10.1109/WACV45572.2020.9093615.

Rodenburg, T. B., van Krimpen, M. M., de Jong, I. C., de Haas, E. N., Kops, M. S., Riedstra, B. J., Nordquist, R. E., Wagenaar, J. P., Bestman, M., and Nicol, C. J. (2013). The prevention and control of feather pecking in laying hens: identifying the underlying principles. *World's Poultry Science Journal*, 69(2):361–374. https://doi.org/10.1017/S0043933913000354.

Rowe, E., Dawkins, M. S., and Gebhardt-Henrich, S. G. (2019). A systematic review of precision livestock farming in the poultry sector: Is technology focussed on improving bird welfare? *Animals*, 9(9):614. https://doi.org/10.3390/ani9090614.

Roy, A. G., Conjeti, S., Navab, N., and Wachinger, C. (2019). Bayesian quicknat: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage*, 195. https://doi.org/10.48550/arXiv.1811.09800.

Ryan, M. (2023). Labour and skills shortages in the agro-food sector. https://doi.org/10.1787/ed758aab-en.

Sabottke, C. F. and Spieler, B. M. (2020). The effect of image resolution on deep learning in radiography. *Radiology. Artificial intelligence*, 2(1):e190015. https://doi.org/10.1148/ryai.2019190015.

Sadiq, B. O., Muhammad, B., Abdullahi, M. N., Onuh, G., Muhammed, A. A., and Babatunde, A. E. (2020). Keyframe extraction techniques: A review. *ELEKTRIKA-Journal of Electrical Engineering*, 19(3):54–60. https://doi.org/10.2174/2213275911666180719111118.

Schilling, B., Bahmani, K., Li, B., Banerjee, S., Smith, J. S., Moshier, T., and Schuckers, S. (2018). Validation of biometric identification of dairy cows based on udder nir images. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7. IEEE. https://doi.org/10.1109/BTAS.2018.8698553.

Schneider, S., Taylor, G. W., Linquist, S., and Kremer, S. C. (2019). Past, present and future approaches using computer vision for animal re–identification from camera trap data. *Methods in Ecology and Evolution*, 10(4):461–470. https://doi.org/10.1111/2041-210X.13133.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823. IEEE.

https://doi.org/10.1109/CVPR.2015.7298682.

Seeland, M. and Mäder, P. (2021). Multi-view classification with convolutional neural networks. *PloS one*, 16(1):e0245230. https://doi.org/10.1371/journal.pone.0245230.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626. https://doi.org/10.1109/ICCV.2017.74.

Sensoy, M., Kaplan, L., and Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31. https://doi.org/10.48550/arXiv.1806.01768.

Sergeant, D., Boyle, R., and Forbes, M. (1998). Computer visual tracking of poultry. *Computers and Electronics in Agriculture*, 21(1):1–18. https://doi.org/10.1016/S0168-1699(98)00025-8.

Shao, W., Kawakami, R., Yoshihashi, R., You, S., Kawase, H., and Naemura, T. (2020). Cattle detection and counting in uav images based on convolutional neural networks. *International Journal of Remote Sensing*, 41(1):31–52. https://doi.org/10.1080/01431161.2019.1624858.

Sihalath, T., Basak, J. K., Bhujel, A., Arulmozhi, E., Moon, B. E., and Kim, H. T. (2021). Pig identification using deep convolutional neural network based on different age range. *Journal of Biosystems Engineering*, 46(2):182–195. https://doi.org/10.1007/s42853-021-00098-7.

Steinfeld, H., Gerber, P., Wassenaar, T. D., Castel, V., and De Haan, C. (2006). *Livestock's long shadow: environmental issues and options*. Food & Agriculture Org.

Stygar, A. H., Gómez, Y., Berteselli, G. V., Dalla Costa, E., Canali, E., Niemi, J. K., Llonch, P., and Pastell, M. (2021). A systematic review on commercially available and validated sensor technologies for welfare assessment of dairy cattle. *Frontiers in veterinary science*, 8. https://doi.org/10.3389/fvets.2021.634338.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR. https://doi.org/10.48550/arXiv.1905.11946.

Teletchea, F. (2019). *Animal domestication: A brief overview*. IntechOpen. https://doi.org/10.5772/intechopen.86783.

Thobe, P., Chibanda, C., and Behrendt, L. (2021). Steckbriefe zur tierhaltung in deutschland: Mastgeflügel. *Thünen-Institut für Betriebswirtschaft: Braunschweig, Germany*. Available online: https://literatur.thuenen.de/digbib_extern/dn064308.pdf (accessed on 20.01.2024).

Tian, J., Cheung, W., Glaser, N., Liu, Y.-C., and Kira, Z. (2020a). Uno: Uncertainty-

aware noisy-or multimodal fusion for unanticipated input degradation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5716–5723. IEEE. https://doi.org/10.1109/ICRA40945.2020.9197266.

Tian, Y., Krishnan, D., and Isola, P. (2020b). Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer. https://doi.org/10.1007/978-3-030-58621-8_45.

Tian, Y., Li, D., and Xu, C. (2020c). Unified multisensory perception: Weakly-supervised audio-visual video parsing. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision – ECCV 2020*, volume 12348 of *Lecture Notes in Computer Science*, pages 436–454. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-58580-8_26.

Tian, Y., Zorron Cheng Tao Pu, L., Liu, Y., Maicas, G., Verjans, J. W., Burt, A. D., Shin, S. H., Singh, R., and Carneiro, G. (2024). Chapter 15 - detecting, localizing and classifying polyps from colonoscopy videos using deep learning. In *Deep Learning for Medical Image Analysis (Second Edition)*, The MICCAI Society book Series, pages 425–450. Academic Press, second edition edition. https://doi.org/10.1016/B978-0-32-385124-4.00026-X.

Tilbrook, A. J. and Fisher, A. D. (2021). Stress, health and the welfare of laying hens. *Animal Production Science*, 61(10):931–943. https://doi.org/10.1071/an19666.

van Veen, L. A., van den Oever, A. C., Kemp, B., and van den Brand, H. (2023). Perception of laying hen farmers, poultry veterinarians, and poultry experts regarding sensor-based continuous monitoring of laying hen health and welfare. *Poultry Science*, 102(5). https://doi.org/10.1016/j.psj.2023.102581.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. https://doi.org/10.48550/arXiv.1706.03762.

Verdoja, F. and Kyrki, V. (2021). Notes on the behavior of mc dropout. In *ICML Workshop on Uncertainty & Robustness in Deep Learning*. https://doi.org/10.48550/arXiv.2008.02627.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., and Vercauteren, T. (2019). Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*. https://doi.org/10.48550/arXiv.1807.07356.

Wang, J., Wang, N., Li, L., and Ren, Z. (2020). Real-time behavior detection and judgment of egg breeders based on yolo v3. *Neural Computing and Applications*, 32(10):5471–5481. https://doi.org/10.1007/s00521-019-04645-4.

Wang, M., Larsen, M. L., Liu, D., Winters, J. F., Rault, J.-L., and Norton, T. (2022a).

Towards re-identification for long-term tracking of group housed pigs. *Biosystems Engineering*, 222:71–81. https://doi.org/10.1016/j.biosystemseng.2022.07.017.

Wang, S., Liu, J., Yu, G., Liu, X., Zhou, S., Zhu, E., Yang, Y., Yin, J., and Yang, W. (2022b). Multiview deep anomaly detection: A systematic exploration. *IEEE transactions on neural networks and learning systems*, PP. https://doi.org/10.1109/TNNLS.2022.3184723.

Wang, Y., Geng, Z., Jiang, F., Li, C., Wang, Y., Yang, J., and Lin, Z. (2021). Residual relaxation for multi-view representation learning. *Advances in Neural Information Processing Systems*, 34:12104–12115. https://doi.org/10.48550/arXiv.2110.15348.

Wang, Y., Xu, X., Wang, Z., Li, R., Hua, Z., and Song, H. (2023). Shufflenet-triplet: A lightweight re-identification network for dairy cows in natural scenes. *Computers and Electronics in Agriculture*, 205. https://doi.org/10.1016/j.compag.2023.107632.

Wang, Z. and Liu, T. (2022). Two-stage method based on triplet margin loss for pig face recognition. *Computers and Electronics in Agriculture*, 194. https://doi.org/10.1016/j.compag.2022.106737.

Wei, X., Zhang, T., Li, Y., Zhang, Y., and Wu, F. (2020). Multi-modality cross attention network for image and sentence matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10938–10947. IEEE. https://doi.org/10.1109/CVPR42600.2020.01095.

Wojke, N. and Bewley, A. (2018). Deep cosine metric learning for person re-identification. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 748–756. IEEE. https://doi.org/10.1109/WACV.2018.00087.

Wojke, N., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE. https://doi.org/10.1109/ICIP.2017.8296962.

Wu, D., Ying, Y., Zhou, M., Pan, J., and Cui, D. (2023). Improved resnet-50 deep learning algorithm for identifying chicken gender. *Computers and Electronics in Agriculture*, 205. https://doi.org/10.1016/j.compag.2023.107622.

Xuan, H., Stylianou, A., and Pless, R. (2020). Improved embeddings with easy positive triplet mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2474–2482. https://doi.org/10.1109/WACV45572.2020.9093432.

Yang, X., Bist, R., Subedi, S., Wu, Z., Liu, T., and Chai, L. (2023). An automatic classifier for monitoring applied behaviors of cage-free laying hens with deep learning. *Engineering Applications of Artificial Intelligence*, 123:106377. https://doi.org/10.1016/j.engappai.2023.106377.

Yang, Z., Xiong, H., Chen, X., Liu, H., Kuang, Y., and Gao, Y. (2019). Dairy cow tiny

face recognition based on convolutional neural networks. In Sun, Z., He, R., Feng, J., Shan, S., and Guo, Z., editors, *Biometric Recognition*, volume 11818 of *Lecture Notes in Computer Science*, pages 216–222. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-31456-9_24.

Zhang, X., Diakogiannis, F. I., Dodson, R., and Wicenec, A. (2022). Radio transient detection with closure products and machine learning. *Instrumentation and Methods for Astrophysics*. https://doi.org/10.48550/arXiv.2204.01958.

Zhang Feiyang, Hu Yueming, Chen Liancheng, Guo Lihong, Duan Wenjie, and Wang Lu (2016). Monitoring behavior of poultry based on rfid radio frequency network. *International Journal of Agricultural and Biological Engineering*, 9(6):139–147. https://doi.org/10.25165/ijabe.v9i6.1568.

Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., and Tian, Q. (2016). Mars: A video benchmark for large-scale person re-identification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 868–884. Springer. https://doi.org/10.1007/978-3-319-46466-4_52.

Zhu, X., Chen, C., Zheng, B., Yang, X., Gan, H., Zheng, C., Yang, A., Mao, L., and Xue, Y. (2020). Automatic recognition of lactating sow postures by refined two-stream rgb-d faster r-cnn. *Biosystems Engineering*, 189:116–132. https://doi.org/10.1016/j.biosystemseng.2019.11.013.

Zhuang, X., Bi, M., Guo, J., Wu, S., and Zhang, T. (2018). Development of an early warning algorithm to detect sick broilers. *Computers and Electronics in Agriculture*, 144:102–113. https://doi.org/10.1016/j.compag.2017.11.032.

Zhuang, Y., Rui, Y., Huang, T. S., and Mehrotra, S. (1998). Adaptive key frame extraction using unsupervised clustering. In *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269)*, pages 866–870. IEEE Comput. Soc. https://doi.org/10.1109/ICIP.1998.723655.

Summary

The steadily growing world population and changing consumer behavior have significantly increased the global demand for animal products in recent decades. This trend is expected to continue in the coming years, with the demand for poultry products, in particular, driving this development. A consequence and simultaneous reinforcement of this trend is the increasing intensification and industrialization of animal husbandry, which promotes the affordable availability of animal products. In this context, technological innovations, especially the automation of individual processes, play a crucial role, as they can address current challenges such as the shortage of skilled workers in the agricultural sector and biosecurity concerns.

While automated solutions for feed and water supply or waste transportation are now standard in modern livestock facilities, regular animal monitoring typically remains a manual task. As a result, the quality of animal inspection heavily depends on the qualifications, experience, and motivation of the respective inspector. In addition, it usually relies on samples of individual animals taken at regular intervals. This approach does not allow for continuous monitoring of the entire herd, making it difficult to intervene quickly in emergencies. Given the high importance of monitoring for both animal welfare and productivity, there is a huge potential for optimization. While wearable sensors for continuous monitoring of vital parameters, especially in the dairy sector, have already been established, their cost makes them economically unfeasible for implementation in large herds, such as in the poultry sector.

In this regard, there is significant potential in utilizing camera-based systems for non-contact monitoring of animals, which eliminates the necessity for individual sensors for each animal. Additionally, cameras can monitor multiple parameters simultaneously. The rise of deep learning methods in image processing further enables complex tasks, such as evaluating behavioral patterns or detecting sick chickens. Although a variety of prototypes for various monitoring tasks in animal husbandry have been developed in the past, they have so far seen little practical application. One main reason for this is the lack of reliability of image-based methods under the challenging environmental conditions in commercial livestock farms. These include factors such as varying illumination, contamination of cameras, or changes in the visual appearance of the animals. To become established in

practice, monitoring systems must be robust against these influences. Moreover, the predictions of the underlying models must be transparent and reliable so that management decisions for the welfare of the animals can be made based on them.

Therefore, the aim of this thesis was to develop robust methods for image-based animal monitoring that are specifically designed for practical use. While the use of deep learning algorithms was a central element of this thesis, plumage conditions scoring of laying hens served as a representative case for the development and evaluation of the various methods. A detailed explanation of feather scoring as a relevant and non-trivial task in poultry farming can be found in the general introduction in Chapter 1. As a result of this research, a monitoring framework has been developed, consisting of four different sub-modules, detailed in Chapters 2 to 5.

The deep learning model ChickenNet, which was the focus of Chapter 2, can be seen as the backbone of the entire framework. This model enables the recognition and simultaneous scoring of individual animals, aiming to detect individual hens within a larger flock while providing a robust assessment of their feather condition. Additionally, the approach aimed to be easily transferable to other applications where the recognition and assessment of an animal with regard to a specific characteristic is required. This was implemented using a neural network, which has outputs for recognition and segmentation, as well as a regression output for predicting a numerical value. For the training of this model, image data from a commercial layer farm was used, supplemented with artificial effects to simulate adverse environmental conditions. In the experiments of this chapter, the impact of image resolution on the model's accuracy was evaluated and it was investigated whether the use of depth enhances recognition and evaluation accuracy. This is particularly relevant for practical applications because increased model complexity results in higher computational requirements, leading to increased application costs. Here, it was shown that a resolution of 896x896 pixels yielded the best results, with a detection accuracy of over 98% and a feather condition assessment accuracy of 92%. Additional utilization of depth information did not result in a general enhancement in accuracies.

Chapter 3 dealt with the estimation of uncertainties in the predictions of the ChickenNet model. The aim was to extend the model so that, in addition to predicting the assessment scores, it can also predict the uncertainty associated with the assessment. This addressed a weakness of classic deep learning models, which do not provide any indication of the reliability of the given prediction. Especially under adverse environmental conditions, such as those prevailing in livestock facilities, influencing factors such as high animal density or camera contamination can affect the predictions of the model, making them unreliable. To enable reliable animal monitoring, it is crucial to identify these uncertain predictions. In this chapter, three different methods for estimating uncertainty were integrated into the original model and compared against each other. The first approach was based on an indirect estimation of the assessment uncertainty by predicting the level of occlusion

for each detected animal. This approach was based on the assumption that assessing a heavily occluded animal is more uncertain than that of a fully visible one. In addition, two methods were integrated for the direct estimation of epistemic and aleatory uncertainty. Epistemic uncertainty refers to uncertainty arising from an inadequate prediction model, while aleatoric uncertainty stems from the quality of the input data. For all three estimation methods, it was hypothesized that the estimated uncertainties would be higher for incorrect assessments compared to correct assessments. Furthermore, it was hypothesized that rejecting uncertain predictions could enhance the overall accuracy of assessments. To evaluate the generalizability of the developed methods beyond the specific use case of plumage condition assessment, they were also tested on an image dataset for human age estimation. Results from both datasets indicated that the estimation of aleatoric and epistemic uncertainties correlated with assessment errors, with higher uncertainties estimated for incorrect predictions than for correct ones. It was further shown that the overall assessment accuracy could be improved by ignoring predictions with high epistemic or aleatoric uncertainty. In contrast, neither hypothesis was confirmed when uncertainty was estimated indirectly through the prediction of occlusion levels.

In Chapter 4, the monitoring framework was enhanced to utilize entire image sequences instead of single images for assessment. The goal here was to combine information from multiple individual frames within a sequence while ignoring irrelevant frames to achieve precise and reliable predictions. The focus of the chapter was on the development of FUSE, an approach to combine multiple predictions that are weighted based on their respective uncertainties. This approach aimed to utilize only informative frames within a sequence, excluding those with occlusions, blurriness, or other issues that could hinder accurate assessment. In this study, the ChickenNet model was utilized to detect and assess the animals, while the previously developed methods for estimating aleatoric and epistemic uncertainty were integrated to weight the individual predictions. However, FUSE was designed to allow for the use of other models and uncertainty estimators to ensure transferability to other monitoring applications. In this chapter, the hypothesis was that assessments of feather condition made on entire sequences using FUSE would be more accurate than those made based on single images. This hypothesis was evaluated by directly comparing both approaches. Furthermore, the experiments investigated how the number of available frames per sequence and the number of sequences per animal affect the accuracy of the assessment. In line with Chapter 3, the method was tested for both the primary application of feather assessment and for human age estimation from images. Results of this study showed that the use of image sequences by FUSE increased assessment accuracy by up to 7% for both datasets compared to conventional single-image assessment. This improvement was also observed when using a reduced number of frames per sequence and a limited number of sequences per animal or person.

Chapter 5 explored the use of deep learning for the re-identification of individual laying hens based on images of their heads. Such re-identification is relevant when different de-

tections within an image sequence must be assigned to a specific animal. Furthermore, it is a prerequisite for any long-term monitoring to address temporal interruptions between individual images. The goal of the approach presented in this chapter was to develop a re-identification method that can be applied in large animal herds, not just for chickens, and does not require manual handling of the animals. For this purpose, a transformer-based neural network was utilized to learn representations from input images. The model was trained in such a way that representations of images of the same animal exhibit a high degree of similarity, while images of different animals result in distinct representations. This representation learning approach enables training the re-identification method without requiring individual training images for each animal. In this chapter, it was hypothesized that the transformer-based method enables the re-identification of individual laying hens while surpassing the performance of traditional CNN-based architectures. To test this hypothesis, the re-identification accuracies of different configurations for both types of architectures were compared. Additionally, the impact of the number of images per chicken and the overall population size on re-identification accuracy was analyzed to evaluate the practical applicability in commercial livestock farms. As this was the first deep learning approach for the re-identification of chickens, it was also investigated which visual features in the head area of the animals are relevant for re-identification by the different models. The experiments demonstrated the general feasibility of re-identifying chickens using the presented method. Transformer-based models outperformed the CNN architectures, achieving a re-identification accuracy of over 90% for small groups of up to five animals. For groups of 100 different chickens, an accuracy of approximately 40% was determined. While this level of accuracy enables the method to be applied to a small number of animals or to allocate detections within an image sequence, it is not currently adequate for commercial use in flocks containing several thousand animals. Despite this, the experiments revealed that the evaluated models learned to prioritize features such as the comb, wattles, and earlobes, often aligning with human perception.

In summary, it can be stated that the primary objective of this thesis, to develop robust methods for image-based animal monitoring tailored for practical applications, has been accomplished through the framework presented. In addition to enhanced resilience against environmental factors in commercial livestock farms, the incorporation of mechanisms for uncertainty assessment into the prediction models was a key contribution of this work. This enhanced the transparency of a monitoring system while also enabling the filtering of unreliable predictions to increase prediction accuracy. A detailed discussion of each chapter's contribution to the overarching goal of the work was conducted in Chapter 6. The methods presented in this work offer a promising foundation for automated animal monitoring in commercial settings. However, they do not yet provide a fully deployable plug-and-play solution for all possible use cases. Therefore, the establishment of automated monitoring systems in practice requires further developments beyond mere software optimization, as also discussed in Chapter 6.

Zusammenfassung

Die stetig wachsende Weltbevölkerung und ein sich veränderndes Konsumverhalten haben die weltweite Nachfrage nach tierischen Produkten in den vergangenen Jahrzehnten drastisch erhöht. Dieser Trend zeichnet sich auch für die kommenden Jahre ab, wobei insbesondere die Nachfrage nach Geflügelprodukten als Haupttreiber dieser Entwicklung zu sehen ist. Eine Folge dieses Trends, die gleichzeitig auch eine Verstärkung darstellt, ist die zunehmende Intensivierung und Industrialisierung der Tierhaltung, welche die preiswerte Verfügbarkeit tierischer Produkte fördert. Eine entscheidende Rolle spielen dabei insbesondere technologische Innovationen und die Automatisierung einzelner Prozesse, mit denen akute Herausforderungen wie der Fachkräftemangel im landwirtschaftlichen Sektor sowie die Biosicherheit adressiert werden können.

Während automatisierte Lösungen für die Futter- und Wasserversorgung oder den Reststofftransport heutzutage Standard in modernen Nutztierstellen sind, gehört die regelmäßige Kontrolle der Tiere üblicherweise zu den Aufgaben, die noch ausschließlich manuell ausgeführt werden. Dies führt dazu, dass die Qualität der Tierkontrolle stark von der Qualifikation, Erfahrung und Motivation des jeweiligen Prüfers abhängt. Zudem basiert sie häufig auf Stichproben einzelner Tiere, die in regelmäßigen Abständen durchgeführt werden. Diese Vorgehensweise ermöglicht keine permanente Überwachung der gesamten Herde und erschwert ein schnelles Eingreifen in Notfällen. Angesichts der hohen Bedeutung des Monitorings sowohl für das Tierwohl als auch für die Produktivität der Tiere ergeben sich hier klare Verbesserungsmöglichkeiten. Zwar haben sich vom Tier getragene Sensoren zur permanenten Überwachung von Vitalparametern insbesondere im Milchviehbereich bereits etabliert, diese sind jedoch aus ökonomischer Sicht für den Einsatz in großen Herden wie im Geflügelbereich kaum relevant.

Erhebliches Potenzial bietet hingegen die Nutzung von kamera-basierten Systemen zur berührungslosen Überwachung der Tiere, wodurch nicht für jedes Tier ein separater Sensor benötigt wird. Gleichzeitig können durch Kameras verschiedene Parameter gleichzeitig überwacht werden. Insbesondere der zunehmende Einsatz von Deep-Learning Methoden in der Bildverarbeitung erlaubt die Umsetzung komplexer Aufgaben, wie beispielsweise die Bewertung von Verhaltensmustern oder die Erkennung kranker Hühner. Obwohl in der Vergangenheit bereits eine Vielzahl von Prototypen für ver-

schiedenste Monitoring-Aufgaben in der Tierhaltung entwickelt wurden, finden diese in der Praxis bislang kaum Verbreitung. Ein Hauptgrund dafür ist die fehlende Zuverlässigkeit der bildbasierten Verfahren unter den herausfordernden Umgebungsbedingungen in kommerziellen Nutztierställen. Dazu gehören beispielsweise Faktoren wie wechselhafte Lichtbedingungen, Verschmutzung der Kamerasysteme oder Veränderungen im optischen Erscheinungsbild der Tiere. Um sich im praktischen Einsatz zu etablieren, müssen Monitoringsysteme robust gegenüber diesen Einflüssen sein. Gleichzeitig müssen die Vorhersagen der zugrundeliegenden Modelle transparent und verlässlich sein, damit auf ihrer Grundlage Managemententscheidungen zum Wohl der Tiere getroffen werden können.

Das Ziel dieser Doktorarbeit bestand daher darin, robuste Methoden für bildbasiertes Tiermonitoring zu entwickeln, die speziell für den Einsatz im praktischen Kontext ausgelegt sind. Ein Schwerpunkt lag dabei auf der Nutzung moderner Deep-Learning-Algorithmen. Als repräsentativer Anwendungsfall für die Entwicklung und Evaluierung der verschiedenen Methoden diente die Gefiederbonitur von Legehennen, welche eine wichtige und gleichzeitig nichttriviale Aufgabe in der Geflügelhaltung darstellt. Eine detaillierte Erläuterung der Gefiederbonitur und deren Relevanz als Anwendungsfall dieser Arbeit ist in der allgemeinen Einleitung in Kapitel 1 zu finden. Als Ergebnis der Arbeit ist ein Monitoring-Framework entstanden, das aus insgesamt vier verschiedenen Teilmodulen besteht, deren Entwicklung in den Kapiteln 2 bis 5 detailliert betrachtet wurde.

Als Rückgrat des gesamten Frameworks kann dabei das Deep-Learning Modell ChickenNet gesehen werden, welches in Kapitel 2 im Fokus stand. Dieses Modell erlaubt die Erkennung und gleichzeitige Bonitur einzelner Tiere. Ziel war es hier, einzelne Hennen innerhalb einer größeren Herde zu detektieren und gleichzeitig eine robuste Bewertung des Gefiederzustands zu liefern. Ebenso sollte eine einfache Übertragbarkeit des Ansatzes auf andere Anwendungen, bei denen die Erkennung und Bewertung eines Tieres in Bezug auf eine bestimmte Eigenschaft erforderlich ist, gewährleistet sein. Dies wurde mittels eines neuronalen Netzes umgesetzt, welches sowohl über Ausgaben für die Erkennung und Segmentierung als auch über eine Regressionsausgabe zur Vorhersage eines numerischen Wertes verfügt. Für das Training dieses Modells wurden Bilddaten aus einer kommerziellen Legehennenfarm genutzt, die darüber hinaus mit künstlichen Effekten ergänzt wurden, um widrige Umgebungseinflüsse abzubilden. In den Experimenten dieses Kapitels wurde der Einfluss der Bildauflösung auf die Genauigkeit des Modells evaluiert und untersucht, ob die Nutzung von Tiefenbildern zur Verbesserung der Erkennungs- und Bewertungsgenauigkeit führt. Dies ist insbesondere im Hinblick auf praktische Anwendungen relevant, da eine höhere Komplexität des Modells zu einem höheren Rechenaufwand und damit gesteigerten Anwendungskosten führen. Hier lieferte eine Auflösung von 896x896 Pixeln die besten Ergebnisse, wobei eine Erkennungsgenauigkeit von über 98% festgestellt wurde, während der Gefiederzustand mit einer Genauigkeit von 92% bewertet werden konnte. Eine zusätzliche Nutzung von Tiefeninformationen führte nicht zu einer generellen

Verbesserung der Genauigkeiten.

Kapitel 3 thematisierte die Schätzung von Unsicherheiten in den Vorhersagen des ChickenNet Models. Ziel war es, das Modell so zu erweitern, dass neben der eigentlichen Bewertung auch eine Unsicherheit dieser Bewertung vorhergesagt werden kann. Damit wurde eine Schwachstelle klassischer Deep-Learning Modelle adressiert, die keine Aussage zur Zuverlässigkeit der gegebenen Vorhersage zulassen. Insbesondere unter widrigen Umgebungsbedingungen, wie sie in Nutztierställen vorherrschen, können Einflussfaktoren wie eine hohe Tierdichte oder die Verschmutzung der Kamera jedoch die Vorhersagen des Modells beeinträchtigen und unzuverlässig machen. Um ein verlässliches Tiermonitoring zu ermöglichen, gilt es diese unsicheren Vorhersagen zu identifizieren. Insgesamt wurden dazu drei verschiedene Methoden zur Unsicherheitsschätzung in das Originalmodel integriert und verglichen. Der erste Ansatz basierte auf einer indirekten Schätzung der Bewertungsunsicherheit durch die Vorhersage der Verdeckungsgrads der erkannten Tiere. Dies beruhte auf der Annahme, dass die Beurteilung eines stark verdeckten Tieres unsicherer ist als die eines vollständig sichtbaren Tieres. Zusätzlich wurden zwei Methoden zur direkten Schätzung von epistemischer und aleatorischer Unsicherheit integriert. Epistemische Unsicherheit bezeichnet dabei die Unsicherheit, die auf ein unzureichendes Vorhersagemodell zurückzuführen ist, während sich die aleatorische Unsicherheit aus der Qualität der Input-Daten ergibt. Für alle drei Schätzverfahren wurde die Hypothese aufgestellt, dass für falsche Vorhersagen des ChickenNet Modells eine höhere Unsicherheit geschätzt würde als für richtige Vorhersagen. Weiterhin wurde hypothesiert, dass das Ignorieren von unsicheren Vorhersagen die allgemeine Bewertungsgenauigkeit verbessern würde. Um die Generalisierbarkeit der Methoden über den Anwendungsfall der Gefiederbewertung hinaus zu evaluieren, wurden sie zusätzlich auf einem Datensatz zu menschlichen Altersbestimmung anhand von Fotos getestet. Im Ergebnis zeigte sich, dass die Schätzung aleatorischer und epistemischer Unsicherheit mit dem Bewertungsfehler korrelierte und somit die geschätzte Unsicherheit für falsche Vorhersagen höher war als für richtige. Ebenso konnte gezeigt werden, dass die Bewertungsgenauigkeit durch das Ignorieren von Vorhersagen mit hoher epistemischer oder aleatorischer Unsicherheit verbessert wurde. Im Gegensatz dazu wurden beide Hypothesen nicht bestätigt, wenn die Unsicherheiten indirekt durch die Vorhersage des Verdeckungsgrads geschätzt wurden. Diese Ergebnisse konnten sowohl für den Anwendungsfall der Gefiederbewertung als auch für die Altersermittlung demonstriert werden.

In Kapitel 4 wurde das Monitoring-Framework um die Möglichkeit ergänzt, ganze Bildsequenzen anstelle von Einzelbildern für eine Bewertung zu nutzen. Hier bestand das Ziel darin, Informationen aus mehreren Einzelbildern zu kombinieren und gleichzeitig unbrauchbare Bilder innerhalb einer Sequenz zu ignorieren, um so präzise und verlässliche Vorhersagen zu erreichen. Im Fokus des Kapitels stand dabei die Entwicklung von FUSE, einem Ansatz zur Kombination mehrerer unabhängiger Vorhersagen, die anhand ihrer jeweiligen Unsicherheit priorisiert werden. Dieser Ansatz beabsichtigte, dass einzelne

Bilder einer Sequenz, in denen das Tier beispielsweise verdeckt oder nur verschwommen zu sehen ist, keinen Einfluss auf die finale Bewertung dieses Tieres haben und stattdessen aussagekräftige Aufnahmen innerhalb der Sequenz berücksichtigt werden. Dazu wurden in dieser Arbeit das ChickenNet Modell zur Erkennung und Bewertung der Tiere sowie die zuvor entwickelten Methoden zur Schätzung von aleatorischer und epistemischer Unsicherheit eingesetzt. FUSE wurde jedoch so konzipiert, dass es auch die Verwendung anderer Modelle und Unsicherheitsschätzer ermöglicht, um die Übertragbarkeit auf andere Monitoring-Anwendungen zu gewährleisten. In diesem Kapitel war die Hypothese, dass Bewertungen des Gefiederzustands, die mittels FUSE anhand von Bildsequenzen abgegeben wurden, genauer sind als solche, die auf Basis von einzelnen Bildern gemacht wurden. Dies wurde im direkten Vergleich beider Verfahren getestet. Weiterhin untersuchten die Experimente, wie sich die Anzahl der Bilder pro Sequenz sowie die Zahl der Sequenzen pro Tier auf die Bewertungsgenauigkeit auswirkt. Analog zu Kapitel 3 wurde die Methode sowohl für den primären Anwendungsfall der Gefiederbewertung als auch für die Altersschätzung von Menschen anhand von Bildern getestet. Die Ergebnisse dieser Untersuchung, dass die Nutzung von Bildsequenzen durch FUSE für beide Datensätze zu einer bis zu 7% höheren Genauigkeit im Vergleich zur konventionellen Bewertung mittels Einzelbildern führte. Diese Steigerung der Bewertungsgenauigkeit konnte zudem auch für künstliche gekürzte Bildsequenzen sowie eine begrenze Zahl von Sequenzen pro Tier, bzw. pro Mensch beobachtet werden.

Kapitel 5 untersuchte die Anwendung von Deep Learning zur Wiedererkennung einzelner Hühner anhand von Bildaufnahmen des Kopfes. Eine solche Wiedererkennung ist immer dann relevant, wenn verschiedene Aufnahmen innerhalb einer Bildsequenz einem bestimmten Tier zugeordnet werden müssen. Darüber hinaus ist sie Voraussetzung für jedes Langzeitmonitoring, sobald zeitliche Unterbrechungen zwischen einzelnen Aufnahmen auftreten. Das Ziel des in diesem Kapitel vorgestellten Ansatzes bestand darin, eine Wiederkennungs-Methode zu entwickeln, die für die Anwendung in großen Tierherden, nicht nur für Hühner, geeignet ist und kein manuelles Handling der Tiere erfordert. Dazu wurde ein Transformer-basiertes neuronales Netz für das Erlernen von Repräsentationen aus Inputbildern genutzt. Dieses Modell wurde so trainiert, dass Repräsentationen von Aufnahmen des gleichen Tieres eine hohe Ähnlichkeit aufweisen, während Aufnahmen verschiedener Tiere zu unterschiedlichen Repräsentationen führen. Dieses Vorgehen erlaubt eine Wiedererkennung auch ohne das Vorliegen vieler Trainingsdaten für jedes einzelne Tier. In diesem Kapitel wurde die Hypothese aufgestellt, dass die transformerbasierte Methode eine Wiedererkennung ermöglicht und dabei die Leistung traditioneller CNN-basierter Architekturen übertrifft. Dazu wurden verschiedene Konfigurationen beider Modellarchitekturen evaluiert und die Wiedererkennungsgenauigkeiten miteinander verglichen. Darüber hinaus wurde der Einfluss von Herdengröße und Bildanzahl pro Tier auf die Wiedererkennungsleistung analysiert, um die praktische Anwendbarkeit in kommerziellen Nutztierhaltungen zu beurteilen. Da es sich um den ersten Deep-Learning-

Ansatz zur Wiedererkennung von Hühnern handelt, wurde zusätzlich untersucht, welche optischen Merkmale im Kopfbereich der Tiere für die Wiedererkennung durch die verschiedenen Modelle relevant sind. In den Experimenten wurde die grundsätzliche Machbarkeit der Wiedererkennung von Hühnern mittels der vorgestellten Methode gezeigt. Dabei übertrafen die transformer-basierten Modelle die CNN-Architekturen deutlich und erreichten eine Genauigkeit von über 90% für kleine Tiergruppen bis zu fünf Tieren. Für Gruppen von 100 verschiedenen Hühnern wurde eine Genauigkeit von ca. 40% ermittelt. Während diese Genauigkeit den Einsatz der Methodik für eine begrenzte Anzahl von Tieren oder die Zuordnungen von Aufnahmen innerhalb einer Bildsequenz ermöglicht, ist sie für die kommerzielle Anwendung in Herden mit mehreren Tausend Tieren noch nicht ausreichend. Darüber hinaus zeigten die Experimente, dass für die Wiederkennung insbesondere Merkmale wie der Kamm, die Kehllappen und Ohrläppchen eines Huhns relevant waren.

Zusammenfassend lässt sich feststellen, dass das übergeordnete Ziel dieser Arbeit, robuste Methoden für bildbasiertes Tiermonitoring zu entwickeln, die speziell für den Einsatz im praktischen Kontext ausgelegt sind, durch das vorgestellte Framework erreicht wurde. Neben der verbesserten Robustheit gegenüber den Umgebungseinflüssen in kommerziellen Nutztierställen bildete insbesondere die Integration von Mechanismen zur Unsicherheitsbewertung in die Vorhersagemodelle einen zentralen Beitrag dieser Arbeit. Dies steigerte einerseits die Transparenz eines Monitoringsystems und ermöglichte andererseits das Filtern von unzuverlässigen Vorhersagen zur Erhöhung der Vorhersagegenauigkeit. Eine detaillierte Diskussion der einzelnen Kapitel im Hinblick auf ihren Beitrag zum übergeordneten Ziel der Arbeit wurde in Kapitel 6 vorgenommen. Die in dieser Arbeit vorgestellten Methoden bilden eine aussichtsreiche Grundlage für das automatisierte Tiermonitoring im kommerziellen Einsatz. Gleichzeitig stellen sie jedoch noch keine fertig einsatzbereite Plug-and-Play-Lösung für alle denkbaren Anwendungsfälle dar. Die Etablierung automatisierter Monitoringsysteme in der Praxis erfordert daher weiterführende Entwicklungen, die über die reine Optimierung der Software hinausgehen und ebenfalls in Kapitel 6 näher betrachtet und diskutiert wurden.

Acknowledgements

Next week will mark exactly five years since I began my PhD journey and started working in the Radix building. Time flies. Before starting at Wageningen University, I must admit I had not even heard of Wageningen. I was unaware that the world's leading university in the field of agriculture was just 2.5 hours away from my hometown in Germany. I first learned about it when I came across the PhD position in AI and Robotics offered by Big Dutchman in collaboration with WUR. The idea of combining scientific research on state-of-the-art technologies with the opportunity to turn ideas into real products really appealed to me, so I applied for the position and was fortunate to get in shortly after. While this pairing of research and industry was a great motivation, it also posed the biggest challenge in the beginning of my PhD when I tried working out a research topic that fulfills both corporate interest and scientific quality. Initially, it seemed to me that university and industry were speaking two different languages (and it's not only about English and German). However, soon, I learned to translate between both languages and together, with the great support of my supervisors, managed to shape a topic and come up with a project proposal that both sides agreed on. The hybrid structure turned out to be great, with me especially enjoying the engineering part, including the development of new algorithms and the opportunity to test them on real farms.

I have often been told that pursuing a PhD is a journey full of highs and lows, where you frequently question everything and often end up far from your initial proposal. However, I can say that I experienced this journey as being quite smooth. Even though paper writing never became my greatest passion, I did not encounter any major setbacks or unexpected turns. This seamless progress was primarily due to the excellent support I received all along the way. Therefore, I would like to take this opportunity to express my gratitude to those who have supported me over the last five years. First, I would like to thank my co-promotors and promoter for their invaluable guidance, support, and mentorship. Without you, this thesis would not have been possible.

Marjolein, as my first supervisor, you were involved in the project even before I began in Wageningen. From day one, you not only helped me maintain the scientific quality of my work and supported me on a daily basis, but you also had a talent for communicating ideas so that everyone could see the long-term vision. What I really appreciated when working

with you was your pragmatic approach to ensure things moved forward. I could always rely on your thorough feedback, but instead of fine-tuning the tiniest detail of a paper in endless revisions, you'd rather say, "let's give it a try" and submit it, which usually turned out to be the right choice. Thank you for your optimism and the enthusiasm you showed for the PhD project from the very beginning.

Gert, you stepped in as my supervisor when the project's focus shifted to computer vision, and greatly helped me with your expertise in the field. It was always fun to discuss the many ideas you had for adding novelty in my work and advancing the state of the art, even though I couldn't implement all of them within the few years of my PhD. Thank you for your consistently constructive feedback and the effort you put into supervision, even when that meant joining our meetings while on your bike. I must admit that I was sometimes a bit shocked by how much a paper I thought was "almost there" could still be improved. Nevertheless, I know that these revisions had a great positive impact on my research. Still, I'm impressed by how you provided such detailed feedback despite your busy schedule and yet managed to be competitive in the FTE sports sessions.

Peter, as my promotor you always had the bigger picture in mind and an excellent sense of where to position my research, especially in the early stages when I was just starting on the topic and drafting the proposal. It was based on your initiative that we shifted from the initially planned focus on robotics to computer vision and AI - a decision that, looking back, was strategically very beneficial. It directed my work into a field whose popularity has recently exploded, making it highly relevant not only in academic research but also in industry and for Big Dutchman. Over the past few years, I could always count on your thoughtful insights and advice. Thank you for your mentorship and for always being available to provide guidance when needed.

I further want to thank my supervisors and colleagues at Big Dutchman. A special thanks goes to **Felix and Ludger**, who supported me in bridging the gap between theory and practice, which initially appeared bigger to me than it actually was. **Felix**, as my team leader, you always gave me the freedom and time to explore my ideas while ensuring I never lost sight of their practical applications. Whenever support or equipment for experiments was needed, you arranged things quickly and without bureaucratic hassle. Moreover, you were always up for a chat or a beer, which made the work not only productive but also enjoyable. A big thanks for that! **Ludger**, you truly deserve the title "master of the unexpected." Your talent for always coming up with out-of-the-box ideas has been both enlightening and entertaining. Only you knew that sometimes a meeting with all supervisors needs to be interrupted to introduce everyone to the "*Mettfrühstück*" tradition. And no one else could have taught me the tricks of effective communication and presentation to defend my ideas as effectively as you did. Thank you for your great guidance and the unforgettable lessons along the way. I'd also like to express my sincere gratitude to all the colleagues at Big Dutchman who offered their support whenever it

was needed. A special shoutout goes to the apprentices and trainees who helped with data annotation during the early stages of my PhD, back when automated labeling was not a thing. It was way too much work for one person, and your help made it a bit more manageable – big thanks for that!

Despite the ones directly involved in the PhD project, I have met many great people whose support must not be underestimated. Here, I want to mention my colleagues in the Agricultural Biosystems Engineering Group, especially **Akshay, David, Rick, Robert, Cecile, Helena, Henry, Thijs, Bart, Homa, Maria, Salma, Arni, Tim, Ali, Peyman, Laiquan, Daniel, Sun, Wali, Bert, Eldert, Ricardo, Miranda, Areerat, Reniela, Sam, Simon, Rik**, and **Xin**. Thanks for the great time we had together, whether it was by having everyday chats or helping each other out with smaller and larger issues. **Akshay**, you started your PhD journey almost simultaneously with mine and were a great support, especially in the early days, when I faced the typical challenges of a newbie. Since then, I have always enjoyed our regular chitchats in the office. Beyond work, I got to know you as an awesome roommate who generously shared his apartment with me during my time in Wageningen. While writing this, I'd like to take the opportunity to apologize for always being a bit too competitive during our football sessions and take full responsibility for the bruises I left you with. I hope you've recovered both physically and mentally from these matches. It's a pity you're no longer in Wageningen, but I wish you all the best for the future and can't wait to hear about the successes of your future start-up! To my paranymphs **David** and **Rick**, you joined the PhD madness shortly after Akshay and me, and we shared quite some office hours - first on the second floor, and later when Rick and I decided to move our office into the former storage room. David, as you know much more about deep learning and computer vision than I do, our discussions were always enlightening for me, sometimes in relation to work but more frequently on random topics. I learned a lot, not just about neural networks, but also about things like how to recognize the beginning of summer in Spain. Thanks for that! Rick, you taught me much about "typical Dutch things", while I did my best to broaden your horizons with the finest German internet memes from the early 2000s. You also introduced me to Kapsalon during our Monday evening meetings with Bart to prepare for the field robot event – a great culinary experience. Thank you both for supporting me as paranymphs and for all the fun we had during and after work. I really enjoyed the time.

Last but not least, I want to thank **my family**. To **my parents**, for making all this possible and for their unwavering support, not just during the time of the but throughout the last (almost) 30 years. I also extend my gratitude to **my sister** and **grandparents**, especially **my grandmother**, who was always deeply interested in my work but will not be able to witness the defense.

As I close, thanks to **everyone who has supported me** along this journey, whether

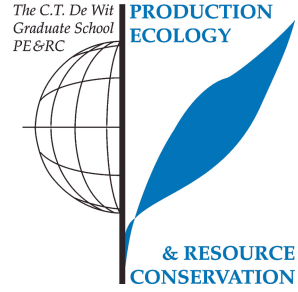
directly or indirectly. If I have unintentionally missed mentioning anyone, please accept my apologies and know that your contribution has been deeply appreciated.
Cheers!

About the author

Christian Lamping was born in Vechta, Germany, on the 11th of January 1995. He studied Industrial Engineering at the University of Hanover and Michigan Technological University (Houghton, USA). In 2019, he obtained his master's degree with a thesis on the development of autonomous indoor navigation for drones. For this thesis, he received the "IPH Future Award", honoring the best thesis of the year. Following his graduation, Christian sought a PhD opportunity in a similar field and found it at Big Dutchman in partnership with Wageningen University & Research. The results of this project are documented in this thesis. Since February 2024, Christian has been employed as Product Owner for AI and Robotics at Big Dutchman.

PE&RC Training and Education Statement

With the training and education activities listed below the PhD candidate has complied with the requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 30 ECTS (= 20 weeks of activities)



Review/project proposal (4.5 ECTS)

- Non-invasive individual animal monitoring using computer vision

Post-graduate courses (4.5 ECTS)

- A25 - Computer Vision by Learning Course, ASCI research school (2022)
- Machine Learning, Coursera (2020)

Deficiency, refresh, brush-up courses (3 ECTS)

- Deep Learning in Data Science, WUR (2020)

Invited review of journal manuscript (1 ECTS)

- Biosystems Engineering: Plumage condition assessment using Deep Learning (2023)

Competence, skills and career-oriented activities (3.1 ECTS)

- How to present online, The floor is yours (2020)
- Critical thinking and argumentation, WUR (2022)
- SCRUM – Master and Product Owner, Big Dutchman (2022)
- Project and Time Management, WUR (2023)

Scientific Integrity/Ethics in science activities (0.8 ECTS)

- Ethics and Animal Science, WUR (2023)

PE&RC Retreat, PE&RC Day, and other PE&RC events (1.5 ECTS)

- PE&RC First year retreat (2020)
- PE&RC Midterm retreat (2022)

National scientific meetings, local seminars, and discussion groups (6.5 ECTS)

- Deep Learning Discussion Group (2019-2020)
- Machine Vision & Robotics Thematic Meeting (2021-2024)
- Big Dutchman Robotics Workshop Navigation + Control (2021)
- Agro-Food Robotics expertise and exchange meetings (2022)
- Symposium on Agricultural Robotics (2022)
- Animal Science Seminar University of Hohenheim (2023)

International symposia, workshops and conferences (3.2 ECTS)

- Netherlands Conference on Computer Vision (NCCV), Volendam (2022)
- Land.Technik Conference, Berlin (2022)

Societally relevant exposure (0.4 ECTS)

- Guest lecture Mobile Robotics (2022, 2023)

Lecturing / supervision of practical's / tutorials (6 ECTS)

- GRS-34806 Deep Learning (2021-2023)
- WIAS: Image and Video Analysis (2021, 2023)

BSc/MSc thesis supervision (15 ECTS)

- Preventing pollution of camera systems in pig poultry houses (BSc.)
- Tracking of Laying Hens in a Farm Environment Using Computer Vision (MSc.)
- Optimization of deep learning-based classification using object tracking (BSc.)
- Design of a module for camera-based plumage condition assessment (MSc.)
- Determination of body orientation in agricultural poultry farming using computer vision (MSc.)

The research described in this thesis was financially supported by Big Dutchman International GmbH, Vechta, Germany

Cover design: Christian Lamping
Printed by: DigiForce — Proefschriftmaken.nl