Enhanced camera-based individual pig detection and tracking for smart pig farms

Computers and Electronics in Agriculture

Guo, Qinghua; Sun, Yue; Orsini, Clémence; Bolhuis, Liesbeth; de Vlieg, Jakob et al

https://doi.org/10.1016/j.compag.2023.108009

Computers and Electronics in Agriculture 211 (2023) 108009

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Enhanced camera-based individual pig detection and tracking for smart pig farms

Qinghua Guo [a], Yue Sun [d,a,*], Clémence Orsini [b], J. Elizabeth Bolhuis [b], Jakob de Vlieg [c], Piter Bijma [b], Peter H.N. de With [a]

[a] *Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, 5612 AP, The Netherlands*
[b] *Department of Animal Sciences, Wageningen University & Research, Wageningen, 6708 PB, The Netherlands*
[c] *Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, 5612 AP, The Netherlands*
[d] *Faculty of Applied Science, Macao Polytechnic University, 999078, Macao Special Administrative Region of China*

ARTICLE INFO

Keywords:
Animal detection
Animal tracking
Camera-based detection and tracking
Multi-object tracking models

ABSTRACT

Negative social interactions are harmful for animal health and welfare. It is increasingly important to employ a continuous and effective monitoring system for detecting and tracking individual animals in large-scale farms. Such a system can provide timely alarms for farmers to intervene when damaging behavior occurs. Deep learning combined with camera-based monitoring is currently arising in agriculture. In this work, deep neural networks are employed to assist individual pig detection and tracking, which enables further analyzing behavior at the individual pig level. First, three state-of-the-art deep learning-based Multi-Object Tracking (MOT) methods are investigated, namely Joint Detection and Embedding (JDE), FairMOT, and YOLOv5s with DeepSORT. All models facilitate automated and continuous individual detection and tracking. Second, weighted-association algorithms are proposed for each MOT method, in order to optimize the object re-identification (re-ID), and improve the individual animal-tracking performance, especially for reducing the number of identity switches. The proposed weighted-association methods are evaluated on a large manually annotated pig dataset, and compared with the state-of-the-art methods. FairMOT with the proposed weighted association achieves the highest IDF1, the least number of identity switches, and the fastest execution rate. YOLOv5s with DeepSORT results in the highest MOTA and MOTP tracking metrics. These methods show high accuracy and robustness for individual pig tracking, and are promising candidates for continuous multi-object tracking for real use in commercial farms.

1. Introduction

With the increasing demand for animal products and growing societal concerns on animal welfare, effective monitoring and analysis of animal welfare become an increasing research focus. It is known that commercial farms may have problems with negative and damaging social interactions between animals. For instance, tail-biting among pigs threatens both animal welfare causing wounds and stress, and economic effects on pig production. Therefore, early detection and prevention of negative animal behaviors are critical and may give means for better animal welfare. This induces a new challenge for the animal science community, which aims to a holistic solution that unites human, animal, and environmental health (Zhang et al., 2019). In large-scale commercial farms, pigs are kept in groups, which increases the difficulty for farmers to monitor individual animals. To achieve an automated and systematic management strategy allowing timely alarms of welfare problems, and enabling interventions to prevent or reduce damaging behavior, continuous video-based monitoring of pigs is desired. Through continuous video-based monitoring, farmers are able to check the living conditions of pigs in real time. For instance, if an aggressive event occurs, pig farmers may get an alarm that includes location and moving trajectory of the aggressor. In this way, appropriate actions can be taken immediately to reduce such behavior. Meanwhile, the record of action trajectories of particular animals can also be used for analyzing individual characteristics, enabling phenotyping that can be potentially used for genetic selection. The possibility to track individual pigs will also benefit and boost the social network analysis, e.g. by providing proximity information. Therefore, the aim is to detect and track individual animals continuously and accurately.

Regarding the existing automated animal tracking systems, there are two categories. First, contact-based methods with attached sensors,

[*] Corresponding author at: Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, 5612 AP, The Netherlands.
E-mail address: joyyuesun@gmail.com (Y. Sun).

https://doi.org/10.1016/j.compag.2023.108009
Received 10 September 2022; Received in revised form 6 April 2023; Accepted 18 June 2023
Available online 8 July 2023

e.g., the Radio Frequency Identification Device (RFID) sensors are generally installed on the ears of pigs and the legs of laying hens (Maselyne et al., 2016; Siegford et al., 2016). Considering the cost of hardware and maintenance on large-scale commercial farms, contact-based solutions for automated tracking are not preferable. Second, contactless monitoring using video cameras, which has a growing popularity, because of the low cost and sustainability compared with the contact-based sensors.

Most 2D camera-based Multi-Object Tracking (MOT) methods are developed for pedestrians, vehicles or static objects. However, limited studies have been performed for animal tracking. Farm animals within a group usually have a similar appearance and can show various motion patterns. To develop an effective monitoring system for individual animals in commercial farms, we first investigate multiple state-of-the-art methods. By training the networks on a large manually annotated pig dataset, our research facilitates the use of 2D camera-based systems in combination with computer vision and advanced deep learning technologies. In this paper, the focus is on tracking of individual animals to enable long-term analysis of their behavior. Long-term animal tracking is usually hampered by animal occlusion, irregular movements, and abrupt change of direction and action. The intention is to design a robust tracking system that can handle such kind of irregularities.

Three types of tracking approaches exist in literature. The baseline of the first type is the Joint Detection and Embedding (JDE) model, which is a one-shot system that realizes detection and tracking with a single network. This is an anchor-based method that utilizes bounding-box clusters (Wang et al., 2020). We propose to combine JDE with the k-means method to acquire more accurate anchor clusters for specific animal (pig) datasets. The detection re-identification is optimized by a proposed weighted association strategy. Regarding the fixed amount of objects per video in our application, we further constrain the number of objects per frame for detection (Guo. et al., 2022). The second type is FairMOT, derived from the JDE model but improved by an anchor-free method, which significantly reduces the number of identity switches and enhances the tracking performance (Zhang et al., 2021). To this end, a re-identification strategy is proposed for the same objects using data association, which is also employed for FairMOT. The third type is based on a two-stage system, which consists of the detector – You Only Look Once (YOLO) latest Version 5 (Redmon et al., 2016; Jocher et al., 2022), and the tracker – Simple Online and Realtime Tracking with a Deep Association Metric (Deep SORT) (Wojke et al., 2017). For this combination, a weighted association strategy is investigated to re-identify objects in DeepSORT. The performances of the proposed methods are evaluated using videos recorded at a commercial pig farm. The tracking results are compared using state-of-the-art MOT evaluation metrics including MOTA, MOTP, IDF1, ID switches, execution rate, etc. (Heindl, 2017).

The motivating baseline models have been selected based on four aspects: tracking performance, identification accuracy, data-collection efficiency, and model execution rate. The first aspect is tracking performance which is essential for evaluation. Evaluation metrics are required for inspecting the individual animal tracking performance. The proposed baseline models are evaluated on the MOT-16 benchmark (adopted from person tracking Milan et al., 2016), which contains 7 challenging real-world videos of both static scenes and moving scenes. The total set of scenes is composed of 759 tracks with 182,326 bounding boxes in the test set. MOTA is a standard tracking metric and represents multi-object tracking accuracy, including false positive, false negative, and the number of ID switches. The three baseline models JDE, FairMOT, and DeepSORT result in MOTA values of 64.4%, 74.9%, and 61.4%, respectively. The second aspect aims at achieving sufficient identification accuracy. Regarding individual animal monitoring, long and stable tracking for the same animal is desired. High-level animal behavior analysis is feasible only if the tracking procedure is nearly error-free, so that a long tracking pattern can offer more information related to the same animal. However, individual animal tracking has several challenges such as occlusion, close distance between multiple objects, and arbitrary moving directions. In these cases, accurate and stable identification accuracy is required. This means that long tracking trajectories of the same animal lead to a low number of identity switches. Besides this number, another metric for tracking performance is the IDF1 score, which is a percentage score of identity association over time. In terms of the MOT-16 benchmark, JDE results in an IDF1 of 55.8% and 1544 ID switches. FairMOT achieves an IDF1 of 72.8% and 1074 ID switches. DeepSORT gives 781 ID switches. Regarding these two aspects of tracking performances and identification accuracy, the proposed baseline models are competitive with the state-of-the-art methods on the MOT-16 pedestrian dataset.

The third aspect is data-collection efficiency which aims to acquire annotated data effectively. The bounding box is annotated, which fits with the input of the object detector. Compared with other annotation formats e.g. keypoints, polygons, and polylines, the bounding box is sufficient to localize and track objects accurately. Because we seek after the object location and accurate identity, rather than perfectly matched object contour. The bounding box requires less labor and offers the desired accuracy.

The fourth aspect is the model execution rate, which aims at a real-time monitoring system. The requirement for an execution rate is faster than 15 fps. For this purpose, we have initially compared state-of-the-art multi-object tracking methods applied to pedestrian datasets (because the amount of state-of-the-art tracking research on animals is limited). JDE achieves an execution rate of 18.8 fps on a single Nvidia Titan XP GPU. FairMOT executes at 25.9 fps on two RTX 2080 Ti GPUs. Both JDE and FairMOT are joint detection and tracking, which test on the MOT-16 benchmark (pedestrian dataset) with an image resolution of 1088 × 1088 pixels (Wang et al., 2020; Zhou et al., 2019). The detector YOLOv5s detects objects in an image of 416 × 416 pixels with an execution time of 140 fps on a Tesla P100 GPU (Jocher et al., 2022). The DeepSORT tracker achieves an execution time of 40 fps on a GeForce GTX-1050 GPU, which tests on the MOT-16 benchmark with an image resolution of 640 × 640 pixels (Wojke et al., 2017). Based on the above inference times, all proposed models are suitable for the application. Therefore, the proposed baseline models — JDE, FairMOT, and YOLOv5 with DeepSORT, have been adopted for further research, on the basis of the above four aspects.

This work mainly contributes to 3 areas, which are as follows. (1) Investigation and comparison of several state-of-the-art MOT methods for pigs in large-scale farms. (2) Development of data association strategies to optimize multi-object re-identification, by expanding the proposed strategy for three deep learning-based models on a pig dataset. (3) Two manually annotated datasets with bounding boxes of location information for videos recorded at a real commercial farm. The enrichment of the dataset is manually annotated and containing two groups: (a) frames with an annotation interval of 2 s, derived from 96 video recordings including 22,384 frames of 12.44 h in 33 days; (b) frames from continuous videos (without any annotation intervals), based on 5 one-hour recordings from 2 pens in two days (in total 5 h). Each one-hour video contains around 54,000 frames.

The remainder of this paper is organized as follows. Section 2 introduces the related works in recent years. Section 3 describes the workflow of data acquisition and annotation, and elaborates the network architectures of three MOT methods and the proposed enhanced re-identification algorithms. The evaluation methods are also introduced. Section 4 shows experimental details and results. Section 5 discusses the findings of the study and the potential future work accordingly. Finally, Section 6 concludes this work.

2. Related works

Three-dimensional (3D) Kinect cameras with a depth sensor have been used for monitoring pigs (Mallick et al., 2014; Kim et al., 2017).

However, the field of view captured by a 3D camera is rather limited, while the computation based on 3D data is expensive (Matthews et al., 2017). Therefore, more state-of-the-art studies employ two-dimensional (2D) camera monitoring, which provides a broader view, a cost-efficient solution and lower computation requirement. Studies were reported on 2D video-based monitoring and giving visual information on e.g. collective detection of golden shiner fish groups (Davidson et al., 2021), but also multiple pig detection and tracking in indoor sheds (Zhang et al., 2019). This motivates the utilization of 2D RGB cameras for our research. Artificial intelligence presents the emerging topic of 2D video-based individual animal detection and tracking, which has the potential to offer high efficiency. Therefore, 2D video-based tracking can contribute to saving costs, while avoiding contacts with animals.

Video-based animal monitoring requires to localize objects at an individual level and recognize animal behavior status. In order to apply deep learning methods to learn features from animal videos, data preparation is needed with corresponding annotations for specific monitoring objectives. Some research provides the objects and contours by segmentation as input for tracking (Fragkiadaki and Shi, 2011; Thombre et al., 2009), which is more accurate for outlining individual objects. However, preparing contour-based manual segmentation is time-consuming, labor intensive and subjective. It also requires significant computational resources during learning.

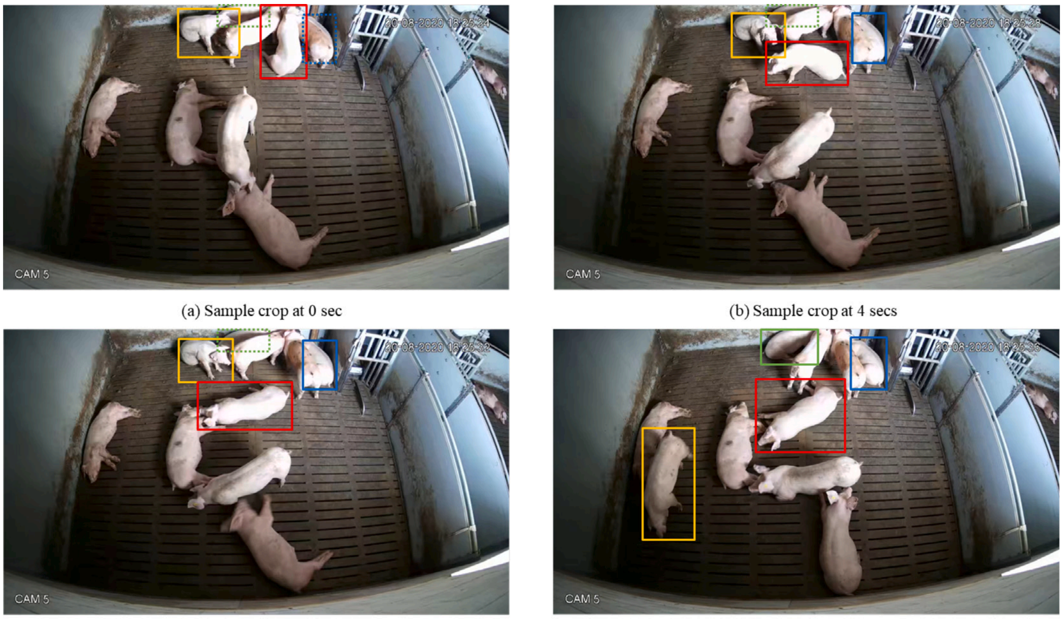
Some research focuses on MOT for pig analysis. Perner (2001) subtract the background from each video frame by segmentation, and group object pixels into an object by the line-coincidence method. A stable monitoring of pig position and its movements are achieved, although the data only records three pigs in one pen. It shows the limitation to be applied on commercial farms, which have 10–11 pigs in one pen. A specific study (van der Zande et al., 2021) applies a tracking-by-detection method of YOLOv3 with SORT on pig datasets. However, it shows a limitation of insufficient training data (4000 frames), which is similar in other research (Riekert et al., 2020) that only has 305 manually annotated frames. In addition, the qualitative evaluation in van der Zande et al. (2021) is not sufficient to validate their tracking performance. For instance, it lacks the state-of-the-art evaluation metrics including IDF1, MOTA and MOTP (Leal-Taixé et al., 2015). Zhang et al. (2019) combine a CNN-based detector with a correlation filter-based tracker, using a novel hierarchical data association algorithm, which achieves a MOTA of 89.58% on 5 testing video segments. However, the evaluation is not based on cross-validation and a comprehensive comparison with the state-of-the-art tracking algorithms is not provided.

All existing studies regarding pig detection and tracking use tracking-by-detection methods, which lead to high computational cost, especially when object numbers increase. In this situation, it is challenging to realize a real-time MOT system for pig tracking. The existing work using MOT for tracking pigs is also limited. In order to explore more MOT methods, we investigate alternative related work developed for other purposes, e.g. pedestrian tracking and vehicle tracking. The Kanade-Lucas-Tomasi (KLT) feature tracker is a widely used vanilla tracking method that is based on the Lucas-Kanade optical flow method and is frequently employed in applications that require the identification of object motion through the association of tracked features between frames (Tomasi and Kanade, 1991). However, the KLT feature tracker's performance may suffer due to the non-rigid motion common in animal behavior and the large deformation resulting from their movements, invalidating its linear motion assumption and resulting in lost tracks. Additionally, the KLT feature tracker lacks the capability to recover lost features or handle occlusions, both of which can lead to tracking failures. For these reasons, we have explored state-of-the-art methods in multi-object tracking that can better address the unique challenges of animal behavior analysis.

As video-based multi-object tracking becomes popular, there are two major directions: (1) tracking-by-detection, and (2) joint detection and tracking.

Tracking-by-detection method is a two-stage system that consists of: (1) object detection, and (2) tracking. In the tracking module, both motion and appearance features are extracted from the detected bounding-box sequence by a re-identification model, followed by associating the detection frame-by-frame. Therefore, tracking results are obtained only in the final stage. Regarding object detection, tremendous work has been investigated utilizing deep learning. Some studies apply the R-CNN (Girshick et al., 2014) and Fast R-CNN (Girshick, 2015). The networks R-CNN and Fast R-CNN use selective search to generate regions for each image and detect objects. However, the regional approach can yield high computation cost. Faster R-CNN (Zhou et al., 2018) utilizes a region proposal network to replace the selective search method, which achieves promising detection accuracy on the PASCAL VOC 2007 (Everingham et al., 2007), 2012 (Everingham et al., 2012), and MS COCO datasets (Lin et al., 2014). Faster R-CNN accelerates the computation compared with R-CNN and Fast R-CNN. However, it is still challenging to reach real-time processing speed in practice. Another popular detector used for object detection is the YOLO detector (Redmon et al., 2016), based on a single network to obtain all bounding-box information on a full image, and subsequently implement regression and classification. The YOLO architecture series are developing fast and have an advantage of fast processing speed by virtue of simpler network architecture. By comparison, Faster R-CNN achieves a Mean Average Precision (MAP) of 87.69%, while YOLOv3 achieves an MAP of 80.17% on the same pill image dataset. However, the detection speed of YOLOv3 is more than eight times faster than that of Faster R-CNN (Tan et al., 2021). The prevailing trackers including SORT (Bewley et al., 2016) and DeepSORT (Wojke et al., 2017), have been widely used for MOT of pedestrians and vehicles (Wojke et al., 2017; Hou et al., 2019; Bathija and Sharma, 2019). SORT is a data association method based on a Kalman filter and a Hungarian algorithm to associate the detected bounding-box results between adjacent frames (Bewley et al., 2016). SORT achieves good performance on the MOT challenge dataset (Leal-Taixé et al., 2015), whereas it has a deficiency in handling occlusions. DeepSORT introduces the comparison of appearance features, which is added to the motion model in SORT. This enhances the performance for a longer duration of occlusion (Wojke et al., 2017). For two-stage tracking-by-detection networks, the optimal models for detection and tracking can be determined individually, whereas the computation cost evaluated for the two individual steps is high to sustain continuous multi-object tracking.

In contrast with tracking-by-detection methods, joining tracking with detection gains significant computational efficiency. Therefore, another research direction is towards a joint detection-and-tracking strategy, also known as a one-shot system. The joint method utilizes a single network to combine detection with tracking, which shares features among these two tasks and saves computational cost. The first category of joint strategies is to perform object detection and re-ID feature extraction in a single network. For instance, TrackR-CNN (Voigtlaender et al., 2019) adds a re-ID branch with Mask R-CNN (He et al., 2017), which obtains mask-based detection with appearance features for each proposal in a regression task. This approach reduces the execution rate. However, the tracking performance is not comparable with two-stage methods. Another research regarding associative embedding is to join object detection and embedding appearance features, used in human-pose estimation (Newell et al., 2017). However, this research can be only applied to single images, while data association in MOT requires association across sequential frames (Newell et al., 2017). The second category is to perform object detection with motion feature extraction in a single network for tracking. For instance, the research on Tracktor (Bergmann et al., 2019) is converting a detector to a model that jointly materializes detection and tracking. It incorporates a single object tracker into a tracked Faster-RCNN detector by extracting the spatial and temporal positions, i.e. trajectories, so it does not rely on training or optimization on tracking data. However,



(a) Sample crop at 0 sec

(b) Sample crop at 4 secs

(c) Sample crop at 8 secs

(d) Sample crop at 12 secs

Fig. 1. Sample frames taken from a period of 12 s with a time difference of 4 s, for pigs recorded at Volmer farm in Germany. Several challenging situations are highlighted. Pigs in the green and blue bounding boxes are occluded by neighbors in the first several frames. The pig in the yellow bounding box moves fast between (c) and (d), which causes a large deformation of the bounding-box size. As can be observed that pigs in the yellow and red bounding boxes have a highly activity level, where the bounding boxes have variable width/height ratios and sizes. Compared to pedestrian data, the head direction of a pig is unpredictable.

Tracktor is only capable of addressing most of the less challenging tracking scenarios (i.e. no small and occluded objects or missing detection). Regarding complex situations, the accuracy is not stable, which may be caused by lacking an additional embedding model to extract appearance features (Wang et al., 2020).

3. Materials and methods

3.1. Data description

The research data involved in this study is based on videos recorded at a real commercial farm. The data is first acquired, followed by manual annotation and data curation. Video recordings are collected at the pig farm of Volmer in Germany. The Animal Welfare Body of Wageningen University & Research (Wageningen, the Netherlands) approved the protocol of the study (211223_LB_IMAGEN). The procedure is in accordance with the Dutch law on animal experimentation, which complies with the European Directive 2010/63/EU on the protection of animals used for scientific purposes. Fig. 1 shows four sample frames taken from a period of 12 s with a time difference of 4 s. Several challenging situations are highlighted. Pigs in the green and blue bounding boxes are occluded by neighbors in the first several frames. The pig in the yellow bounding box moves fast between Fig. 1(c) and (d), which causes a large deformation of the bounding-box size. It can be observed that pigs in the yellow and red bounding boxes have a highly active level, where the bounding boxes have variable width/height ratios and sizes. Compared to pedestrian data, the head direction of a pig is unpredictable. Pigs from in total 10 pens are recorded, where most pens contain 11 pigs with or without sprayed color marks on their backs. The group composition of the pigs usually remains constant unless special situations occur, like e.g. sickness or injury. Most pens are installed with one single camera which captures the side view towards the pen's floor at ceiling height (average pen size (length×width×height): 372.5 × 288.75 × 280 cm^3), covering the entire pen. The cameras used for recording are LOREX 4KSDAI168 with an image resolution of 1280 × 720 pixels, and a frame rate of 15 fps. Pig videos are recorded continuously on a 24/7 basis, and each video is automatically generated and stored per hour.

All manually annotated frames are processed with the Computer Vision Annotation Tool (CVAT) (Openvinotoolkit, 2020), which is a software package for object annotation. Video segments showing active pig movements are selected, followed by annotating the pig location in each video frame with a consistent identity associated for each individual pig. CVAT is capable of labeling object-location information using a rectangular-shape bounding box, and also provides the options for adding occlusion conditions. Aiming at more effective annotation work, we have chosen to annotate bounding boxes rather than object contours. CVAT supports saving the frame ID, object ID, bounding-box location, and size of the object.

3.2. Network architecture overview

3.2.1. Joint Detection and Embedding (JDE)

JDE is aiming at a one-shot system, which combines detection and tracking in a single network by adding a re-ID embedding branch in parallel. The jointly learned features are shared for two objectives, which are to localize the objects and to associate identities between continuous frames with the appearance embedding. In this way, the system reduces the computation cost and enhances the tracking efficiency (Wang et al., 2020). The baseline network in JDE is derived from YOLOv3, which is the DarkNet-53 - a Feature Pyramid Network (FPN). As described in Fig. 2, the feature maps at multiple scales of an input video frame are first acquired. Second, feature maps are fused by a skip connection between the feature maps of the smallest scale and the second smallest scale, similarly to other scales. Ultimately, each fused feature map is attached to a prediction head, which generates a dense prediction map with three branches: box classification, box regression, and appearance embedding. The detection branch is responsible for the first two tasks: foreground/background classification with a cross-entropy loss, and bounding-box regression with a smooth L_1 loss. The learning procedure of appearance embedding in JDE is to derive a small distance measure for detected bounding boxes for the same identity, whereas bounding boxes for different identities achieve a large distance. The experimental results with pedestrian datasets show that the cross-entropy loss gives the best results (Wang et al., 2020). Hence, the

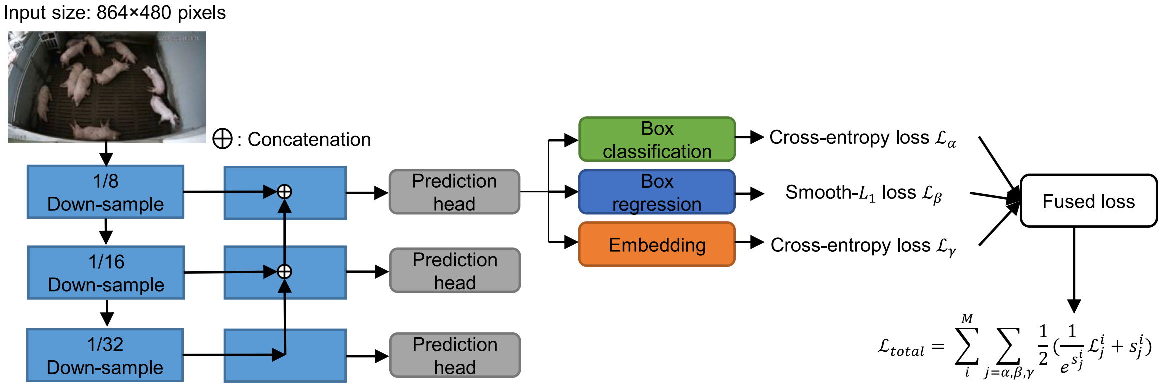

Input size: 864×480 pixels
⊕: Concatenation
1/8 Down-sample
1/16 Down-sample
1/32 Down-sample
Prediction head
Prediction head
Prediction head
Box classification
Box regression
Embedding
Cross-entropy loss $\mathcal{L}_\alpha$
Smooth-L_1 loss $\mathcal{L}_\beta$
Cross-entropy loss $\mathcal{L}_\gamma$
Fused loss
$\mathcal{L}_{total} = \sum_{i}^{M} \sum_{j=\alpha,\beta,\gamma} \frac{1}{2}(\frac{1}{e^{s_j^i}}\mathcal{L}_j^i + s_j^i)$

Fig. 2. JDE network architecture with prediction heads, including the box classification, box regression and embedding (Wang et al., 2020).

appearance embedding learning of JDE is based on using cross-entropy loss. To fuse all the losses, the system adopts an automated learning scheme by using the concept of task-independent uncertainty (Kendall et al., 2018). The total loss shown in Fig. 2 is specified by

$$\mathcal{L}_{total} = \sum_{i}^{M} \sum_{j=\alpha,\beta,\gamma} \frac{1}{2}(\frac{1}{e^{s_j^i}}\mathcal{L}_j^i + s_j^i) , \quad (1)$$

where M is the number of prediction heads, $i = 1, \ldots, M$, while $j = \alpha, \beta, \gamma$ are branch losses. Parameter s denotes the task-dependent uncertainty, which is a learnable parameter that can be adjusted for each loss. When the loss of one task increases, the related learnable parameter decreases. This parameter optimizes one task as much as possible, but without affecting other tasks.

JDE adopts a simple and fast online association algorithm. Each track consists of an appearance state and a motion state. The appearance affinity matrix is calculated by cosine similarity and the motion affinity matrix is computed using the Mahalanobis distance. A buffer pool is set for potential tracks to the subsequent association. For each frame, there is a re-identification calculation between all detection and tracks in the buffer pool. The Hungarian algorithm (Kuhn, 1955) solves the linear assignment to output matched tracks, unmatched tracks and detections. A Kalman filter (Welch et al., 1995) is used to update and predict the locations in the current frame from the existing tracks.

K-means clustering is also applied to the training dataset to recalculate 12 anchors, where each scale has 4 anchors. For the widely used pedestrian datasets in MOT, a filter condition is normally applied to constrain the object aspect ratio to 1/3 (width/height). We remove this constraint because more deformations are expected in the pig datasets. Furthermore, we fix the amount of objects in each frame with a non-maximal suppression (NMS) setting.

3.2.2. Enhanced re-identification association on JDE

The workflow of data association in JDE is depicted drawn in Fig. 3. In total, there are three online association steps to handle the detections and tracks. We adopt all detected bounding boxes in the first frame as the initial tracks. The first association is related to the embedding distance with fused motion. After calculating the Hungarian algorithm (Kuhn, 1955), the unmatched tracks and related detections are further imported to the second association. The original second association applies the Intersection over Union (IoU) comparison. In the third ID association step, the IoU distance is adopted to handle the unconfirmed tracks, which are usually tracked with only one initial frame. A buffer pool is used for storing lost tracks, and the tracks are removed when they are lost for more than a certain frame-count duration (threshold). In the end, the outputs combine all followed tracks, activated tracks, and refined tracks.

The shape of the bounding box for pedestrian detection has a comparative regular width/height ratio. The common width/height ratio of pedestrians is 1/3. However, in the captured videos, the position and posture information of pigs cannot be expected, since they are under a “free”-living mode in the commercial pens (see Fig. 1). In this regard, diverse width/height ratios and sizes of the bounding box from one pig are possible. In this way, animals are more challenging to track than pedestrians. However, in the second step for data association in the original network architecture, only a single IoU calculation is performed, where only the percentage of the overlap between two boxes is considered. The bounding box that outlines a single pig can change substantially between adjacent frames according to pig status, yielding to significant changes of bounding-box shapes. This means the re-ID method of calculating the IoU for associating the pedestrian detection is not appropriate for pig detection. We propose to add another association term based on extracted appearance-embedding features to enhance the re-ID method for associating the pig detection. Therefore, we introduce an enhanced weighting strategy in the second association (see green marks in Fig. 3), specified by

$$dis_{total} = \omega_F \cdot dis_{IoU} + (1 - \omega_F) \cdot dis_{emb} , \quad (2)$$

where dis_{total} is the total distance of the second association, parameter ω_F is a gain parameter. Parameter dis_{IoU} represents the IoU distance and the distance dis_{emb} indicates the embedding distance with fused motion. This distance parameter considers the weighted distance between the IoU and fused embedding with motion, instead of only relying on the IoU distance after the first embedding comparison (Guo. et al., 2022).

3.2.3. FairMOT

Similar to JDE, FairMOT is also a one-shot system. The backbone network used in FairMOT is ResNet-34, which trades-off tracking performance and computing time. To fuse multi-layer features as JDE, a developed version of Deep Layer Aggregation (DLA) (Zhou et al., 2019) is attached to the backbone, as shown in Fig. 4. The tuning of the network adds more skip connections between multiple scales, which is similarly explored in FPN. Moreover, deformable convolution layers in all up-sampling stages are used, to enable dynamic adjustment among object scales and poses. The entire network is called DLA-34.

Compared with JDE, FairMOT addresses three unfair issues caused by anchors, features, and feature dimensions. The unfairness caused by the anchor-based method shows all active anchors around the object center are considered as candidates of re-ID features. The adjacent anchors have a high possibility to be confirmed as being the same identity if their IoU values are high enough, which results in sub-optimal extracted features. FairMOT solves this unfairness by extracting the re-ID feature only from the center of the object. In addition, FairMOT improves the setting of the feature dimension, whereas the performance is higher when the network learns lower-dimensional features. The object detection branch in FairMOT is based on the anchor-free object detection architecture CenterNet. It leaves out the steps for computing clusters from all bounding boxes. As can be observed in Fig. 4, three

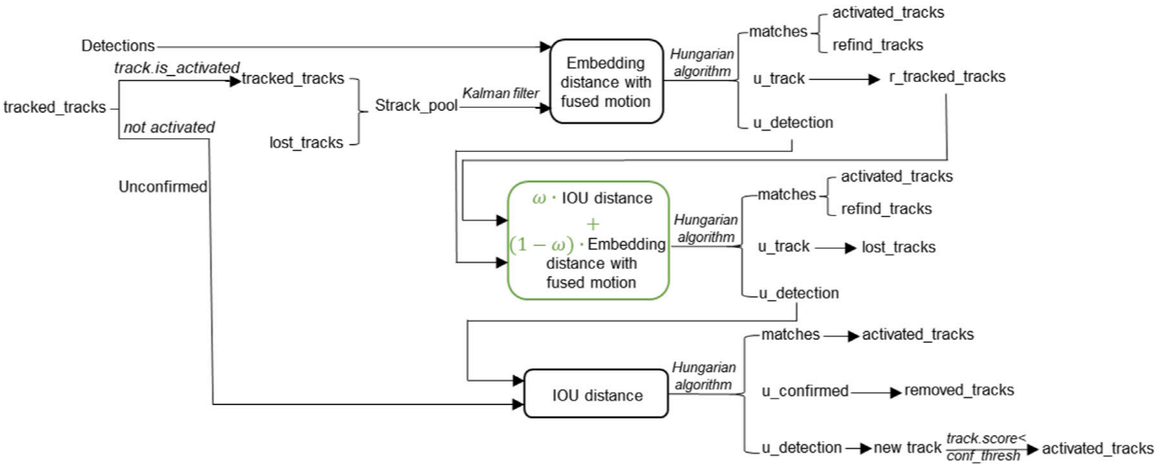


Fig. 3. Enhanced re-identification association workflow of JDE and FairMOT methods. The new proposed weighted association strategy is highlighted in the green block. This distance parameter considers the weighted distance between the IoU and fused embedding with motion, instead of only relying on the IoU distance after the first embedding comparison.

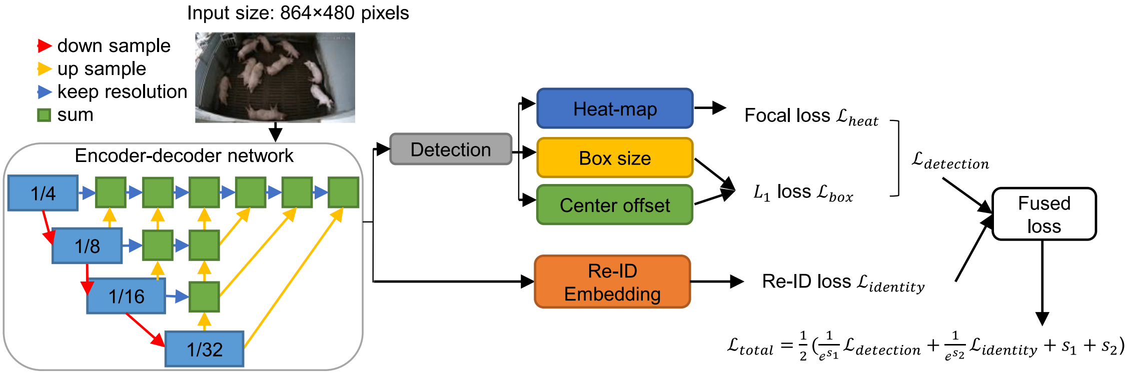

Input size: 864×480 pixels
down sample
up sample
keep resolution
sum
Encoder-decoder network
1/4
1/8
1/16
1/32
Detection
Heat-map
Box size
Center offset
Re-ID Embedding
Focal loss $\mathcal{L}_{heat}$
L_1 loss $\mathcal{L}_{box}$
$\mathcal{L}_{detection}$
Re-ID loss $\mathcal{L}_{identity}$
Fused loss
$\mathcal{L}_{total} = \frac{1}{2}(\frac{1}{e^{s_1}}\mathcal{L}_{detection} + \frac{1}{e^{s_2}}\mathcal{L}_{identity} + s_1 + s_2)$

Fig. 4. FairMOT network architecture with prediction branches, including the heat-map, box size, center offset, and re-ID embedding (Zhang et al., 2021).

parallel heads contribute to the detection branch. The heat-map head predicts the locations of the object centers with a focal loss. The box-offset head and the box-size head are responsible for more accurate localization and estimating the height and width of the target box, optimized by the L_1 loss. As shown in Fig. 4, FairMOT introduces a re-ID branch to generate object features, aiming at distinguishing different objects. The re-ID features are extracted from the feature map, which are derived from a convolution layer with 128 kernels, based on the backbone network. The automated loss balancing in FairMOT is the same as used in the JDE network. It consists of the detection loss and re-ID loss. The detection loss is the sum of $\mathcal{L}_{heat}$ and $\mathcal{L}_{box}$. The total loss shown in Fig. 4 is specified by

$$\mathcal{L}_{total} = \frac{1}{2}\left(\frac{1}{e^{s_1}}\mathcal{L}_{detection} + \frac{1}{e^{s_2}}\mathcal{L}_{identity} + s_1 + s_2\right) , \tag{3}$$

where s is the same as the parameter in the JDE method.

3.2.4. Enhanced re-identification association on FairMOT

The online association strategy in FairMOT takes a prevailing tracking method, similar to the JDE method. We also explore the proposed weighted strategy in the FairMOT approach, which is expected to utilize more appearance embedding features and reduce the number of identity switches during tracking.

3.2.5. YOLOv5 and DeepSORT

Another proposed method is a two-stage system, which is following a tracking-by-detection strategy. The first stage is to localize objects by a detector for each input frame. Meanwhile, appearance and motion features are extracted from the detected bounding-box sequences. Appearance feature association helps to reduce identity switches, while motion features are processed by a Kalman filter to predict object location in the next frame. The second stage obtains identity association between adjacent frames using a data-association algorithm. The strength of a two-stage system is that each of the two stages can be optimized individually. However, it increases computation cost, which is not desired for continuous 24/7 tracking with MOT in practice. In this research, we employ a combination with the You Only Look Once (YOLO) Version 5 detector (Jocher et al., 2022) and the Simple Online and Real-time Tracking with a Deep Association Metric (DeepSORT) (Wojke et al., 2017). We aim at using the YOLO network with lower complexity to improve the model efficiency. Therefore, our baseline employs the small YOLOv5 model (further referred to as YOLOv5s), which reduces the complexity of the network for faster training and inference (see Fig. 5). The core concept of YOLO is to convert the object detection to a regression problem. YOLO utilizes a full image as input to a single deep neural network to derive the bounding boxes. The architecture of YOLO consists of three components: (1) a backbone convolution neural network (CNN) that extracts appearance features with different sizes, (2) a neck network including a set of layers that integrates features and then passes them to the prediction layer, and (3) a head incorporating features along with the bounding-box predictions and then classifying the predictions including regression to finalize the detection stage (Redmon et al., 2016). The architecture of YOLOv5s employs CSPDaeknet as the backbone for appearance-feature extraction, PANet as the neck for generating FPN to pass features to the prediction head, and convolutional layers as the head to output the

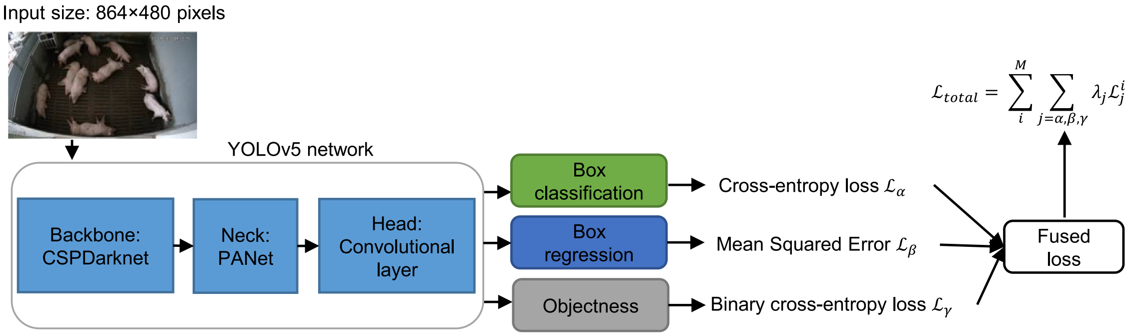

Input size: 864×480 pixels
YOLOv5 network
Backbone: CSPDarknet
Neck: PANet
Head: Convolutional layer
Box classification
Box regression
Objectness
Cross-entropy loss $\mathcal{L}_\alpha$
Mean Squared Error $\mathcal{L}_\beta$
Binary cross-entropy loss $\mathcal{L}_\gamma$
Fused loss
$\mathcal{L}_{total} = \sum_{i}^{M} \sum_{j=\alpha,\beta,\gamma} \lambda_j \mathcal{L}_j^i$

Fig. 5. YOLOv5 network architecture with prediction branches, including the box classification, box regression, and objectness (Jocher et al., 2022).

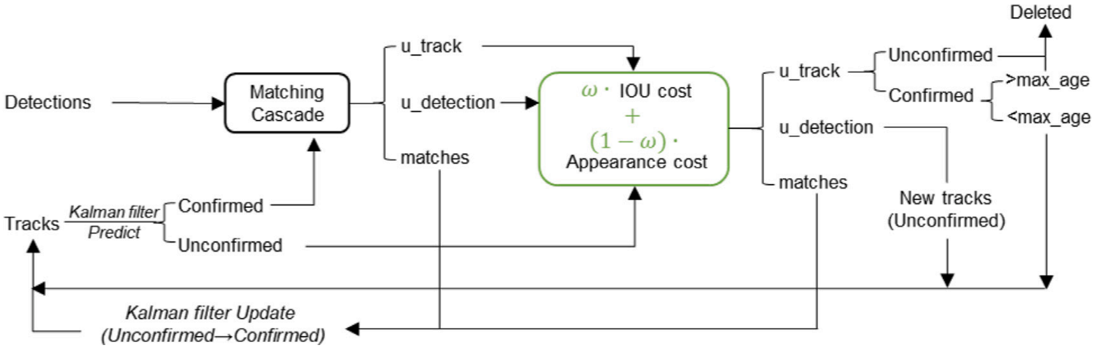


Fig. 6. Enhanced re-identification association workflow of DeepSORT method. The proposed weighted association strategy is highlighted in the green block. The cost matrix considers the weighted distance between the IoU and appearance, instead of only relying on the IoU comparison after the matching cascade.

predicted bounding boxes (Xu et al., 2021). The multi-task loss function of YOLO is specified by

$$\mathcal{L}_{total} = \sum_{i=1}^{M} \sum_{j=\alpha,\beta,\gamma} \lambda_j \mathcal{L}_j^i \ , \tag{4}$$

where M is the number of individual tasks, and $j = \alpha, \beta, \gamma$ are hyperparameters to constrain the gain of the loss.

DeepSORT is a conventional tracker used in two-stage systems (Wojke et al., 2017). It presents the track handling and state estimation to associate frame-by-frame detection. Apart from the Hungarian algorithm (Kuhn, 1955) and the Kalman filter (Welch et al., 1995) which are also used in JDE and FairMOT, the core component in DeepSORT introduces a matching cascade algorithm to effectively solve the presence of occlusions, which reduces the identity loss significantly (Wojke et al., 2017).

3.2.6. Enhanced re-identification association on YOLOv5 and DeepSORT

As shown in Fig. 6, the baseline association strategy calculates the association steps in two stages. Based on all detected bounding boxes in the first frame, we consider those as initial tracks. The first association step uses the matching cascade, related to the appearance-based comparison, weighted by the Mahalanobis distance and the cosine distance. After distribution by the Hungarian algorithm (Kuhn, 1955), the unmatched tracks and detections are further imported to the second association stage. The original second association is only an IOU comparison to sort out the matched tracks, unmatched detection, and track candidates. If the status of a track remains unconfirmed after a certain number of frames, indicated by a threshold, it will be removed.

The motivation of an enhanced re-ID association is illustrated in Section 3.2.1. DeepSORT performs a separate IoU calculation, which enables the use of appearance-embedding features to help data association. Similar to the JDE and FairMOT methods, we propose a weighted strategy to employ feature fusion in YOLOv5s with DeepSORT (see green marks in Fig. 6). Here, we introduce a weighted strategy, specified by

$$Cost_{total} = \omega_D \cdot Cost_{IoU} + (1-\omega_D) \cdot Cost_{appearance} \ , \tag{5}$$

where $Cost_{total}$ is the total cost matrix of the second association, and ω_D is a gain parameter. Variable $Cost_{IoU}$ is the IoU matching cost and $Cost_{appearance}$ is the cost matrix of a matching cascade with the Mahalanobis distance and the cosine distance. The total cost considers the weighted distance at the second step to include richer feature comparison of the IoU cost and the appearance-based cost, instead of only relying on the IoU comparison after the matching cascade.

3.3. Evaluation methods

3.3.1. Evaluation metrics

The proposed methods are evaluated using the metrics derived from the MOT challenge based on the pedestrian datasets (Leal-Taixé et al., 2015), combined with evaluation metrics used in evaluating the JDE (Wang et al., 2020) and FairMOT methods (Zhang et al., 2021). These metrics are employed and listed in Table 1 that illustrates metric terminology for evaluating the proposed systems. The upward arrow after the name of the metric indicates that a higher value of this term is desired, while a downward arrow after the metric highlights that a lower value is better. It should be noted that the different metrics are not equally important. MOTA considers detection performance, using the overlap between the detected box and ground-truth box, and a confidence score to constrain the detected box. In this way, MOTA significantly outweighs the association performance. However, our tracking application does not aim for a large overlap between the detected box and the ground-truth box. Regarding individual animal monitoring, long and stable tracking for the same animal is more important and desirable. Therefore, we mainly focus on the detected identification accuracy IDF1 and the number of ID switches, rather than striving for a high overlap by MOTA.

Table 1
Evaluation metrics and terminology for the proposed MOT methods (arrows indicate preferred value optimization).

Metric	Description
MOTA↑	Multi-object tracking accuracy. This measure combines three error sources: false positives, missed targets, and identity switches.
MOTP↓	Multi-object tracking precision. The misalignment between the annotated and the predicted bounding boxes.
MT↑	Number of mostly tracked trajectories.
PT	Number of partially tracked trajectories.
ML↓	Number of mostly lost trajectories.
IDF1↑	ID F1 score. The ratio of correctly identified detection over the average number of ground-truth and computed detection.
IDs↓	Number of identity switches.
FPS↑	Execution rate, frames per second.

3.3.2. K-fold cross-validation

The proposed MOT systems are evaluated using a K-fold cross-validation. Every sample frame can be utilized in either the training or the testing dataset, which results in a model validation with less bias. We randomly split the entire dataset into K groups, where each group is involved in the training procedure for K − 1 times, and one time for the test procedure. The evaluation metrics are accumulated for each method of each training cycle and finally averaged for measuring the performance (Rodriguez et al., 2009).

3.3.3. T-test

The t-test is utilized to validate the effect of the proposed strategy. The two-tailed paired t-test is a valuable statistical test that is appropriate for comparing the baseline model and enhanced model, due to the evaluation metrics obtained from the same groups. This test determines whether the enhancement brought by the proposed strategy is statistically meaningful, or is merely an occasional event. For this study, a significance threshold of a p-value $p = 0.05$ is used to indicate that above the threshold the null hypothesis cannot be rejected when $p > 0.05$. Conversely, a p-value of $p \leq 0.05$ denotes a statistically significant test result, indicating that the enhancement is statistically significant (Kim, 2015).

4. Results

4.1. Summarization of dataset

Our manually annotated datasets of pigs are divided into two groups. The first group is used for K-fold cross-validation. It consists of 22,384 frames including 250,638 annotated bounding boxes from 10 pens recorded from different daytime periods. The videos are selected from 33 days of video recordings. As shown in Table 2, we have 32 short segments with around 100 frames of 3 min and 20 s, and 64 long segments of around 300 frames of 10 min. The total duration is about 12.44 h for 96 videos. All videos are recorded at a frame rate of 15 fps. A frame step of 30 frames (2 s) is taken during annotation, i.e. to output one frame per 30 frames for effective ground-truth acquisition. All frames are selected from daytime scenes in an uncontrolled farming environment. For evaluating the generalization of the proposed models, videos under various conditions are captured and selected according to the activity levels of pig movements, occurrence of occlusion, or group stacking. The summary shown in Table 3 describes the data distribution of the collected data, which is later used for implementing K-fold cross-validation. We divide the dataset into $K = 8$ groups. Each group is randomly assigned with 4 short segments and 8 long segments, including around 2800 frames in total. Each group is involved in 7 cycles of training, combined with 1 test as a procedure cycle. Each cycle of the procedure uses around 19,600 frames for training, while each test step after this employs around 2800 frames. To test the overall tracking performance, the evaluation metrics are calculated, based on the average values over all testing splits.

The second group of the pig dataset is used for validating the proposed system on long-duration tracking scenarios. The dataset consists of 5 one-hour videos from 2 pens in 2 days. The ground truth of one-hour videos is annotated on continuous video frames, where each video contains around 54,000 frames. The data description is shown in Table 4. Video Nos. 1-4 are recorded at two different time periods. Video No. 5 is recorded at one time period. The activity levels among pigs are illustrated in Table 4.

Table 2
Summary of dataset distribution for K-fold cross-validation.

Video data	No. of frames	Duration	No. of videos
Short segments	≈100/segment	≈3 m 20 s/segment	32
Long segments	≈300/segment	≈10 m/segment	64
Overall	22,384	≈12.44 h	96

Table 3
Summarization of the dataset in each fold for K-fold cross-validation, K = 8.

Dataset	No. of frames	Duration	No. of videos
Training	≈19,600	≈653.33 m	84
Testing	≈2800	≈94.67 m	12

Table 4
Summarization of the long-duration dataset, including 5 one-hour video segments.

Video No.	Channel No.	Date	Time	Activity level
v1	ch2	2022-10-22	08:00–09:00	Inactive
v2	ch2	2022-10-22	16:00–17:00	Medium
v3	ch2	2022-10-23	08:00–09:00	Mild
v4	ch2	2022-10-23	16:00–17:00	Highly
v5	ch3	2022-10-22	08:00–09:00	Mild

4.2. Implementation overview

This section gives an overview of the implementation details of the proposed methods. The resolution of the input frames for all models is set to 864 × 480 pixels for a fair comparison. All experiments are executed on a RTX 2080Ti GPU.

4.2.1. JDE

The backbone network of JDE is DarkNet-53 (Redmon and Farhadi, 2018). Twelve clusters of anchor boxes are derived from all training bounding boxes by a k-means clustering method (Lloyd, 1982). Three key parameters – learning rate, batch size and epoch count – are determined, experimentally considering the best loss convergence and the highest accuracy. The training model is based on a learning rate of 0.0001, optimizing under standard SGD. The training is performed for 50 epochs with a batch size of unity. The input video frames are resized to 864 × 480 pixels. The weight for enhanced association is set to 0.8.

4.2.2. FairMOT

The backbone network of FairMOT is DLA-34 (Zhang et al., 2021). The model is trained with a learning rate of 0.0001, optimized with the Adam optimizer. The training is performed for 50 epochs with a batch size of 2. For the sake of efficiency, the resolution of input images is resized to 864 × 480 pixels. The gain parameter w for the weighted association is 0.8.

Table 5
Comparison between the original re-ID association method and the proposed weighted-association strategy.

Testing data	Method	Re-ID association method	IDF1↑	MOTA↑	GT	IDs↓
8-fold testing data	JDE	Original	66.19	**83.56**	128	554
		Weighted	**66.60**	82.74		**514**
	FairMOT	Original	78.97	88.44		259
		Weighted	**80.94**	**88.55**		**213**
	YOLOv5s with DeepSORT	Original	**66.46**	**88.99**		**486**
		Weighted	65.76	88.74		499

Table 6
Average tracking results of enhanced JDE and FairMOT, and original YOLOv5s+DeepSORT using 8-fold cross-validation.

Method	IDF1↑ ± SD↓	MOTA↑ ± SD↓	MOTP↑ ± SD↓	GT	MT↑	PT	ML↓	IDs↓	FPS↑
Enhanced JDE	66.60 ± 3.85	82.74 ± 3.13	81.00 ± 1.01	128	113	15	0	514	36.13
Enhanced FairMOT	**80.94** ± 4.81	88.55 ± 3.45	82.60 ± 1.70	128	**123**	5	0	**213**	**48.87**
Original YOLOv5s+DeepSORT	66.46 ± 9.80	**88.99** ± 8.40	**89.78** ± 0.89	128	117	9	2	486	22.24

4.2.3. *YOLOv5s with DeepSORT*

YOLOv5s with DeepSORT includes two stages. The first stage is the small YOLOv5 architecture, which consists of the backbone of CSP-Daeknet, the neck of PANet, and the head of convolutional layers. The training model has a learning rate of 0.01, optimizing under standard SGD. The training is performed for 100 epochs with a batch size of 2. The second stage on tracking utilizes the association metric that was learned based on OSNet (Zhou et al., 2021) for the purpose of person re-ID. The input video frames are also resized to 864 × 480 pixels. The enhanced association uses a weight of 0.9.

4.3. *Results*

This section illustrates the achieved results in 3 areas: (1) the benefit of the proposed weighted-association algorithm, (2) overall short-duration results of 8-fold cross-validation, and (3) validated results on long-duration videos.

4.3.1. *Benefit of weighted association*

The models obtained with the proposed methods are evaluated using 8-fold cross-validation. Table 5 shows the comparison of the original association and the weighted association, in terms of IDF1, MOTA and the number of ID switches (bold numbers are best). The evaluation metrics are averaged over all 8-fold testing sets. In terms of the JDE method, the weighted strategy outperforms the original association on IDF1. The number of identity switches is 40 lower than the original association. However, the value of MOTA decreases by 0.82% with the weighted association. The overall performance of FairMOT is better than JDE, especially concerning IDF1, MOTA, and the number of identity switches. As can be observed in Table 5, FairMOT with a weighted-association strategy reduces 46 identity switches, which is the most effective tracking system. The weighted association in YOLOv5s with DeepSORT does not perform as desired, because the performances of most evaluation metrics decrease. The execution rates between original and weighted association on each method are similar. Through the comparison above, it is demonstrated that the proposed algorithms based on JDE and FairMOT benefit the tracking performance. It can be concluded that the added appearance features proposed in the weighted association improve the tracking performance. However, the weighted strategy shows limitations when it is applied on YOLOv5s with DeepSORT.

4.3.2. *8-fold cross-validation results*

From Table 5, the best performing version of each algorithm is chosen for further comparison with average numbers. This further comparison is shown in Table 6, which contains the average 8-fold cross-validation results of the best version of the proposed methods. To show the reliability of the proposed tracking methods, the Standard Deviation (SD) of 8-fold cross-validation results is calculated for IDF1, MOTA and MOTP. The tracking performance of the enhanced JDE method is not optimal, while the execution rate is faster than the video frame rate. The enhanced FairMOT outperforms the enhanced JDE and original YOLOv5s with DeepSORT in most evaluation metrics, especially in two aspects. (1) The number of identity switches that is around 300 less than enhanced JDE and 273 less than original YOLOv5s with DeepSORT. (2) The execution rate of the enhanced FairMOT is 1.35 times faster than enhanced JDE and 2.20 times faster than original YOLOv5s with DeepSORT. Original YOLOv5s with DeepSORT achieves the highest MOTA and MOTP metrics among all methods. However, the original YOLOv5s with DeepSORT obtains a large number of identity switches, and an undesired result of IDF1.

Summarizing, the enhanced FairMOT provides the best results among the proposed methods. To prove the effectiveness of the enhanced re-identification association strategy, there are multiple comparisons to highlight the proposed strategy for 8 consecutive cycles. Figs. 7 and 8 present IDF1 results and the number of ID switches for the original FairMOT and the weighted FairMOT. It can be observed that the weighted strategy version of FairMOT reduces the ID switches, thereby improving tracking performance for virtually each cycle. Regarding the comprehensive data of each fold in Figs. 7 and 8, we perform two paired t-tests for comparing the performance of the conventional and weighted FairMOT. Notably, one t-test is deployed for evaluating the IDF1 scores, while the other is inspecting the number of ID switches. The statistical results verify that the p-values corresponding to IDF1 and ID switches are as low as 0.0018 and 0.0045, respectively. A p-value less than 0.05 indicates a significant difference, whereas a p-value larger than 0.05 represents no substantial difference. Thus, the enhanced FairMOT has a noticeable positive impact on multi-object tracking performance for pigs, as confirmed by the remarkably significant results ($p = 0.0018$ and $p = 0.0045$).

To assess the effectiveness of our proposed methods, we compare them against two state-of-the-art tracking-by-detection benchmark models for multi-object tracking. We use the YOLOXs detector to obtain detection results at the first stage (Ge et al., 2021). We incorporate the two-stage MOT systems by employing two trackers, SORT and ByteTrack (Bewley et al., 2016; Zhang et al., 2022). To ensure a fair comparison, we use the same dataset and partitioned folds as our proposed models, and conduct eightfold cross-validation for the comparative experiments. We present a comparison between our proposed methods and the two additional state-of-the-art methods, as demonstrated through eightfold cross-validation in Table 7. YOLOXs+SORT achieves the highest MOTA among all the evaluated models, scoring 89.4%, but it results in a higher number of ID switches, leading to a lower IDF1 value. Compared to the one-shot methods, the two-stage methods have a significantly faster inference time. Our enhanced FairMOT outperforms all other state-of-the-art baseline models, achieving both the highest IDF1 value and the lowest number of ID switches.

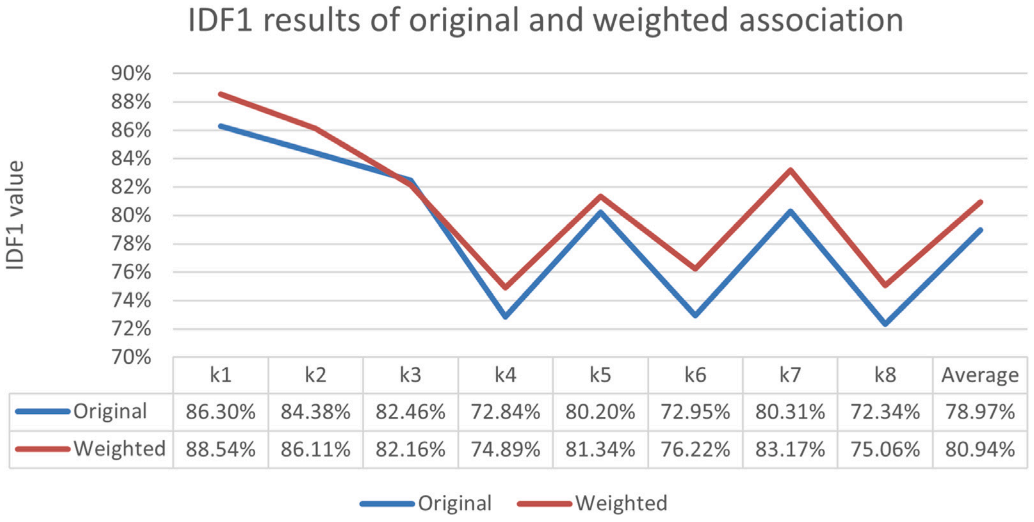


Fig. 7. IDF1 results of FairMOT with the original and weighted re-ID association (higher value is expected, k1–k8 are the names of each test fold used in the eightfold cross-validation).

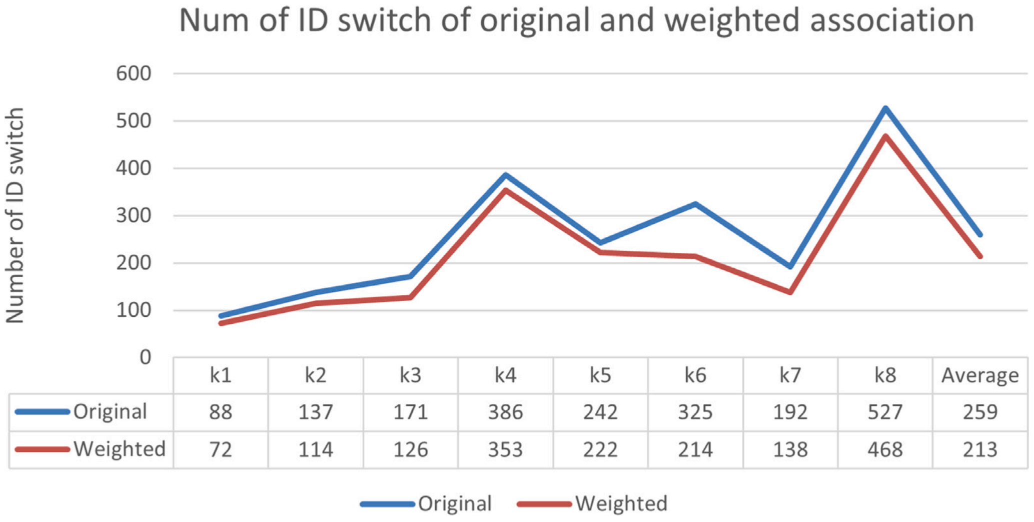


Fig. 8. Number of ID switch results of FairMOT with the original and weighted re-ID association (lower value is expected, k1–k8 are the names of each test fold used in the eightfold cross-validation).

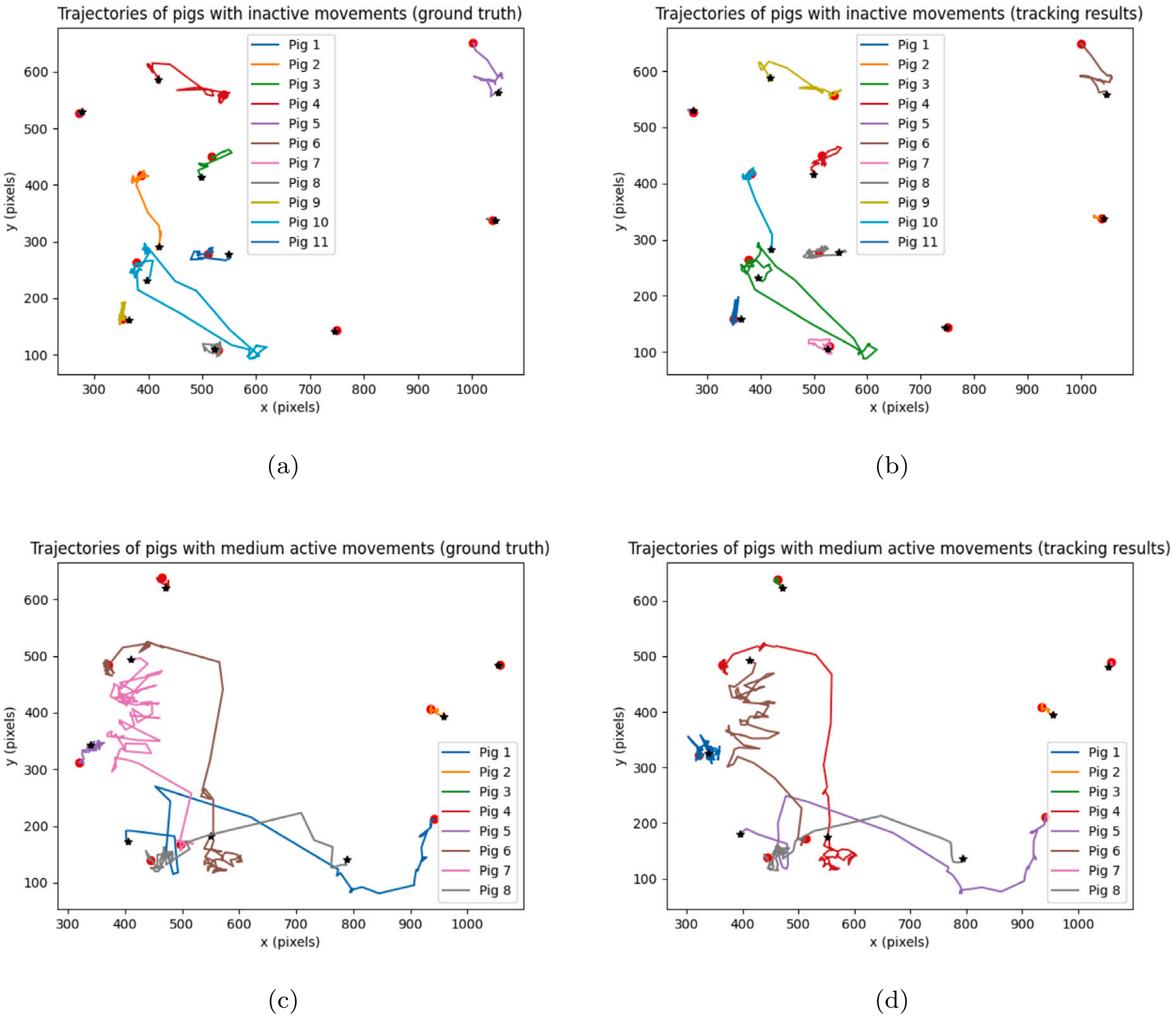
Table 7
Comparison of the state-of-the-art methods under 8-fold cross-validation.

Method	IDF1↑	MOTA↑	MOTP↑	GT	MT↑	PT	ML↓	IDs↓	FPS↑
JDE	66.19	83.56	81.41	128	112	16	0	554	36.13
FairMOT	78.97	88.44	82.80	128	122	6	0	256	**48.87**
YOLOv5s+DeepSORT	66.46	88.99	**89.78**	128	117	9	2	486	22.24
YOLOXs+SORT	69.60	**89.48**	85.02	128	116	12	0	537	14.35
YOLOXs+ByteTrack	73.18	86.80	81.77	128	120	8	0	425	14.65
Enhanced JDE	66.60	82.74	81.00	128	113	15	0	514	36.13
Enhanced FairMOT	**80.94**	88.55	82.60	128	**123**	5	0	**213**	**48.87**

Apart from the quantitative evaluation metrics, we also visualize pig trajectories (see Fig. 9). The enhanced FairMOT yields the least number of identity switches, as shown in Table 6, thus we apply it on two short video segments of 3 minutes and 20 s. Pigs recorded in these two segments have different activity levels. The red point represents the pig departure position and the black star symbol points to the pig destination position. Fig. 9(a) and (b) depict a pen housing 11 pigs, with most of them displaying mild active movements around the area nearby their initial position. Fig. 9(c) and (d) portrays a pen containing 8 pigs. Among them, there are 4 pigs (Pig 4, 5, 6, and 8 in Fig. 9(d)) exhibit medium-activity and move away from their initial positions, while the rest remain relatively static. The enhanced FairMOT demonstrates highly accurate tracking results, as evidenced by the visualization comparison of the ground truth and the tracked trajectories, with continuous trajectories and no instances of identity switching in both videos.

4.3.3. *Validation on long-duration recordings*

To validate the feasibility and reliability of the tracking model, we adopt the model that shows the best performance from the k-fold cross-validation for a further validation. Five one-hour video segments are used to validate the model robustness and reliability for long-duration situations, with the aim of continuous real-time monitoring in real practical use. The enhanced FairMOT is chosen for the experiments. The results of the original FairMOT are also presented for comparison with the weighted FairMOT model. Pigs are normally mildly active in the morning, and Video Nos. 1, 3, and 5 are selected for this condition. Video Nos. 2 and 4 record medium and highly active movements among pigs between 16:00 and 17:00 h. When pigs become active, more occlusions and faster changes happen, which cause more identity switches and lower IDF1 values, as shown in Table 8. The average



(a)

(b)

(c)

(d)

Fig. 9. Visualization of pig trajectories with (a) ground truth of inactive pigs, (b) tracking results of inactive pigs, (c) ground truth of medium-active pigs, and (d) tracking results of medium-active pigs (pig identities are not consistent between ground truth and tracking results because they are randomly assigned).

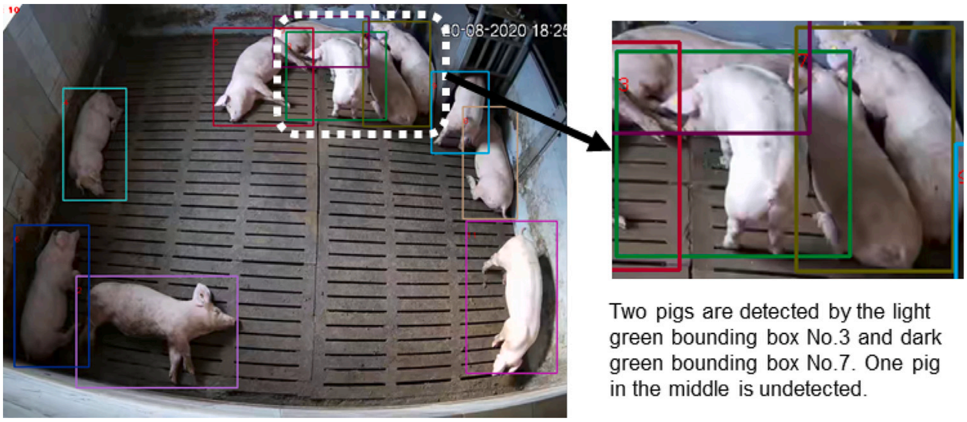
Table 8
Tracking results of 5 long-duration video segments (1 h) using the original and enhanced FairMOT method, average execution rate 48.87 fps.

Video No.	Re-ID association method	IDF1↑	MOTA↑	MOTP↑	GT	MT↑	PT	ML↓	IDs↓
v1	Original	89.90	99.80	95.10	11	11	0	0	19
	Weighted	**94.00**	99.80	95.10	11	11	0	0	19
v2	Original	74.70	89.40	94.50	11	9	2	0	54
	Weighted	**78.70**	89.40	94.50	11	9	2	0	**50**
v3	Original	92.20	100.00	96.10	11	11	0	0	6
	Weighted	92.20	100.00	96.10	11	11	0	0	6
v4	Original	52.60	99.80	95.60	11	11	0	0	107
	Weighted	52.60	99.80	95.60	11	11	0	0	**105**
v5	Original	**89.40**	99.90	94.70	11	11	0	0	18
	Weighted	88.30	99.90	94.70	11	11	0	0	**11**
Overall value (Mean ± SD)	Original	79.96 ± 16.69	97.78 ± 4.19	95.20 ± 0.59	55	53	2	0	204
	Weighted	**81.16** ± 15.23	97.78 ± 4.19	95.20 ± 0.59	55	53	2	0	**191**

one-hour tracking performance using enhanced FairMOT achieves an IDF1 of 81.16% which is 1.2% higher than the original FairMOT, as the number of ID switches of 291 is 13 less than the original FairMOT. Since individual comparisons for each video are provided in Table 8, it can be seen that the enhanced FairMOT results in higher performance, especially for IDF1 and the number of ID switches. The enhanced FairMOT method yields a similar and consistent conclusion compared with the validation on short-duration videos.

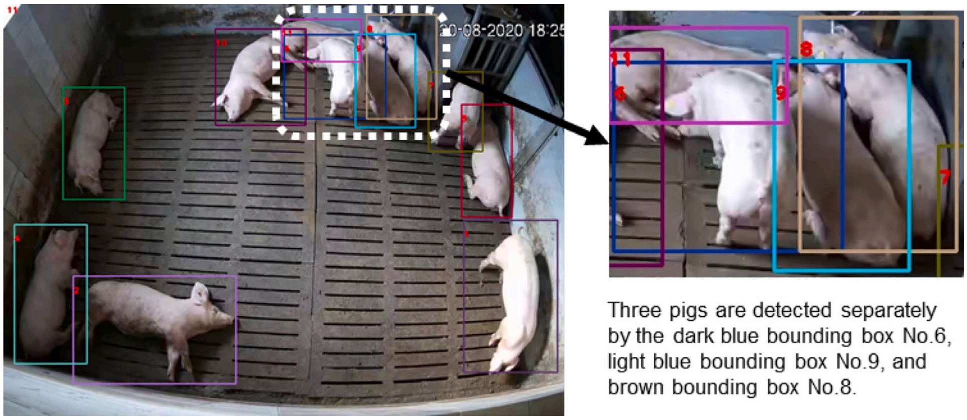
5. Discussion

In terms of the manual annotation effort, the procedure for collecting the appropriate amount of annotation as ground truth is very time-consuming. Considering the difference in moving speed between pedestrians and pigs, we have adopted an annotation interval of 2 s for pigs to improve annotation efficiency. Continuous annotation is expected to yield a more precise tracking system. To highlight the



Two pigs are detected by the light green bounding box No.3 and dark green bounding box No.7. One pig in the middle is undetected.

Fig. 10. Sample frame of tracking results from the anchor-based (JDE) method. The correct number of detections is 10 out of 11, since one anchor box (dark green box No. 7) detects two pigs. It shows the unfairness caused by the anchor-based method, which indicates that one anchor box can be assigned to multiple objects.



Three pigs are detected separately by the dark blue bounding box No.6, light blue bounding box No.9, and brown bounding box No.8.

Fig. 11. Sample frame of tracking results from the anchor-free (FairMOT) method. All 11 pigs are detected using the anchor-free method, because the extra prediction (not in the anchor-based methods) from the object center facilitates more accurate detection.

feasibility of long-duration tracking, we perform the enhanced FairMOT on 5 one-hour videos without an annotation interval. Regarding Tables 6 and 8, the average evaluation metrics of long-duration video segments are comparable. Additionally, our ultimate objective is to achieve good continuous animal tracking, so longer video recordings are required to be tested for all developed models, combined with more thorough evaluation. The obtained tracking performance is promising for processing live-streaming videos considering tracking accuracy and execution time.

Through the explanation of the anchor-based method (JDE and YOLOv5s) and anchor-free method (FairMOT) in Section 3.2.3, we have found that an anchor-free method obtains better capability for pig detection. As can be observed in Table 6, the number of identity switches achieved in enhanced FairMOT is around 300 less than enhanced JDE, and 273 less than original YOLOv5s with DeepSORT. We also demonstrate the visualized comparison between the anchor-based (see Fig. 10) and the anchor-free (see Fig. 11) methods show corresponding tracking results on the same frame. Regarding the anchor-based method, it is possible that one bounding box can be assigned to multiple objects. Therefore, JDE detects two pigs in one anchor box, as shown in the dark green bounding box of Fig. 10, which causes a missed detection and thus a tracking error. It can be observed in Fig. 10 that the detection count is 10 out of 11 because one anchor box detects two pigs. Through the anchor-free method in FairMOT, each pig is predicted from the center which avoids the unfairness caused by the anchor-free method. In this way, FairMOT detects and tracks the same pigs accurately, as shown in Fig. 11. In contrast, the anchor-free method provides a larger possibility to solve the challenge of close distance between objects during detection and tracking. Furthermore, original YOLOv5s with DeepSORT results in higher MOTA and MOTP values. Therefore, the idea to combine YOLO with the anchor-free method is expected in future work, which is also implemented in a recent publication on YOLOX (Ge et al., 2021). In future work, we will apply the detector YOLOX and different trackers such as SORT and DeepSORT, with the weighted-association algorithm proposed for further improvement.

The weighted association benefits both JDE and FairMOT tracking models, as explained in Section 4.3.1. The weighted association method applied on JDE reduces on average 40 identity switches. The original FairMOT performs the least number of identity switches among all methods, while the performance of weighted association on FairMOT is even better. The enhanced FairMOT decreases the number of identity switches with 46 counts lower than the original FairMOT. We have found that the imported appearance embedding features compensate data association compared to only relying on IOU calculation. However, the expected improvements employed on YOLOv5s with DeepSORT are not realized. One of the possible reasons is that DeepSORT has a two-step association process, which filters out the unconfirmed detections and tracklets at an early stage. In this scenario, tracking association becomes constrained, and unconfirmed detections and tracks with insufficient matching opportunities may be discarded prior to fully completing the association process. Conversely, JDE and FairMOT have proposed three-step association processes, which offer enhanced opportunities to match unconfirmed detections and tracklets. This approach has the potential to improve tracking performance and increase accuracy. Therefore, improving tracking integrity in DeepSORT is necessary. Some future work ideas include adding an additional association step prior to the final IoU matching step. The cost function utilized in the extra association could be an IoU cost or an appearance similarity cost. In the interim processing between the first and final step, a buffer pool may be established to memorize the unmatched detections and tracks following the matching cascade. A larger buffer pool size can then store more candidates and provide more potential matching tracklets.

To reduce the number of identity switches, we are working on using multiple cameras to acquire more features of objects from different perspectives. More features can enhance detection accuracy and help further data association. Apart from investigating deep learning models, we also consider to combine the video-based tracking results with contact-based sensor information such as RFID. In this way, we can explore an alternative input information channel to assist animal re-identification and thereby improve the long-duration tracking performance.

The visualization of the trajectory of pigs with the enhanced FairMOT method is shown in Fig. 9. Through this systematic visualization, it is possible to observe pig movements and habits that facilitate the combination of behavior and genetic analysis. These two examples show more inactive behavior (e.g. resting) in pigs, whereas there are also some situations, where pigs are extremely active or even fight with each other. These challenging cases introduce multiple difficulties to MOT systems.

We have collected 238,924 video frames of 96 videos with manually annotated bounding boxes for K-fold cross-validation. There are also 5 long-duration video segments for longer continuous tracking evaluation. The model generalization has been tested for different pens in different days, and various activity levels. The recordings are selected on uncontrolled lighting conditions, including morning and afternoon periods. In future work, the proposed models can be further tested using recordings from different farms.

Different from prevailing pedestrian datasets that people are coming and leaving continuously in the scene, pigs on real farms are growing in fixed pens. The camera of each pen captures all individuals. From the perspective of monitoring, a consistent identity is crucial to check the behavior of the specific pig. Therefore, the number of identity switches has an important role among all evaluation items.

As described in Section 4.2, all experiments are carried out on an RTX 2080Ti GPU. However, this leads to many limitations for parameter settings of machine learning. Apart from the sake of efficiency, the main reason we have selected the input image resolution as 864 × 480 pixels is that the GPU cannot support more expensive computations for higher image resolution during JDE implementation. Therefore, we expect better comparisons and more effective training of various A.I. algorithms by implementing modern high-performance computing solutions based on open source and scalable cloud native computing architectures in the near future.

6. Conclusion

We have investigated three state-of-the-art automated multi-object tracking methods on 2 pig datasets. Both datasets contain manual annotations of pigs in real farms. The video segments have diverse challenging conditions such as occlusion, active and high-speed movements. In this way, the generalization and robustness of the tracking models are evaluated based on K-fold cross-validation. We have proposed a weighted-association strategy to enhance the association algorithm of animal re-ID on JDE and FairMOT methods, which increase the performance of IDF1 by 1.97% at most, and reduces the mean number of identity switches by 46 at most. It can be concluded that the enhanced FairMOT performs the best in terms of multi-object tracking, indicated by an IDF1 of 80.94%, MOTA of 88.55%, MOTP of 82.60%, number of identity switches of 213. All tracking systems achieve nearly and/or real-time execution rate. For the purpose of a continuous MOT system, all proposed methods are sufficient in terms of the execution rate.

In conclusion, the experimental results of evaluation metrics demonstrate the effectiveness and robustness of the three proposed methods on multi-object tracking systems. FairMOT with the proposed weighted-association strategy achieves the best tracking performance for individual pigs in a real farm.

CRediT authorship contribution statement

Qinghua Guo: Conceptualization, Methodology, Software, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Yue Sun:** Supervision, Conceptualization, Methodology, Writing – review & editing, Project administration, Funding acquisition. **Clémence Orsini:** Data curation. **J. Elizabeth Bolhuis:** Resources, Writing – review, Project administration. **Jakob de Vlieg:** Writing – review. **Piter Bijma:** Resources, Writing – review, Project administration. **Peter H.N. de With:** Supervision, Conceptualization, Methodology, Writing – review & editing, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work is supported by the Dutch NWO project IMAGEN [P18-19 Project 1] of the research program Perspectief. The Volmer facility in Germany was offered by Topigs Norsvin in Helvoirt, the Netherlands. The authors would like to thank all researchers and student assistants from Wageningen University & Research and Eindhoven University of Technology for assistance with the tracking ground-truth annotations.

References

Bathija, A., Sharma, G., 2019. Visual object detection and tracking using yolo and sort. Int. J. Eng. Res. Technol. 8 (11).

Bergmann, P., Meinhardt, T., Leal-Taixe, L., 2019. Tracking without bells and whistles. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 941–951.

Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B., 2016. Simple online and realtime tracking. In: 2016 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 3464–3468.

Davidson, J.D., Sosna, M.M., Twomey, C.R., Sridhar, V.H., Leblanc, S.P., Couzin, I.D., 2021. Collective detection based on visual information in animal groups. J. R. Soc. Interface 18 (180), 20210142.

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2007. The PASCAL visual object classes challenge 2007 (VOC2007) results.

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2012. The PASCAL visual object classes challenge 2012 (VOC2012) results.

Fragkiadaki, K., Shi, J., 2011. Detection free tracking: Exploiting motion and topology for segmenting and tracking under entanglement. In: CVPR 2011. IEEE, pp. 2073–2080.

Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J., 2021. Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430.

Girshick, R., 2015. Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1440–1448.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 580–587.

Guo., Q., Sun., Y., Min., L., van Putten., A., Knol., E., Visser., B., Rodenburg., T., Bolhuis., J., Bijma., P., N. de With., P., 2022. Video-based detection and tracking with improved re-identification association for pigs and laying hens in farms. In: Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP,. SciTePress, ISBN: 978-989-758-555-5, pp. 69–78. http://dx.doi.org/10.5220/0010788100003124.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969.

Heindl, C., 2017. Benchmark multiple object trackers (mot) in python.

Hou, X., Wang, Y., Chau, L.-P., 2019. Vehicle tracking using deep sort with low confidence track filtering. In: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance. AVSS, IEEE, pp. 1–6.

Jocher, G., et al., 2022. Ultralytics/yolov5: v6.1 - TensorRT, TensorFlow edge TPU and OpenVINO export and inference. http://dx.doi.org/10.5281/zenodo.6222936.

Kendall, A., Gal, Y., Cipolla, R., 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7482–7491.

Kim, T.K., 2015. T test as a parametric statistic. Korean J. Anesthesiol. 68 (6), 540–546.

Kim, J., Chung, Y., Choi, Y., Sa, J., Kim, H., Chung, Y., Park, D., Kim, H., 2017. Depth-based detection of standing-pigs in moving noise environments. Sensors 17 (12), 2757.

Kuhn, H.W., 1955. The hungarian method for the assignment problem. Nav. Res. Logist. Q. 2 (1–2), 83–97.

Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K., 2015. Motchallenge 2015: Towards a benchmark for multi-target tracking. arXiv preprint arXiv:1504.01942.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: European Conference on Computer Vision. Springer, pp. 740–755.

Lloyd, S., 1982. Least squares quantization in PCM. IEEE Trans. Inform. Theory 28 (2), 129–137.

Mallick, T., Das, P.P., Majumdar, A.K., 2014. Characterizations of noise in kinect depth images: A review. IEEE Sens. J. 14 (6), 1731–1740.

Maselyne, J., Adriaens, I., Huybrechts, T., De Ketelaere, B., Millet, S., Vangeyte, J., Van Nuffel, A., Saeys, W., 2016. Measuring the drinking behaviour of individual pigs housed in group using radio frequency identification (RFID). Animal 10 (9), 1557–1566.

Matthews, S.G., Miller, A.L., PlÖtz, T., Kyriazakis, I., 2017. Automated tracking to measure behavioural changes in pigs for health and welfare monitoring. Sci. Rep. 7 (1), 1–12.

Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K., 2016. MOT16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831.

Newell, A., Huang, Z., Deng, J., 2017. Associative embedding: End-to-end learning for joint detection and grouping. Adv. Neural Inf. Process. Syst. 30.

Openvinotoolkit, O., 2020. Powerful and efficient computer vision annotation tool (CVAT). GitHub, AccessedOctober4,2020.

Perner, P., 2001. Motion tracking of animals for behavior analysis. In: International Workshop on Visual Form. Springer, pp. 779–786.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788.

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Riekert, M., Klein, A., Adrion, F., Hoffmann, C., Gallmann, E., 2020. Automatically detecting pig position and posture by 2D camera imaging and deep learning. Comput. Electron. Agric. 174, 105391.

Rodriguez, J.D., Perez, A., Lozano, J.A., 2009. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans. Pattern Anal. Mach. Intell. 32 (3), 569–575.

Siegford, J.M., Berezowski, J., Biswas, S.K., Daigle, C.L., Gebhardt-Henrich, S.G., Hernandez, C.E., Thurner, S., Toscano, M.J., 2016. Assessing activity and location of individual laying hens in large groups using modern technology. Animals 6 (2), 10.

Tan, L., Huangfu, T., Wu, L., Chen, W., 2021. Comparison of YOLO v3, faster R-CNN, and SSD for real-time pill identification.

Thombre, D., Nirmal, J., Lekha, D., 2009. Human detection and tracking using image segmentation and Kalman filter. In: 2009 International Conference on Intelligent Agent & Multi-Agent Systems. IEEE, pp. 1–5.

Tomasi, C., Kanade, T., 1991. Detection and tracking of point. Int. J. Comput. Vis. 9, 137–154.

Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., Leibe, B., 2019. Mots: Multi-object tracking and segmentation. In: Proceedings of the Ieee/Cvf Conference on Computer Vision and Pattern Recognition. pp. 7942–7951.

Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S., 2020. Towards real-time multi-object tracking. In: European Conference on Computer Vision. Springer, pp. 107–122.

Welch, G., Bishop, G., et al., 1995. An introduction to the Kalman filter.

Wojke, N., Bewley, A., Paulus, D., 2017. Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 3645–3649.

Xu, R., Lin, H., Lu, K., Cao, L., Liu, Y., 2021. A forest fire detection system based on ensemble learning. Forests 12 (2), 217.

van der Zande, L.E., Guzhva, O., Rodenburg, T.B., 2021. Individual detection and tracking of group housed pigs in their home pen using computer vision. Front. Animal Sci. 2, 669312.

Zhang, L., Gray, H., Ye, X., Collins, L., Allinson, N., 2019. Automatic individual pig detection and tracking in pig farms. Sensors 19 (5), 1188.

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X., 2022. Bytetrack: Multi-object tracking by associating every detection box. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII. Springer, pp. 1–21.

Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W., 2021. Fairmot: On the fairness of detection and re-identification in multiple object tracking. Int. J. Comput. Vis. 129 (11), 3069–3087.

Zhou, X., Wang, D., Krähenbühl, P., 2019. Objects as points. arXiv preprint arXiv:1904.07850.

Zhou, Z., Xing, J., Zhang, M., Hu, W., 2018. Online multi-target tracking with tensor-based high-order graph matching. In: 2018 24th International Conference on Pattern Recognition. ICPR, IEEE, pp. 1809–1814.

Zhou, K., Yang, Y., Cavallaro, A., Xiang, T., 2021. Learning generalisable omni-scale representations for person re-identification. IEEE Trans. Pattern Anal. Mach. Intell..