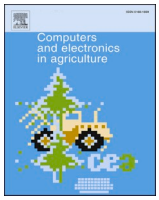
Computers and Electronics in Agriculture 234 (2025) 110182

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag



Original papers

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

Christian Lamping *, Marjolein Derks, Gert Kootstra

Farm Technology Group, Wageningen University & Research, 6700 AA Wageningen, the Netherlands

ARTICLE INFO

Keywords:
Uncertainty
Multi-view
Image sequences
Assessment
Robustness

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.

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

* Corresponding author.
E-mail address: christian.lamping@wur.nl (C. Lamping).

https://doi.org/10.1016/j.compag.2025.110182
Received 20 February 2024; Received in revised form 22 October 2024; Accepted 23 February 2025
Available online 22 March 2025

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.

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.

1.1.1. 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 (Molchanov 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, 2015; Postels et al., 2019).

Other methods for estimating uncertainty include ensemble methods (Lakshminarayanan et al., 2016; Gawlikowski et al., 2021), evidential approaches (Charpentier et al., 2020; Sensoy et al., 2018; Amini et al.) and test-time augmentation methods (Molchanov 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.

1.1.2. 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 summarizes 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 (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., 2022), multi-view representation learning (Tian et al., 2019; 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; Han et al., 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., 2022), 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. Fusion approaches were also developed in (Zhou et al., 2022) and (Zhou et al., 2024), where information from different branches of a network architecture were fused to improve or refine the final model output. In (Tian et al., 2019), various uncertainty measures were considered. 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 combine 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., 2020; Wei et al., 2020; He et al., 2021; 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., 2020; Chen et al., 2019; Pei et al., 2016; Peng et al., 2017). For instance, attention mechanisms have been incorporated into CNNs in order to recognize facial expressions from image sequences (Li et al., 2020) or for classification of pedestrian attributes from surveillance camera videos (Chen et al., 2019). Study (Pei et al., 2016) combines 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.

2. Method: Uncertainty-aware multi-view assessment

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. Fig. 1 provides a visual representation of the intended clustering procedure for this method when applied to a sequence capturing the movement of a chicken.

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 2.1. Subsequently, an appearance-based clustering method is used to group all visually similar detections together and separate dissimilar ones (Section 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 2.3). An overview of the complete method is provided in Fig. 2.

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 Chapter 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; Molchanov et al., 2020; Gal and Ghahramani, 2015), which allows 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 (He et al., 2017) 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

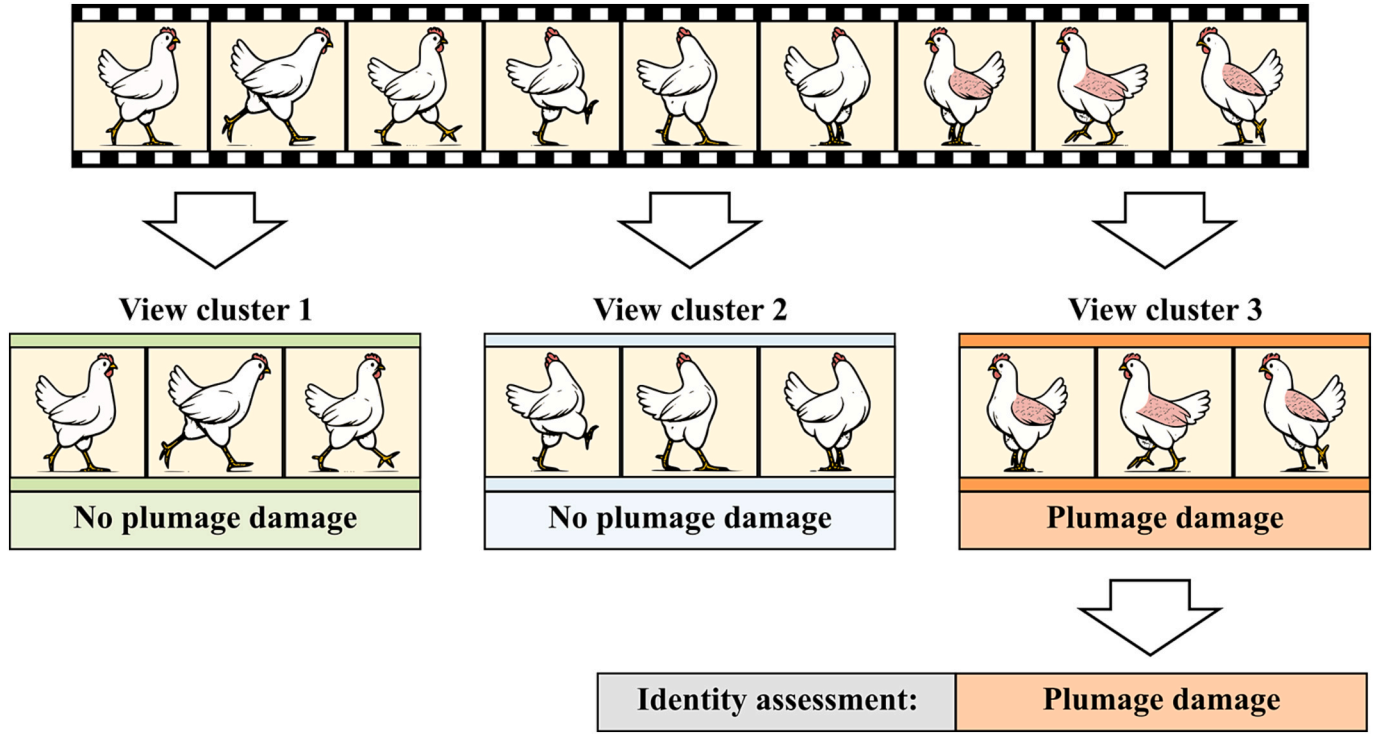


Fig. 1. Intended procedure for identifying distinct viewpoints from a sequence of detection through clustering. For the given example of a moving chicken, the three cluster represent views from the right, rear, and left sides of the animal. While view clusters 1 and 2 do not exhibit any plumage damages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

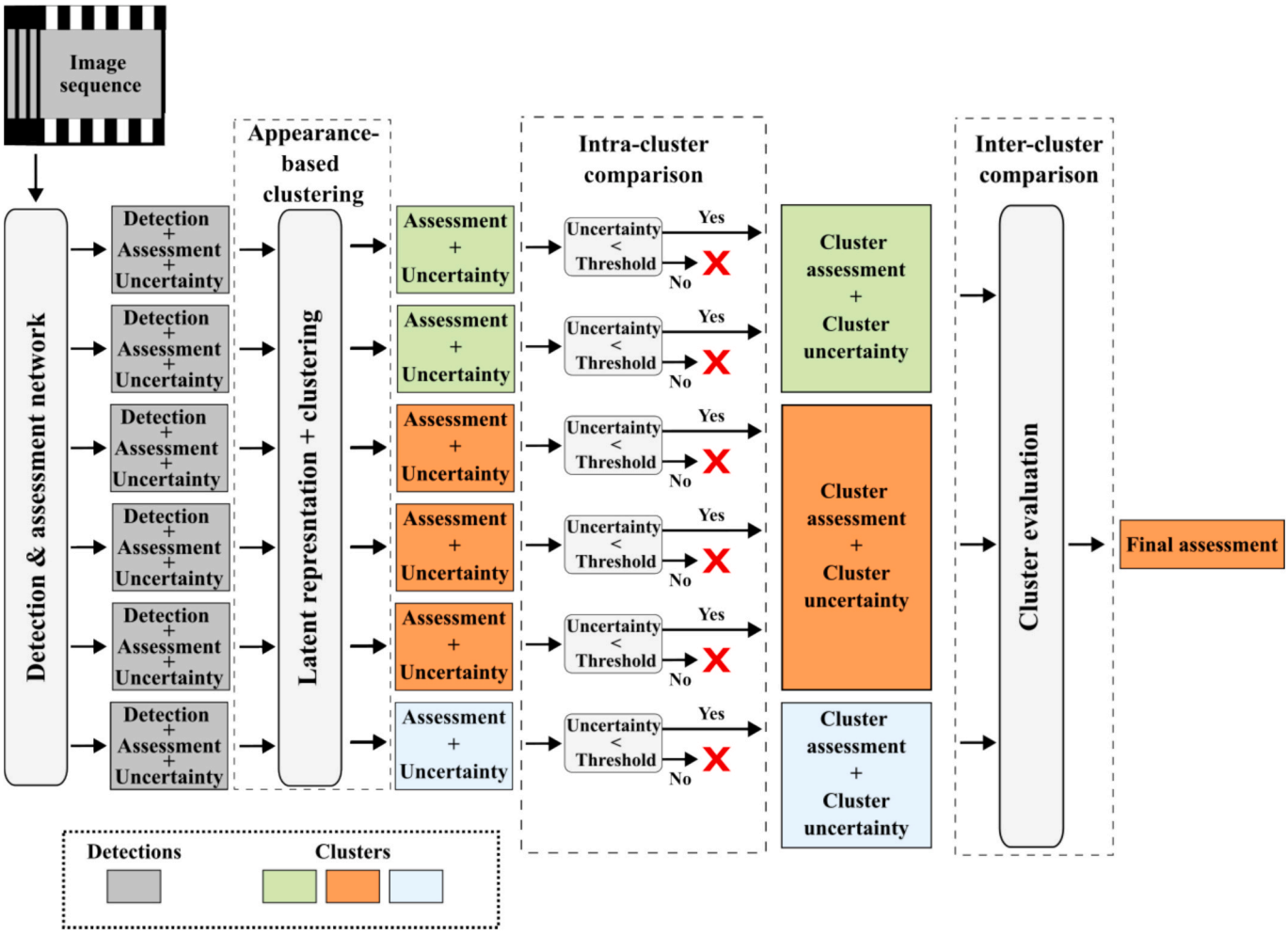


Fig. 2. Generic pipeline of the developed method for image sequence assessment, consisting of a detection and assessment network for the assessment on instance-level, followed by appearance-based clustering of the individual detections. Subsequently, assessments within a cluster are evaluated to obtain a single assessment for each cluster. Those are then compared with each other, leading to a final assessment.

the number of samples, the loss function is defined as:

$$L_{score} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left(y_i^{score} - \hat{y}_i^{score}\right)^2}{2\sigma_i^2} + \frac{1}{2}\sigma_i^2 \quad (1)$$

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 (1). Therefore, following the approach of (Di 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:

$$L_{score\_calib} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left(y_i^{score} - \hat{y}_i^{score}\right)^2}{2\sigma_i^2} + \frac{1}{2}\sigma_i^2 + \left|\sigma_i^2 - \left(y_i^{score} - \hat{y}_i^{score}\right)^2\right| \quad (2)$$

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, 2015). 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.

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.

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.

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 $\widehat{y}_i$ 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 $a_i = \{\widehat{y}_i, \sigma_i^2\}$, 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 Fig. 3.

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, as further explained in Section 2.3.1. Subsequently, in the second step, detailed in Section 2.3.2, we determined the most representative cluster while considering the uncertainty associated with each cluster.

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 assessments 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 cluster c was defined as:

$$\widehat{Y}_{c_j} = \frac{\sum_{i=1}^{n_j} w_i \widehat{y}_i}{\sum_{i=1}^{n_j} w_i}, w_i = \frac{1}{\sigma_i^2} \tag{3}$$

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:

$$\Sigma_{c_j}^2 = \frac{1}{\sum_{i=1}^{n_j} \frac{1}{\sigma_i^2}} \tag{4}$$

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

Fig. 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.

results in an assessment prediction, $\widehat{Y}_{c_j}$, and uncertainty, $\Sigma^2_{c_j}$, per cluster.

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 $\widehat{Y}_{c_j}$ and an associated uncertainty $\Sigma^2_{c_j}$. 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 prioritization 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:

$$Y = \widehat{Y}_{c_k}, k = \underset{0<j<J}{\operatorname{argmax}} \frac{\widehat{Y}_{c_j}}{\Sigma^2_{c_j}} \tag{5}$$

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:

$$Y = \widehat{Y}_{c_k}, k = \underset{0<j<J}{\operatorname{argmax}} \frac{1}{\Sigma^2_{c_j}} \tag{6}$$

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, as outlined in Section 2.4.2, focused the direct comparison with the standard implementation of ChickenNet. Following that, experiment 2, detailed in Section 2.4.3, 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 as presented in Section 2.4.4. 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.

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 to 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 to 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 Fig. 4.

An instance denotes a detection of an identity at a certain timestep of the sequence.

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.

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 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:

2.4.2.1. 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 $k = 0$.

2.4.2.2. 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.

2.4.2.3. 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.

2.4.2.4. 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.

Fig. 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

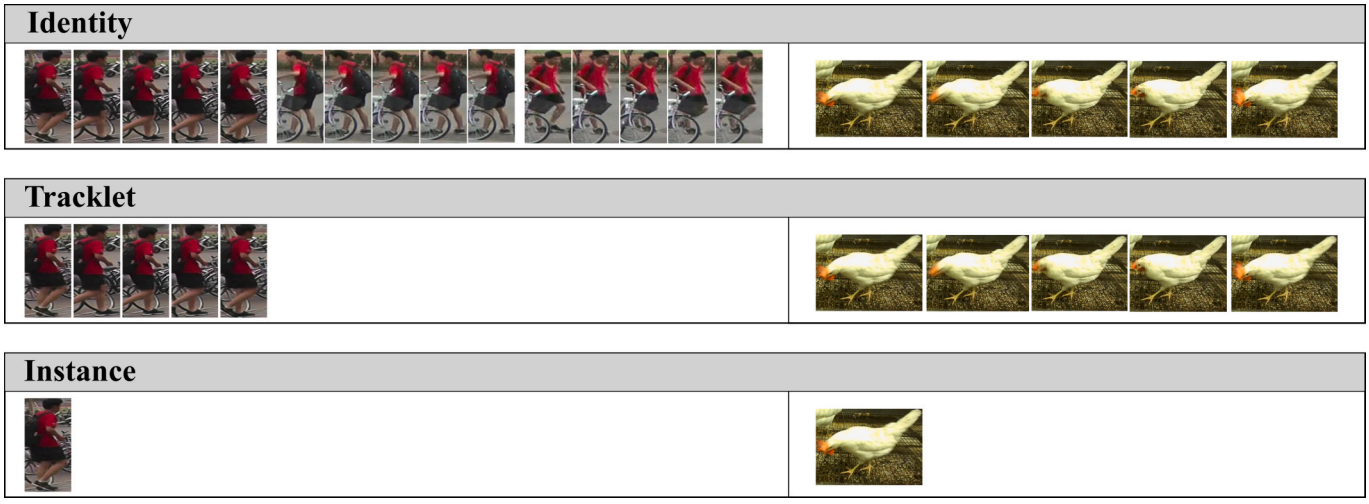


Fig. 4. Structure of the MARS-Attributes dataset (left) and chicken dataset (right). Per identity, the MARS-Attributes contains multiple tracklets, while the chicken dataset consists of a single tracklet per identity.

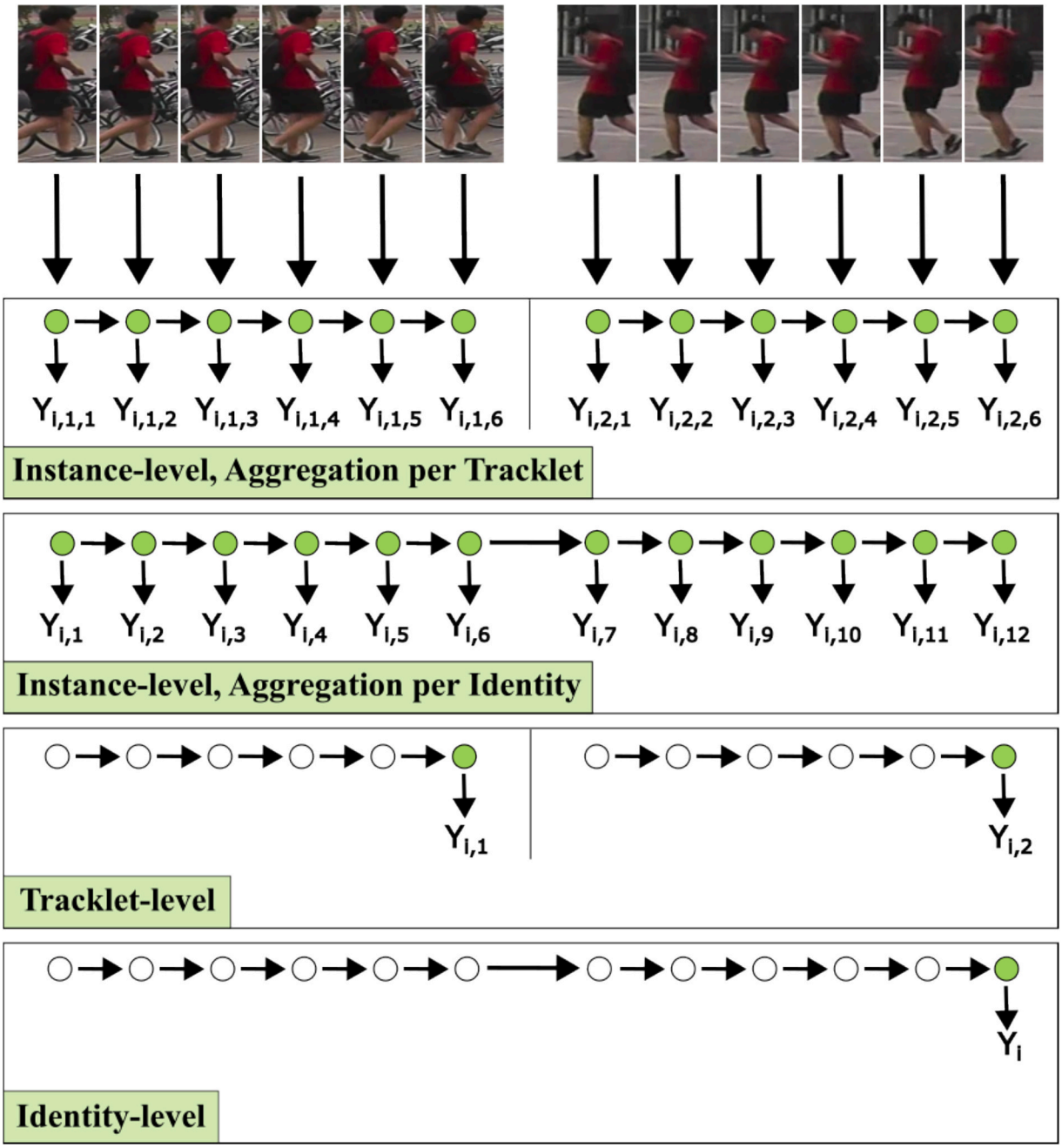


Fig. 5. Evaluation approaches on instance-, tracklet-, and identity-level. Instance-level approaches result in a prediction for each instance and can be obtained by either aggregating all instances of a tracklet or all instances of an identity. Tracklet-level approaches result in one prediction per tracklet 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.

correct predictions among the total number of samples. Additionally, we analyzed the mean squared error (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 Fig. 5.

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.

2.4.4. Experiment 3 – alternative uncertainty quantification

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 the experiment described in Section 2.4.2.

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.

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 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.

Table 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**

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.

Fig. 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 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-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.

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. Fig. 7 visualizes the accuracies for the different evaluation approaches presented in Section 2.4.2 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. Fig. 8 illustrates those accuracies obtained for the MARS-Attributes dataset. It presents the accuracies for

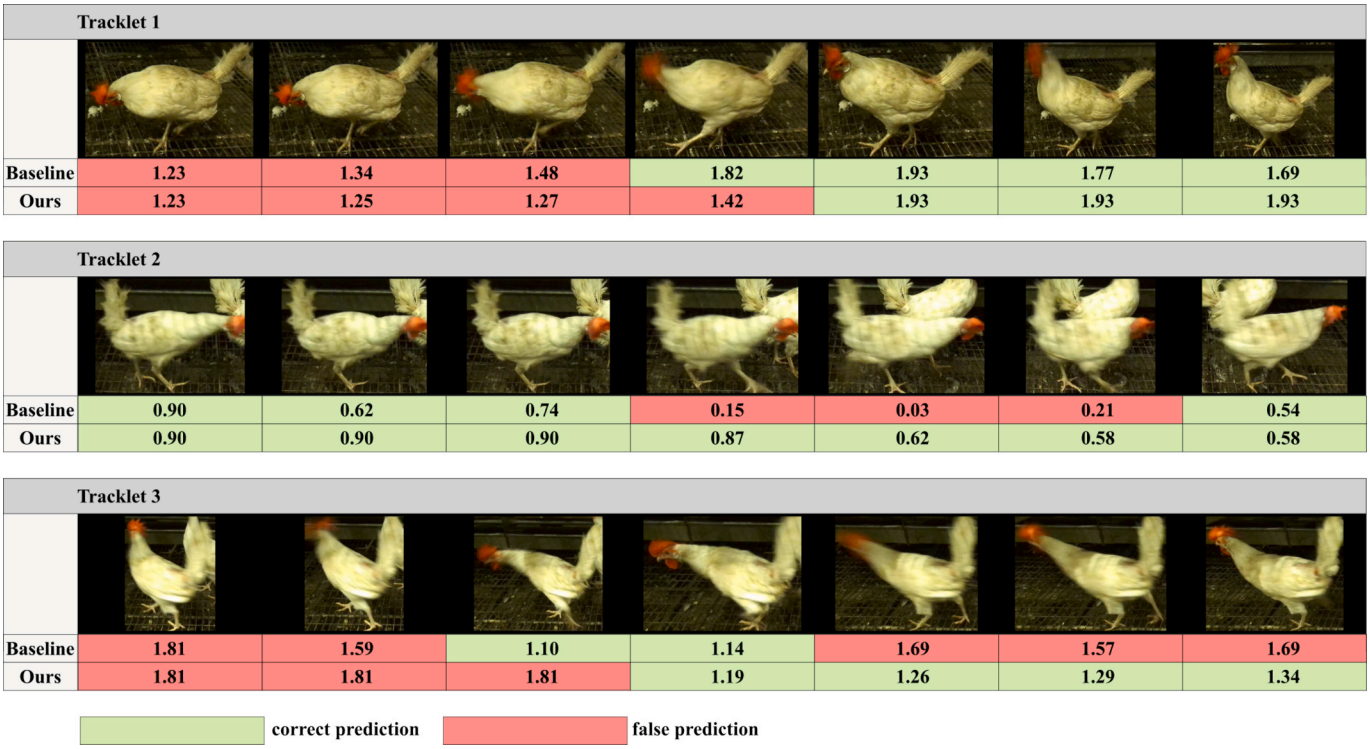


Fig. 6. Instance-wise assessment score predictions of the standard ChickenNet (baseline) model compared to assessments provided by our method for consecutive instances from three tracklets of the chicken dataset. Colours indicate whether the predicted score was correct or not. A correct prediction in the final frame of a tracklet implies a correct assessment of the entire tracklet.

Table 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 %)

instances ranging from 3 to 20 per tracklet, but also for 1–100 tracklets per identity.

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 number 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

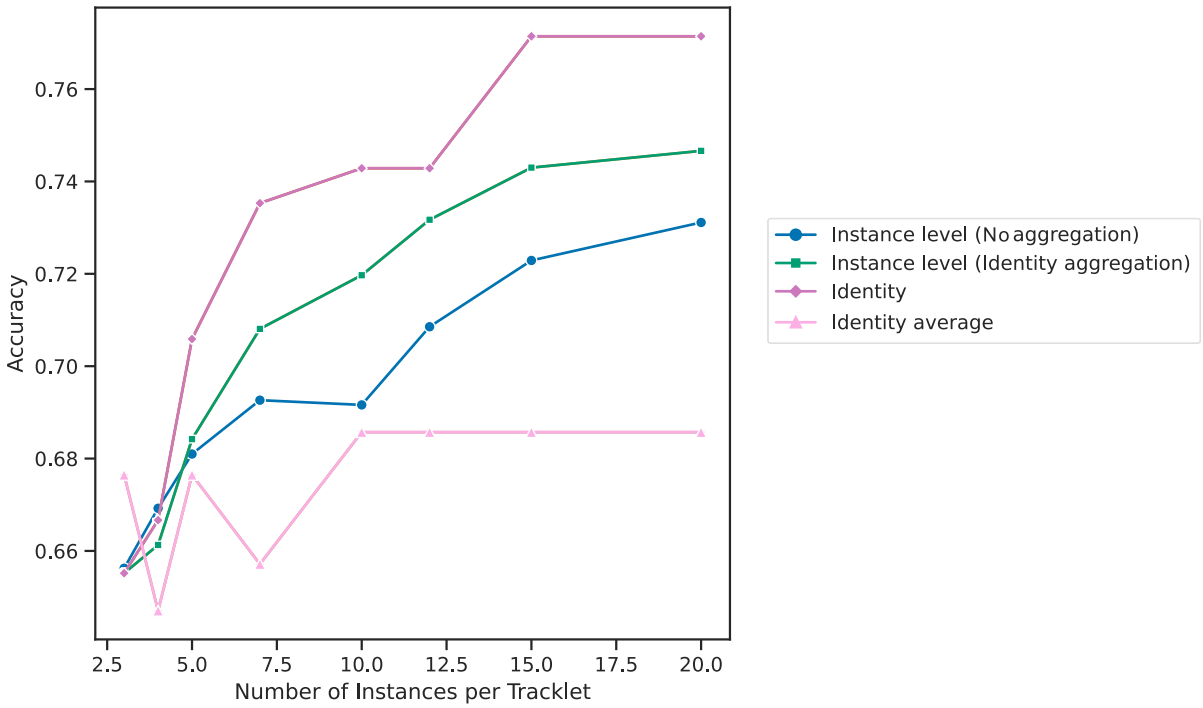


Fig. 7. Assessment accuracies on the chicken dataset for varying numbers of instances per tracklet considered by our method. As the chicken dataset consists of a single tracklet per identity, corrections per tracklet and per identity are equal.

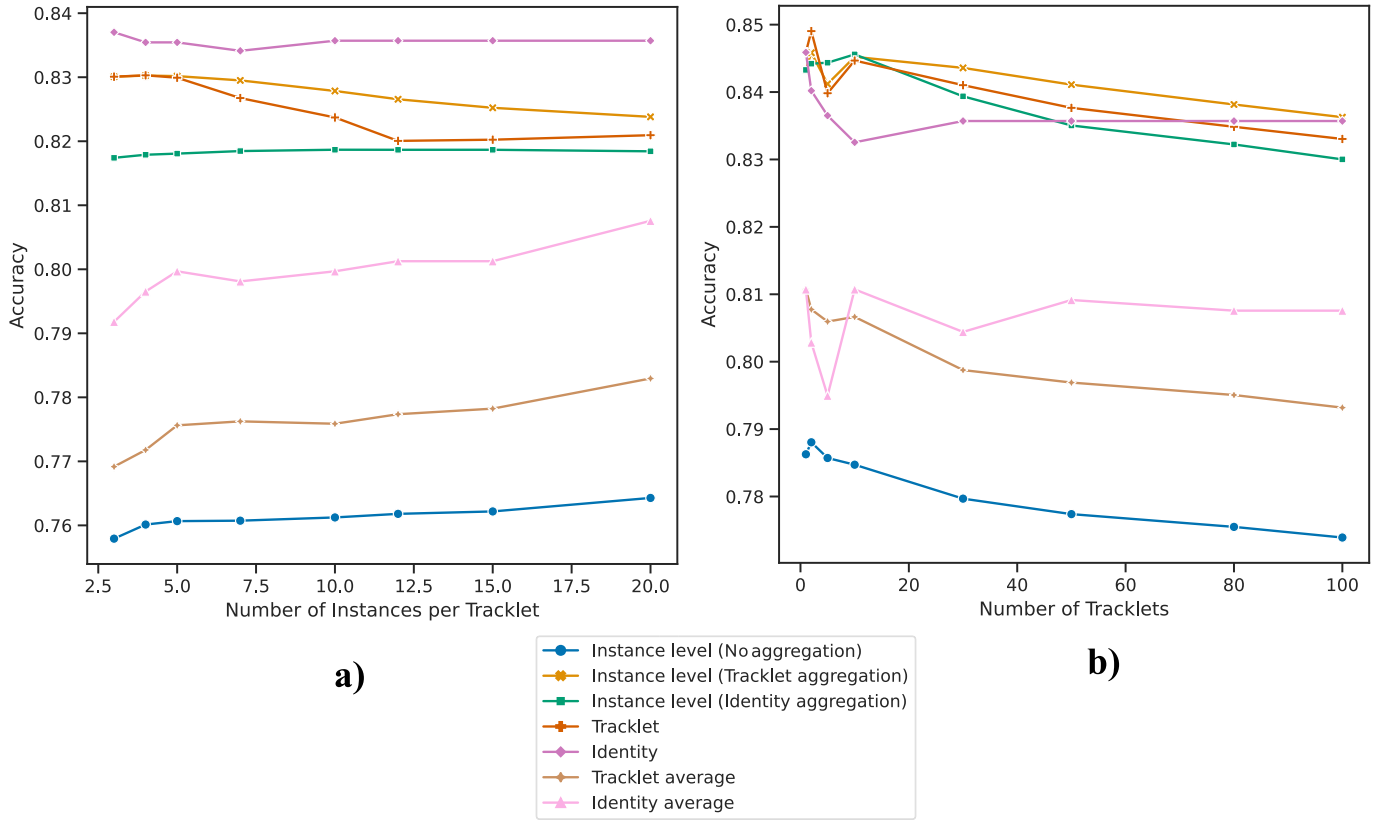


Fig. 8. Assessment accuracies on the MARS-Attributes dataset for varying numbers of a) instances per tracklet and b) tracklets per identity considered by our method.

impact between limited instances per tracklet and limited tracklets per identity.

3.3. Alternative uncertainty quantification

Using epistemic uncertainty to weight individual instance predictions yielded similar results to using aleatoric uncertainty. Table 3

and 4 present the results of the experiments on the chicken dataset and the MARS-Attributes dataset.

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 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. 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., 2016), 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., 2019), 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.

Table 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**

Table 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 %)

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. Especially the assessment model is important, as it determines the input for all subsequent processing steps. For example, a perfectly accurate uncertainty estimator that indicates false assessments becomes redundant if all assessments are consistently inaccurate and would lead to low quality results. 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. As a general guideline, the accuracy of assessment models with low-uncertainty predictions becomes increasingly critical when fewer images are available per sequence. When a high number of images is present, inaccurate predictions can be compensated by the subsequent uncertainty estimation and filtering without impairing the final assessment output.

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. Thus, in use cases including highly similar frames, a feature descriptor specialized to differentiate between perspectives is expected to improve the overall assessment results of the presented approach. 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 Fig. 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 pretrained 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.

4.2. 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 2.4.2. 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.

4.3. 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 to provide better flexibility across different datasets.

Moreover, while this work provides a solid foundation for enhancing transparency in the assessment process through uncertainty estimation, it is important to acknowledge that there remains room for further improvement in the transparency of deep learning-based monitoring systems. A logical first step could involve providing more information on how the uncertainty estimates influence the final predictions within the presented approach. One indication here could be the number of assessments that were rejected due to high uncertainty. This is especially relevant if uncertain predictions are not equally distributed among the different assessment categories, so that an increasing rejection rate could lead to a change in the distribution of the assessments. Dealing with this could be a topic addressed by future research. Beyond the estimation of uncertainty, a transparent monitoring system could further focus on offering explainable recommendations for the users, allowing them to better understand and critically assess the system's decisions.

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.

5. Conclusion

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.

CRediT authorship contribution statement

Christian Lamping: Writing – original draft, Visualization, Software, Methodology, Investigation, Conceptualization. **Marjolein Derks:** Writing – review & editing, Methodology, Conceptualization. **Gert Kootstra:** Writing – review & editing, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Christian Lamping reports financial support was provided by Big Dutchman International GmbH. Christian Lamping reports a relationship with Big Dutchman International GmbH that includes: employment.

Data availability

Data will be made available on request.

References

Amini, A., Schwarting, A. W., Soleimany, Rus, D., Deep Evidential Regression.

Arnez, F., Espinoza, H., Radermacher, A., Terrier, F., 2020. *A Comparison of Uncertainty Estimation Approaches in Deep Learning Components for Autonomous Vehicle Applications*.

Bachman, P., Hjelm, R. D., Buchwalter, W., 2019. *Learning Representations by Maximizing Mutual Information Across Views*.

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

Charpentier, B., Zügner, D., Günnemann, S., 2020. *Posterior Network: Uncertainty Estimation without OOD Samples via Density-Based Pseudo-Counts*.

Chen, Z., Li, A., Wang, Y., 2019. *A Temporal Attentive Approach for Video-Based Pedestrian Attribute Recognition*.

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

Chen, K., Yao, L., Zhang, D., Wang, X., Chang, X., Nie, F., 2020. A semisupervised recurrent convolutional attention model for human activity recognition. *IEEE Trans. Neural Networks Learn. Syst.* 31 (5), 1747–1756. https://doi.org/10.1109/TNNLS.2019.2927224.

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

Di Feng, L., Rosenbaum, C., Glaeser, F.T., Dietmayer, K., 2019. *Can we trust you? On Calibration of a Probabilistic Object Detector for Autonomous Driving*. https://doi.org/10.48550/arXiv.1909.12358.

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

Gal, Y., Ghahramani, Z., 2015. *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., Zhu, X. X., 2021. *A Survey of Uncertainty in Deep Neural Networks*.

Han, Z., Zhang, C., Fu, H., Zhou, J. T., 2021. *Trusted Multi-View Classification*.

Han, Z., Zhang, C., H. Fu, Zhou, J. T., 2022. *Trusted Multi-View Classification with Dynamic Evidential Fusion*.

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

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. *Mask R-CNN*.

Heo, J., Lee, H. B., Kim, S., Lee, J., Kim, K. J., Yang, E., Hwang. S. J., 2018. *Uncertainty-Aware Attention for Reliable Interpretation and Prediction*.

Jeon, H.Y., Tian, L.F., 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.

Kendall, A., Gal, Y., 2017. *What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?*.

Kiela, D., Grave, E., Joulin, A., Mikolov, T., 2018. *Efficient Large-Scale Multi-Modal Classification*.

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

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2016. *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles*.

Lamping, C., Derks, M., Groot Koerkamp, P., Kootstra, G., 2022. ChickenNet - an end-to-end approach for plumage condition assessment of laying hens in commercial farms using computer vision. *Comput. Electron. Agric.* 194, 106695. https://doi.org/10.1016/j.compag.2022.106695.

Lamping, C., Kootstra, G., Derks, M., 2023. Uncertainty estimation for deep neural networks to improve the assessment of plumage conditions of chickens. *Smart Agric. Technol.* 5, 100308. https://doi.org/10.1016/j.atech.2023.100308.

Li, J., Jin, K., Zhou, D., Kubota, N., Ju, Z., 2020. Attention mechanism-based CNN for facial expression recognition. *Neurocomputing* 411, 340–350. https://doi.org/10.1016/j.neucom.2020.06.014.

Liong, V. E., T. Nguyen, N. T., Widjaja, S., Sharma, D., Chong, Z. J., 2020. *AMVNet: Assertion-based Multi-View Fusion Network for LiDAR Semantic Segmentation*.

Lu, X., Lin, Z., Shen, X., Mech, R., Wang, J.Z., 2015. Deep Multi-patch Aggregation Network for Image Style, Aesthetics, and Quality Estimation. In: In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 990–998. https://doi.org/10.1109/ICCV.2015.119.

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

Molchanov, D., Lyzhov, A., Molchanova, Y., Ashukha, A., Vetrov, D., 2020. *Greedy Policy Search: A Simple Baseline for Learnable Test-Time Augmentation*.

Morvant, E., Habrard, A., Ayache, S., 2014. *Majority Vote of Diverse Classifiers for Late Fusion*.

Nahian, M. A., Iftekhar, A. S. M., Islam, M. T., Rahman, S. M. M., Hatzinakos, D., 2017. *CNN-Based Prediction of Frame-Level Shot Importance for Video Summarization*.

Pei, W., Baltrušaitis, T., Tax, D. M.J., Morency, L.-P., 2016. *Temporal Attention-Gated Model for Robust Sequence Classification*.

Peng, Y., Zhao, Y., Zhang, J., 2017. *Two-stream Collaborative Learning with Spatial-Temporal Attention for Video Classification*.

Postels, J., Ferroni, F., Coskun, H., Navab, N., Tombari, F., 2019. *Sampling-free Epistemic Uncertainty Estimation Using Approximated Variance Propagation*.

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

Sadiq, B.O., Muhammad, B., Abdullahi, M.N., Onuh, G., Muhammed, A.A., Babatunde, A. E., 2020. Keyframe extraction techniques: a review. *Elektrika* 19 (3), 54–60. https://doi.org/10.11113/elektrika.v19n3.221.

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

Sensoy, M., Kaplan, L., Kandemir, M., 2018. *Evidential Deep Learning to Quantify Classification Uncertainty*.

Tian, J., Cheung, W., Glaser, N., Liu, Y.-C., Kira, Z., 2019. *UNO: Uncertainty-aware Noisy-Or Multimodal Fusion for Unanticipated Input Degradation*.

Tian, Y., Krishnan, D., Isola, P., 2019. *Contrastive Multiview Coding*.

Tian, Y., Li, D., Xu, C., 2020. Unified Multisensory Perception: Weakly-Supervised Audio-Visual Video Parsing. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox and Jan-Michael Frahm, Eds. Lecture Notes in Computer Science. Springer International Publishing, Cham, 436–454. DOI: 10.1007/978-3-030-58580-8_26.

Wang, Y., Geng, Z., Jiang, F., Li, C., Wang, Y., Yang, J., Lin, Z., 2021. *Residual Relaxation for Multi-view Representation Learning*.

Wang, S., Liu, J., Yu, G., Liu, X., Zhou, S., Zhu, E., Yang, Y., Yin, J., Yang, W., 2022. Multiview Deep Anomaly Detection: A Systematic Exploration. *IEEE transactions on neural networks and learning systems* PP. DOI: 10.1109/TNNLS.2022.3184723.

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

Wojke, N., Bewley, A., 2018. Deep Cosine Metric Learning for Person Re-Identification 10, 748–756. https://doi.org/10.1109/WACV.2018.00087.

Wojke, N., Bewley, A., Paulus, D., 2017. *Simple Online and Realtime Tracking with a Deep Association Metric*.

Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q., 2016. MARS: A Video Benchmark for Large-Scale Person Re-Identification. In . Springer, Cham, 868–884. DOI: 10.1007/978-3-319-46466-4_52.

Zhou, Q., Shi, H., Xiang, W., Kang, B., Wu, X., Latecki, L. J., 2022. *DPNet: Dual-Path Network for Real-time Object Detection with Lightweight Attention. DOI:* 10.48550/arXiv.2209.13933.

Zhou, Q., Wang, L., Gao, G., Kang, B., Ou, W., Lu, H., 2024. Boundary-guided lightweight semantic segmentation with multi-scale semantic context. *IEEE Trans. Multimedia* 26, 7887–7900. https://doi.org/10.1109/TMM.2024.3372835.

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