Label Efficient Lifelong Multi-View Broiler Detection

Thorsten Cardoen, Sam Leroux, Pieter Simoens
IDLab, Department of Information and Technology
Ghent University - imec, Ghent, Belgium
firstname.lastname@ugent.be

Abstract

Broiler localization is crucial for welfare monitoring, particularly in identifying issues such as wet litter. We focus on multi-camera detection systems since multiple viewpoints not only ensure comprehensive pen coverage but also reduce occlusions caused by lighting, feeder and drinking equipment. Previous multi-view detection studies localize subjects either by aggregating ground plane projections of single-view predictions or by developing end-to-end multi-view detectors capable of directly generating predictions. However, single-view detections may suffer from reduced accuracy due to occlusions, and obtaining ground plane labels for training end-to-end multi-view detectors is challenging. In this paper, we combine the strengths of both approaches by using the readily available aggregated single-view detections as labels for training a multi-view detector. Our approach alleviates the need for hard-to-acquire ground-plane labels. Through experiments on a real-world broiler dataset, we demonstrate the effectiveness of our approach.

1. Introduction

In modern agricultural practices, ensuring and maintaining the health and well-being of broilers has become an increasingly vital concern [29]. Automatic video-based systems have garnered significant attention as a non-intrusive, low-cost, and effective approach for monitoring welfare. Central to such video monitoring systems is the accurate position detection of broilers confined within pens of varying sizes and layouts. This capability can assist farmers to identify and address issues such as wet litter and temperature differentials, while also facilitating subsequent tasks such as tracking and action recognition, thereby enhancing the overall broiler welfare and farm efficiency. Conventional broiler detection techniques [21, 22, 31, 40] use a single camera for single-view detection. However, these techniques often face significant challenges in real-world settings, particularly when dealing with occlusion caused by the dense clustering of broilers as well as feeders, drinkers and other equipment [14, 30].

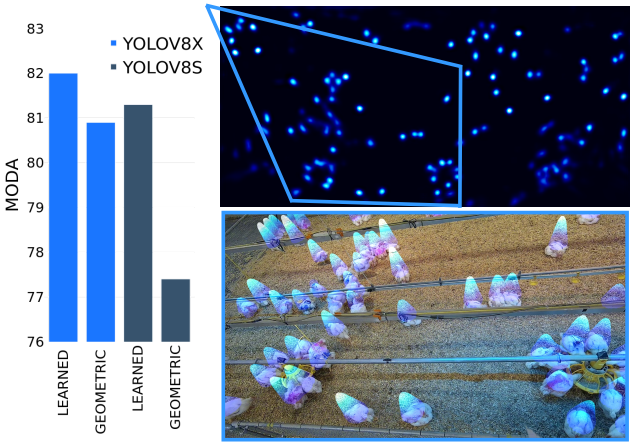

Figure 1. Training an end-to-end multi-view detection model using pseudo labels can improve the performance. Additionally, when the object detector used to generate the labels is smaller/noisier, the approach can be more beneficial. We show the performance improvement in MODA for both the larger YOLOV8X and smaller YOLOV8S object detection models. The top-right image depicts an example of the probability occupancy map (POM) obtained by fusing information across four cameras. The bottom-right image is an example of an input image on which the POM is back-projected. The boundaries of the projected input image are also shown by the blue quadrilateral in the POM.

Recent advancements in multi-view object detection have offered promising solutions to mitigate the challenges posed by occlusion and address the problem of overlap between cameras by producing a unified ground plane detection map. The majority of multi-view detection research has focused on pedestrian detection due to the availability of pedestrian detection benchmarks [10, 17, 39]. These multi-view detection techniques can be categorized into two groups, two-stage and end-to-end detectors. The more advanced end-to-end methods take multiple viewpoints as input and produce a ground plane occupancy map by taking

cues from all inputs at the same time [16, 17, 39]. However, it is worth noting that broiler detection poses unique challenges compared to pedestrian detection, firstly due to the higher density of subjects within the pen and secondly due to the high similarity in appearance of broilers compared to pedestrians [44]. Additionally, a major bottleneck for practical deployment lies in the acquisition of ground plane labels. Unlike bounding box annotations, which can be readily obtained through various annotation tools [8, 37] and object detection pipelines, ground plane position labels remain elusive due to the lack of robust annotation tools.

By using a two-stage approach that first detects objects within each view, and subsequently fuses the projected detections on the ground plane, existing object detection pipelines can be harnessed. These approaches, however, rely on various techniques [2, 11, 34] to fuse per-view results on the ground plane which are prone to calibration errors. In this paper, we utilize the outputs (pseudo labels) of a two-stage multi-view broiler detection pipeline with a geometric fusion technique [7], which we refer to as the geometric-fusion model, to train the Multi-View Detection (MVDet) [17] architecture end-to-end, which we refer to as the learned-fusion model. Without requiring additional ground plane labels for training, we demonstrate significant performance improvements over the geometric-fusion model. This enhancement can be attributed to two key factors: firstly, the model's ability to disregard wrongly detected instances from individual views, and secondly, its capacity to learn and correct calibration inconsistencies inherent in multi-view setups. Due to the changing appearance of the chickens throughout their various growth stages, we introduce age-specific Gaussian filters, which allow the model to learn different features for different ages. We present compelling evidence of the efficacy of our proposed approach, achieving improvements with a MODA of 81.3% compared to 77.4% obtained with the geometric-fusion model, see Figure 1. We also compare the model across different age groups and investigate the effect of degrading pseudo labels. In summary, we make the following contributions:

- We show the performance improvements over the geometric-fusion model and the effect of bounding box quality by using various YOLOV8 object detector sizes.
- We present age-specific kernels that significantly improve the performance by allowing the model to learn different features for different growth stages.
- We detail several modifications required to apply the popular MVDet [17] architecture effectively for broiler detection.

2. Related work

Single-view Detection algorithms for broiler welfare monitoring systems were initially designed with traditional image processing techniques. These pipelines involved converting images to grayscale, applying blurring filters, using the Otsu thresholding algorithm for segmentation, and enhancing broiler regions through dilation operations. Earlier methods often employed a grid-based or density-based approach to estimate occupation scores for different zones within a pen [21, 31]. However, such approaches lacked detailed information, prompting a shift from group-level to individual-level analysis [22], where the exact positions of individual broilers are determined [13, 29]. While these traditional computer vision methods showed efficacy in controlled settings, they faced limitations in adapting to various environmental conditions and struggled with complex backgrounds and lighting variations. In contrast, recent advancements in machine learning have led to more robust and flexible broiler detection systems [6, 23, 38, 42]. Notably, there's a shift towards adopting deep learning object detection models for broiler detection [40, 43] such as SSD [27], Faster R-CNN [33] and YOLO [19, 32], which can locate and classify broilers within images directly. This transition from handcrafted features to learned representations has shown promising results in broiler detection tasks. Although recent work [40] aims to reduce the problem of mutual occlusion by using super-resolution reconstruction, other occlusions, such as feeder and drinking equipment, remain a problem.

Multi-view detection has been an active area of research due to its ability to handle occlusions and improve detection accuracy in crowded scenes. There is a scarcity of research directly addressing the specific problem of multi-view animal detection [7]. Therefore, we instead present relevant techniques from the field of multi-view pedestrian detection and discuss their applicability to broiler detection.

Two-Stage Approaches: Early approaches relied on per-view background subtraction to compute likelihoods over a discrete grid on the ground plane. Conditional random field or mean-field inference is then employed to capture spatial relationships [1, 3, 11]. However, these methods struggle with increased crowd density as background subtraction becomes less effective. As discussed before, recent machine learning advances can address these limitations of background subtraction by training object detection models [19, 27, 32, 33] on large labelled datasets. Using these models, detections can be generated in each view and subsequently projected to the common ground plane [2, 9, 41]. The accurate aggregation of these projections heavily relies on the accuracy of the calibration, which can have inconsistencies and limit the performance of these approaches.

End-to-end Approaches: Recently, multi-view pedestrian detection models produce Probability Occupancy Maps (POMs) directly by training on top-down annotations. MVDet [17] is one such example, where multi-view features are projected onto a ground plane and processed

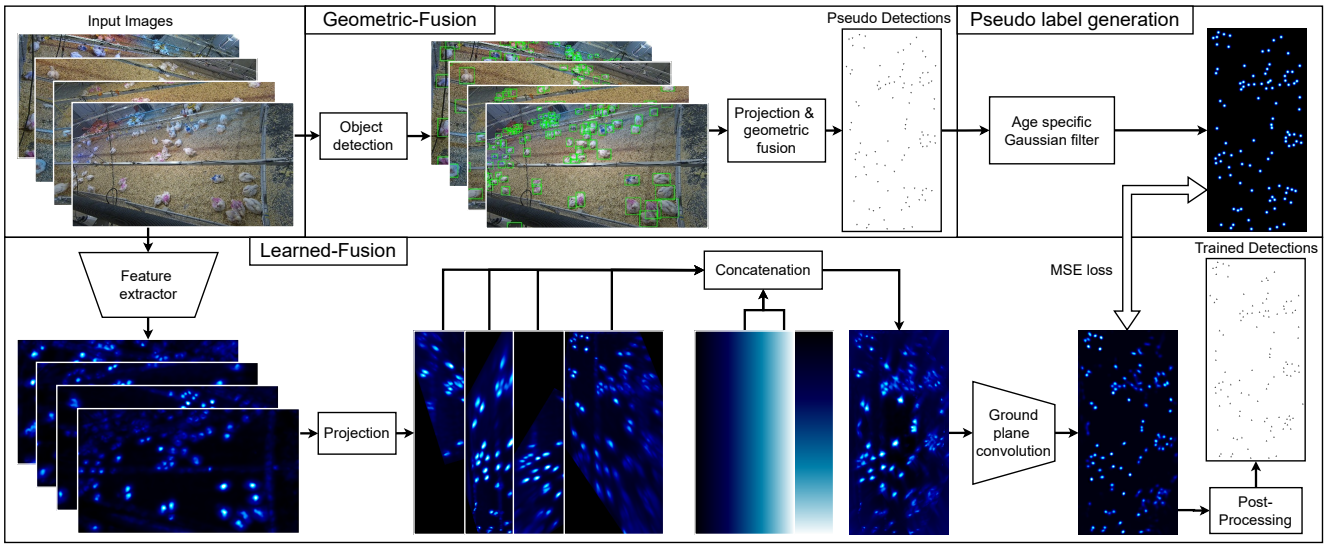

Figure 2. Architecture overview of the broiler detection framework. The top row illustrates the geometric-fusion pipeline for obtaining pseudo detections and the subsequent label generation utilized for model training, while the bottom row depicts the learned-fusion model. The geometric-fusion model initially detects bounding boxes in each image using an object detector, followed by projection onto the ground plane and fusion using a geometric approach. The learned-fusion model (consisting of a feature extractor and ground plane convolutions) can subsequently be trained using these pseudo-detections. Feature maps are extracted and projected onto a common ground plane. Next, they are concatenated together with x and y coordinate maps. The ground plane convolutions then use large kernel sizes for joint occupancy decisions, integrating spatial neighbour information. Finally, we use NMS to post-process the occupancy map.

through convolutional layers to predict a final occupancy map. This work highlights the benefit of fusing projected image features opposed to single-view results or the input images (raw RGB values). Due to the low-camera angles and relative height of pedestrians, shadows complicate the fusion process. To address these issues MVDeTr [16] builds upon MVDet by incorporating deformable transformers for improved feature aggregation. Similarly, SHOT employs multiple homographies at various heights to enhance projection quality [18, 36]. Generalized Multi-View Detection (GMVD) [39] leverages the strengths of existing multi-view pedestrian detection methods while addressing the limitations of generalizability to real-world deployments. It proposes a novel dataset specifically designed to evaluate generalization capabilities in multi-view detection tasks and addresses the three key forms of generalization: varying number of cameras, varying camera positions, and generalizability to new scenes. These methods, however, require large annotated datasets, which prove hard to obtain. Recent work [24] leverages the GMVD approach to reduce the labelling effort by training the model using different source and target labels. However, adapting this method to broiler detection is non-trivial due to the changing subject size. Efforts have also been made to remove the labelling effort altogether and go for a fully unsupervised approach [25] but at a reduced performance.

3. Approach

In this section, we outline our approach, which is divided into three main components: pseudo-label generation, multi-view architecture, and post-processing. Firstly, we discuss our pseudo-label generation, where we utilize a two-stage process for broiler detection. Next, we detail our multi-view architecture, which combines information from multiple views while handling occlusions using anchor-free multi-view aggregation and feature perspective transformation techniques. Lastly, we describe our post-processing methods, including thresholding the output and applying Non-Max Suppression (NMS) to reduce redundant predictions. In Figure 2, we show the complete framework.

3.1. Pseudo label generation

For pseudo-labelling the input data, we utilize the multi-view broiler detection framework presented in [7]. This work uses a two-stage approach whereby the broilers are first detected within each viewpoint using the YOLOV8 object detector [19]. We experiment with YOLOV8 object detection models of different sizes to investigate the effect of the bounding-box quality:

- X: 68.2 million parameters
- L: 43.7 million parameters
- M: 25.9 million parameters
- S: 11.2 million parameters

The second stage begins by approximating the position

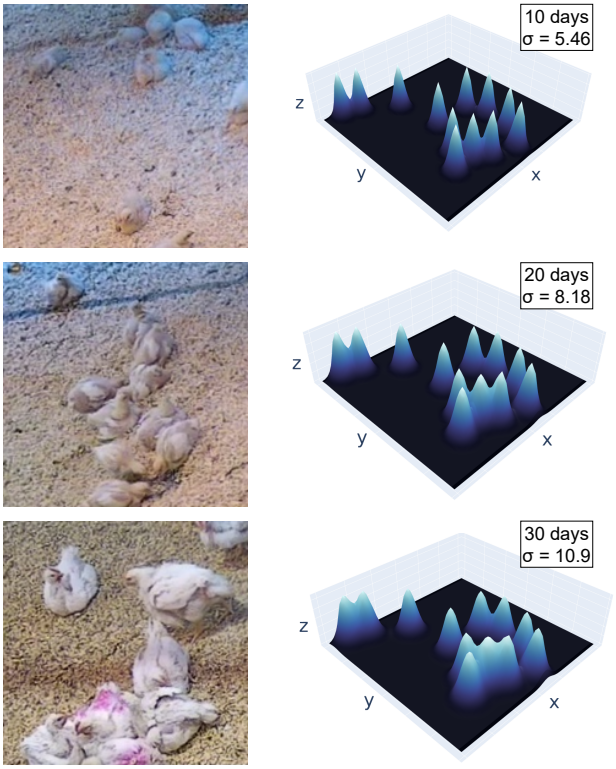

Figure 3. The left shows the crops of input images at various ages, and the right column shows the corresponding kernel applied to a map of broiler detections. Using adaptive kernel size based on the age of the broiler can allow for learning age-specific features.

between the broiler's feet within each bounding box and projecting them onto the ground plane. In order to consolidate broiler detections from various viewpoints, a graph representation is utilized [28]. Detections are linked into a connected component if their Euclidean distance falls below a predetermined fusion radius and originate from different cameras. The arithmetic mean of the projected points from different cameras is computed to determine the location of each broiler. We refer to this procedure as the geometric-fusion model.

A Gaussian kernel is then applied to the detections using a convolution operation to produce the ground truth target for training the learned-fusion model. Let $\mathbf{x} := [k_x, k_y]^T$ be the position in the kernel, σ^2 being the variance of the multivariate Gaussian distribution, we then generate the corresponding kernel using:

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}\right), \quad \Sigma = \sigma^2\mathbf{I}.$$

with $\mathbf{I}$ being the identity matrix, following [17].

To address the varying characteristics of broilers at different ages, we introduce Gaussian kernels with different standard deviation values (σ). This allows us to generate kernels tailored to specific ages, enabling the model to learn distinct features corresponding to each age category. Figure 3 illustrates the impact of employing different sigma values on kernel generation. Notably, by employing varying kernel sizes, we try to approximate the effect of various bounding box sizes inherent in single-view labels, which enables object detectors to learn diverse features for subjects of different sizes.

3.2. Multi-view architecture

A core challenge in multi-view detection is effectively combining information from multiple views while accounting for potential occlusions. We adapt the MVDet architecture that addresses this challenge through anchor-free [17, 20] multiview aggregation and feature perspective transformation.

Using a feature extractor network, image features are computed for each of the synchronized input images. MVDet uses the Resnet18 [15] model and replaces the last 3 strided convolutions with dilated convolutions to maintain a high spatial resolution. We found, however, that using an input dimension of (720, 1280) resulting in image features with spatial dimensions (90, 160) is too low, especially for smaller/younger chickens [7, 40]. To this end, we remove the max-pooling layer before the first residual block, enlarging the spatial dimension of the image features to (180,320). These features are then finally up-sampled to a fixed feature map height and width. Like MVDet, we use an image size reduction of 4 (1920/4, 1080/4) for the feature map height and width. If the full input size were to be used, this up-sampling step could be omitted, however, due to memory limitations, this is impractical.

Using each camera's intrinsic and extrinsic parameters, these image features can be projected to the ground plane. All projected maps are then concatenated together with X and Y coordinate maps, giving the subsequent convolution access to its own input coordinates [26]. The last step involves aggregating the multi-view information. Incorporating information from a larger surrounding area is particularly important for broiler detection in dense environments with multiple occlusions. MVDet addresses this requirement through spatial aggregation using large kernel convolutions. The large receptive field of these convolutions allows the network to consider the context of neighbouring regions, make more informed decisions about broiler occupancy, and overcome small projection errors stemming from imperfect calibration parameters. The model can then be trained end-to-end using Mean Squared Error (MSE).

Being originally developed for pedestrian detection, MVDet uses a secondary loss for head and foot detection,

which are approximated by the centers of the top and bottom lines of the bounding boxes, respectively. The authors conclude, however, that only the head point detection slightly benefits the model's performance due to the heterogeneous supervision compared to the foot detection. We omit this extra learning objective for two reasons: due to the larger camera angle with respect to the ground plane for the broiler camera setup, the head and foot approximation used for pedestrians does not correspond well with the head and feet of the broilers. The second reason is that we do not possess ground truth bounding boxes but only predictions by single-view detectors.

3.3. Post processing

After obtaining the final output map of the model, post-processing is still required to obtain the final detections. After thresholding, the output, NMS [5], is applied to filter out redundant detections. It works by comparing the distances between the detected positions and removes those that are close to each other, keeping only the most confident one. This prevents multiple detections of the same object and helps in producing a more accurate and concise output. This, however, requires a hyperparameter that defines the distance with which predictions are removed. Pedestrian detection techniques employ the same NMS distance threshold as is used to calculate the metrics, which is 0.5 meters [10, 16, 17, 39]. However, we empirically find that using the same threshold for NMS leads to suboptimal results. We therefore use a linearly increasing threshold and regard the coefficients of the linear function as two hyperparameters.

4. Experimental setup

This section introduces the experimental setup, covering three main aspects: dataset, metrics, and implementation details. We briefly outline the dataset used in this study, followed by an overview of the evaluation metrics employed. Lastly, we provide insight into the implementation specifics.

4.1. Dataset

We use the dataset detailed in [7], which aims to automate welfare assessment using multi-modal sensor data and explores the impact of pen infrastructure conditions on chicken welfare. It comprises around 800 hours of video footage from four cameras over the six-week lifespan. The pen accommodates an average of 140 chickens and is divided logically into zones for various activities. Because of the various lighting conditions, different ages and occlusions, see Figure 5, this dataset is very challenging, We indicate the drinker and feeding equipment on the ground plane using black circles in Figure 4 and display the number of views that can view each location. Ground truth data on the position of individual chickens within the pens was manually annotated, totalling 51 sets of four frames, to ensure a comprehensive evaluation of the pipeline's performance. The dataset consists of 4,739 labelled broilers for validation and 1,967 for testing, enabling thorough analysis of multi-view detection algorithms. The training data is generated by exporting 8000 frames for each camera, sampled every 10 seconds. These sets are split up into three age groups depending on the broiler's dietary stage.

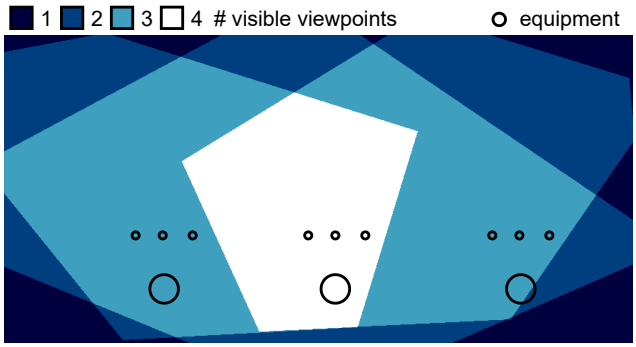

Figure 4. Overlap between camera views shown on the ground plane, the colors indicate the number of viewpoints from which the location is visible. The circles indicate the 9 drinkers and 3 feeders.

- Starter: $<$10 days
- Grower: 10 to 23 days
- Finisher: $>$23 days

4.2. Metrics

To evaluate the performance of the models, we report the commonly used Multi-Object Detection Accuracy (MODA), Multi-Object Detection Precision (MODP), precision and recall following [12]. We use MODA as the primary metric as it takes both false negatives and false positives into account. The threshold distance for determining true positives should reflect the physical size of the animals. For pedestrian detection, a single threshold value of 0.5 m is used as the approximated average width of a human body [10]. During the 6-week lifespan in the pen, the broilers keep growing, increasing the average width. We determine the threshold as a linear function of age. The coefficient values were empirically determined based on real-life measurements of broilers at different ages.

$$\text{Distance Threshold} = 0.3\text{cm} * age + 5\text{cm}$$

4.3. Implementation details

Similar to MVDet, we downsize the images to 720x1280 but remove the first max-pooling layer to keep a higher spatial resolution and facilitate the detection of the small broiler. We train the model for 2000 iterations per epoch

		MODA		MODP		Precision		Recall		Δ MODA
	YOLOV8	Learned	Geometric	Learned	Geometric	Learned	Geometric	Learned	Geometric	
Fixed σ	X	81.4	80.9	64.3	64.2	90.6	90.0	90.9	91.1	+ 0.5
	L	79.9	80.6	63.3	64.1	89.1	90.3	91.3	90.4	- 0.7
	M	79.4	77.3	63.2	63.8	91.7	90.2	87.3	86.8	+ 2.1
	S	79.0	77.4	63.1	63.9	90.1	88.9	89.0	88.7	+ 1.6
Age-specific σ	X	**82.0**	80.9	**65.9**	64.2	**91.4**	90.0	90.6	**91.1**	+ 1.1
	L	80.3	**80.6**	**66.2**	64.1	87.9	**90.3**	**93.1**	90.4	- 0.3
	M	**80.1**	77.3	**65.7**	63.8	90.0	**90.2**	**90.3**	86.8	+ 2.8
	S	**81.3**	77.4	**64.8**	63.9	**91.0**	88.9	**90.4**	88.7	+ 3.9

Table 1. Comparison between various YOLOV8 object detector sizes and the effect of adding age-specific kernels. The biggest improvement can be seen in the model trained on the labels generated by the geometric-fusion model using the YOLOV8S object detector. The overall best-performing model is trained using the YOLOV8X pseudo labels, achieving 82% MODA.

	Starter		Grower		Finisher		Total	
	Learned	Geometric	Learned	Geometric	Learned	Geometric	Learned	Geometric
MODA	78.8	76.8	83.3	83.2	83.9	82.7	82.0	80.9
MODP	63.3	63.2	66.5	64.3	67.9	65.1	65.9	64.2
Precision	91.4	86.3	91.7	91.2	91.1	92.5	91.4	90.0
Recall	87.1	91.4	91.7	92.1	93.0	90.0	90.6	91.1

Table 2. Comparison between the geometric-fusion model using the YOLOV8X detector and the learned-fusion model trained with age-specific kernels using these labels. We show the performance for various age groups. We obtain the highest MODA for the broilers at the finisher stage.

with a batch size of three for a total of 10 epochs. Training these models took about 32 hours using two V100 GPUs. We use a gridsize of 430x880 with per frame average detections of 122.9 broilers, where each cell represents a one-by-one cm square. Due to computational restrictions, we reduce the gridsize by a factor of two, further reductions in gridsize empirically show a degradation in performance because of the high broiler density. We follow the MVDet paper [17] for the remainder of hyperparameters, such as the choice of optimizer, momentum, L2-normalisation, learning rate scheduler and maximum learning rate.

5. Experiments

This section presents quantitative and qualitative results. First, we present performance differences between object detector sizes and the impact of age-specific kernels. Then we compare the learned-fusion model with the geometric-fusion model using the YOLOV8S object detector in more detail. Lastly, we showcase several advantages and limitations of using pseudo labels for training.

5.1. Quantitive results

For each model, we tune the post-processing hyperparameters on the validation set and report the results achieved on the test set. We implement early stopping to alleviate overfitting on the pseudo labels.

From table 1, it is evident that training on the pseudo labels generally leads to performance improvements across different model sizes. The magnitude of improvement varies depending on the model size and the presence of additional age-specific kernels. For instance, models of size M and S exhibit notable improvements, especially when age-specific kernels are added. Conversely, the impact on larger models like X is less pronounced and even experiences slight performance degradation with the M model, although achieving a considerably higher recall.

We also compare the learned-fusion and geometric-fusion models using the YOLOV8X object detector for the different dietary stages in table 2. The error of the mid-feet position approximation within the bounding box has a bigger effect on the geometric-fusion as the broilers and their bounding boxes grow in size, leading to better fusion for the younger broilers. On the other hand, larger broilers are much easier to detect. As a result of these two factors, for the geometric-fusion model, the best MODA of 83.2% is achieved for the grower phase. In contrast, for the learned-fusion model, the best MODA of 83.9% is achieved for the finisher phase. This highlights the improved fusion performance over geometric-fusion specifically for the larger broilers.

Due to the downsampling of the input images for the learned-fusion model compared to the object detector input (width of 1280 vs 1920), the model struggles to ac-

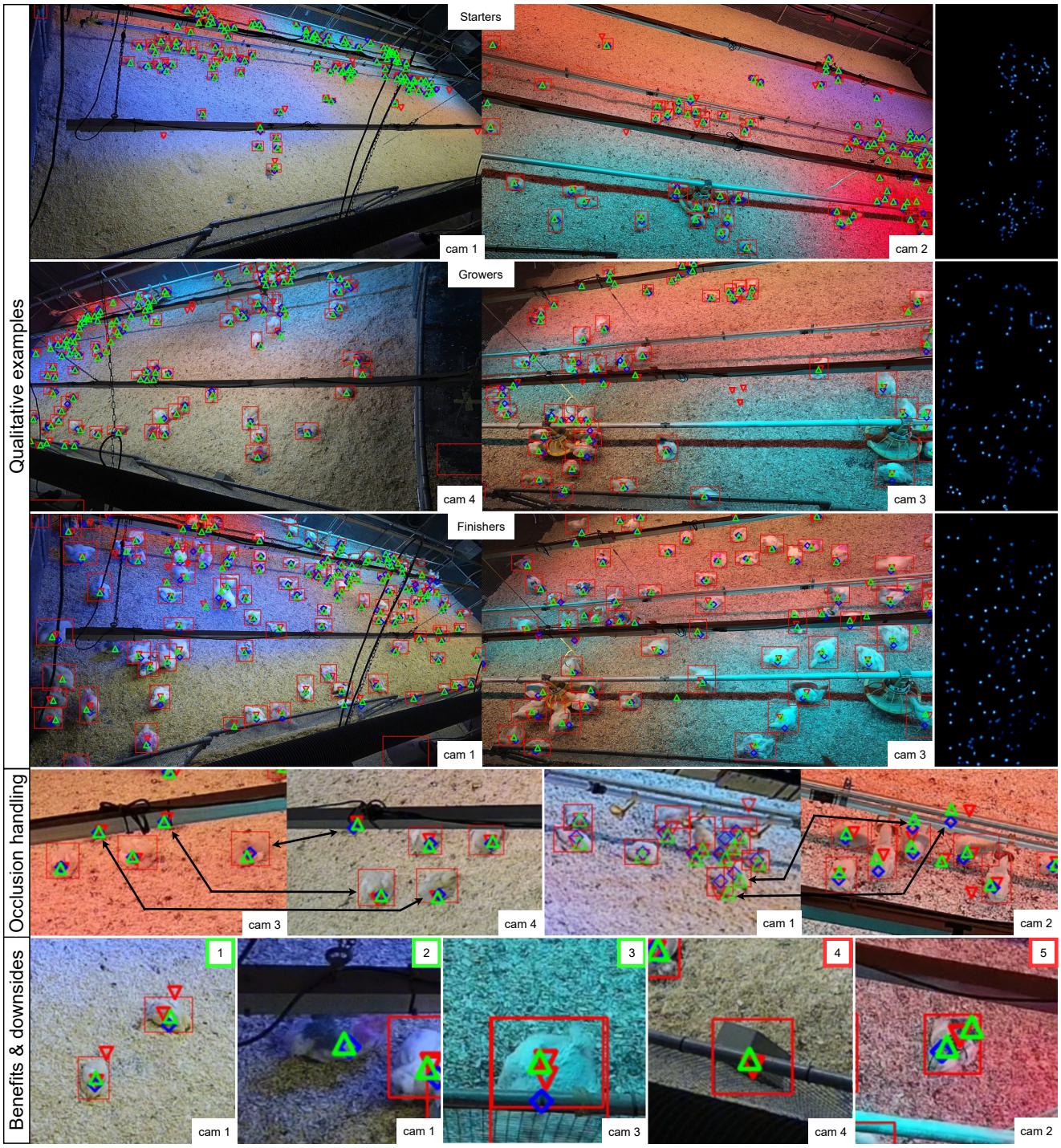

Figure 5. This figure indicates the ground truth labels with ◇ , the pseudo labels (from geometric-fusion model) using ▽ and the model's predictions using △. The bounding boxes □ are the single view results produced by the YOLO model. In the bottom-right corner, we detail which camera produced the image. The first three rows each show two viewpoints for the different age groups. We add the POM produced by the model to the right of each sample. The second to last row demonstrates the model's ability to overcome occlusions caused by the equipment using two examples. In the last row, we show five highlights to indicate the benefits and downsides of using the proposed approach. Note that not all of the cropped images in the last two rows come from the three displayed samples. Figure is best viewed in color.

curately detect the smaller broilers during the starter and grower phases. This leads to a significant decrease in recall and highlights the downside of downsampling [7, 40].

5.2. Qualitative results

In Figure 5, we show one test sample for each age category using the YOLOV8S object detector. During the starter stage, the broilers group together more, which can be seen in the POMs on the right of the input images, increasing the complexity. In the fourth row, we showcase the multi-camera setup's capability to handle occlusions. In the example on the left, two broilers are occluded by lighting equipment in camera 3 but remain visible in camera 4. Conversely, in camera 4, one broiler is occluded, while it remains visible in camera 3. In the last row, we show several upsides as well as failure cases of using pseudo labels for training an end-to-end model. In highlight one, we show that multiple detections due to the incorrect fusion of the per-view projected bounding boxes can be ameliorated. The second highlight indicates that the learned fusion model can improve upon the features learned by the YOLO object detector to detect broilers. Finally, the learned-fusion model can avoid double detections by taking features from every view simultaneously, as seen in the third highlight. However, this approach also has downsides, one of which is that it can pick up features of consistently mispredicted objects by the object detector, an example is shown in highlight 4. Another limitation is that due to the need for post-processing the POM using NMS, which relies on a single distance threshold for the whole ground plane map, multiple detections for a single chicken can occur, see highlight 5.

Due to this limitation, we also experimented with several more advanced post-processing techniques such as Soft-NMS [4] and DB-NMS [35]. However, applying these more sophisticated techniques comes with added difficulties when applying them on POMs. For bounding boxes, a matched ground truth object is determined using the intersection over union (IoU), while euclidean distance is used on the POM. We leave the application of more advanced NMS techniques as future work.

6. Conclusion

In conclusion, we have presented several key changes to improve lifelong multi-view broiler detection, addressing the crucial task of ensuring the welfare of chickens in the poultry industry. By leveraging recent advances in multi-view detection and combining them with easier-to-obtain bounding box annotations, we have demonstrated significant improvements in performance. Our proposed method, which integrates pseudo labelling with the MVDet architecture and employs age-specific kernels, achieves notable enhancements in broiler detection accuracy. Through comprehensive experiments and evaluations, we have showcased several reasons for the improvements of using an end-to-end trained model. Additionally, we highlight settings where our approach can be most beneficial. These advancements hold promise for enhancing automated broiler welfare monitoring systems, ultimately contributing to better practices and outcomes in poultry farming.

7. Acknowledgments

The imec.icon project WISH is a research project bringing together academic researchers and industry partners. Project WISH is co-financed by imec and receives financial support from Flanders Innovation & Entrepreneurship (project nr. HBC.2021.0664). We gratefully acknowledge the support of ILVO Vlaanderen and Ghent University Faculty of Veterinary Medicine - Poultry Health Sciences, who organized the experimental setup, took care of the broilers, and provided us with domain knowledge. Sam Leroux and Pieter Simoens acknowledge the financial support of the Flanders AI Research (FAIR) program.

References

[1] Alexandre Alahi, Laurent Jacques, Yannick Boursier, and Pierre Vandergheynst. Sparsity driven people localization with a heterogeneous network of cameras. *J. Math. Imaging Vis.*, 41(1-2):39–58, 2011. 2

[2] Pierre Baqué, François Fleuret, and Pascal Fua. Deep occlusion reasoning for multi-camera multi-target detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 271–279. IEEE Computer Society, 2017. 2

[3] Jérôme Berclaz, François Fleuret, Engin Türetken, and Pascal Fua. Multiple object tracking using k-shortest paths optimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(9): 1806–1819, 2011. 2

[4] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-nms - improving object detection with one line of code. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5562–5570. IEEE Computer Society, 2017. 8

[5] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. 5

[6] Liangben Cao, Zihan Xiao, Xianghui Liao, Yuanzhou Yao, Kangjie Wu, Jiong Mu, Jun Li, and Haibo Pu. Automated chicken counting in surveillance camera environments based on the point supervision algorithm: Lc-densefcn. *Agriculture*, 11(6), 2021. 2

[7] Thorsten Cardoen, Sam Leroux, and Pieter Simoens. Multi-camera detection framework for lifelong broiler flock monitoring. *Available at SSRN 4685972*. 2, 3, 4, 5, 8

[8] J. Cartucho, R. Ventura, and M. Veloso. Robust object recognition through symbiotic deep learning in mobile robots. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2336–2341, 2018. 2

[9] Tatjana Chavdarova and François Fleuret. Deep multi-camera people detection. In *16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017, Cancun, Mexico, December 18-21, 2017*, pages 848–853. IEEE, 2017. 2

[10] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur M. Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. WILDTRACK: A multi-camera HD dataset for dense unscripted pedestrian detection. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5030–5039. Computer Vision Foundation / IEEE Computer Society, 2018. 1, 5

[11] François Fleuret, Jérôme Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30 (2):267–282, 2008. 2

[12] John S. Garofolo, Rachel Bowers, Dennis E. Moellman, Rangachar Kasturi, Dmitry Goldgof, and Padmanabhan Soundararajan. Performance evaluation protocol for face, person and vehicle detection & tracking in video analysis and content extraction (vace-ii) clear - classification of events, activities and relationships. 2006. 5

[13] Yangyang Guo, Lilong Chai, Samuel E. Aggrey, Adelumola Oladeinde, Jasmine Johnson, and Gregory Zock. A machine vision-based method for monitoring broiler chicken floor distribution. *Sensors*, 20(11):3179, 2020. 2

[14] Yangyang Guo, Samuel E. Aggrey, Adelumola Oladeinde, Jasmine Johnson, Gregory Zock, and Lilong Chai. A machine vision-based method optimized for restoring broiler chicken images occluded by feeding and drinking equipment. *Animals*, 11(1), 2021. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[16] Yunzhong Hou and Liang Zheng. Multiview detection with shadow transformer (and view-coherent data augmentation). In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 1673–1682. ACM, 2021. 2, 3, 5

[17] Yunzhong Hou, Liang Zheng, and Stephen Gould. Multiview detection with feature perspective transformation. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VII*, pages 1–18. Springer, 2020. 1, 2, 4, 5, 6

[18] Jinwoo Hwang, Philipp Benz, and Taehoon Kim. Boostershot: Boosting stacked homography transformations for multiview pedestrian detection with attention. *CoRR*, abs/2208.09211, 2022. 3

[19] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, TaoXie, Jiacong Fang, imyhxy, Lorna, (Zeng Yifu), Colin Wong, Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, Victor Sonck, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Dhruv Nair, Max Strobel, and Mrinal Jain. ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, 2022. 2, 3

[20] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyound anchor-based object detection. *IEEE Trans. Image Process.*, 29:7389–7398, 2020. 4

[21] Guoming Li, Yang Zhao, Joseph L. Purswell, Qian Du, Gray D. Chesser, and John W. Lowe. Analysis of feeding and drinking behaviors of group-reared broilers via image processing. *Computers and Electronics in Agriculture*, 175: 105596, 2020. 1, 2

[22] N. Li, Z. Ren, D. Li, and L. Zeng. Review: Automated techniques for monitoring the behaviour and welfare of broilers and laying hens: towards the goal of precision livestock farming. *Animal*, 14(3):617–625, 2020. 1, 2

[23] Ximing Li, Zeyong Zhao, Jingyi Wu, Yongding Huang, Jiayong Wen, Shikai Sun, Huanlong Xie, Jian Sun, and Yuefang Gao. Y-BGD: broiler counting based on multi-object tracking. *Comput. Electron. Agric.*, 202:107347, 2022. 2

[24] João Paulo Lima, Diego Thomas, Hideaki Uchiyama, and Veronica Teichrieb. Toward unlabeled multi-view 3d pedestrian detection by generalizable AI: techniques and performance analysis. *CoRR*, abs/2308.04515, 2023. 3

[25] Mengyin Liu, Chao Zhu, Shiqi Ren, and Xu-Cheng Yin. Unsupervised multi-view pedestrian detection. *CoRR*, abs/2305.12457, 2023. 3

[26] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9628–9639, 2018. 4

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, pages 21–37. Springer, 2016. 2

[28] Alejandro López-Cifuentes, Marcos Escudero-Viñolo, Jesús Bescós, and Pablo Carballeira. Semantic-driven multi-camera pedestrian detection. *Knowl. Inf. Syst.*, 64(5):1211–1237, 2022. 4

[29] Renan Vilas Novas and Fábio Luiz Usberti. Live monitoring in poultry houses: A broiler detection approach. In *30th SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI 2017, Niterói, Brazil, October 17-20, 2017*, pages 216–222. IEEE Computer Society, 2017. 1, 2

[30] Cedric Okinda, Innocent Nyalala, Tchalla Korohou, Celestine Okinda, Jintao Wang, Tracy Achieng, Patrick Wamalwa, Tai Mang, and Mingxia Shen. A review on computer vision systems in monitoring of poultry: A welfare perspective. *Artificial Intelligence in Agriculture*, 4, 2020. 1

[31] Alberto Peña Fernández, Tomas Norton, Emanuela Tullo, Tom van Hertem, Ali Youssef, Vasileios Exadaktylos, Erik Vranken, Marcella Guarino, and Daniel Berckmans. Real-time monitoring of broiler flock's welfare status using camera-based technology. *Biosystems Engineering*, 173: 103–114, 2018. Advances in the Engineering of Sensor-

based Monitoring and Management Systems for Precision Livestock Farming. 1, 2

[32] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 779–788. IEEE Computer Society, 2016. 2

[33] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017. 2

[34] Gemma Roig, Xavier Boix, Horesh Ben Shitrit, and Pascal Fua. Conditional random fields for multi-camera object detection. In *2011 International Conference on Computer Vision*, pages 563–570, 2011. 2

[35] Li Rui, Xue-Song Tang, and Kuangrong Hao. DB-NMS: improving non-maximum suppression with density-based clustering. *Neural Comput. Appl.*, 34(6):4747–4757, 2022. 8

[36] Liangchen Song, Jialian Wu, Ming Yang, Qian Zhang, Yuan Li, and Junsong Yuan. Stacked homography transformations for multi-view pedestrian detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6029–6037. IEEE, 2021. 3

[37] Tzutalin. Labelimg https://github.com/tzutalin/labelimg, 2015. 2

[38] Jerine A.J. Van der Eijk, Oleksiy Guzhva, Alexander Voss, Matthias Möller, Mona F. Giersberg, Leonie Jacobs, and Ingrid C. De jong. Seeing is caring – automated assessment of resource use of broilers with computer vision techniques. *Frontiers in Animal Science*, 3, 2022. 2

[39] Jeet Vora, Swetanjal Dutta, Kanishk Jain, Shyamgopal Karthik, and Vineet Gandhi. Bringing generalization to deep multi-view pedestrian detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV 2023 - Workshops, Waikoloa, HI, USA, January 3-7, 2023*, pages 110–119. IEEE, 2023. 1, 2, 3, 5

[40] Zhenlong Wu, Tiemin Zhang, Cheng Fang, Jikang Yang, Chuang Ma, Haikun Zheng, and Hongzhi Zhao. Super-resolution fusion optimization for poultry detection: A multi-object chicken detection method. *Journal of Animal Science*, 101:skad249, 2023. 1, 2, 4, 8

[41] Yuanlu Xu, Xiaobai Liu, Yang Liu, and Song-Chun Zhu. Multi-view people tracking via hierarchical trajectory composition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4256–4265. IEEE Computer Society, 2016. 2

[42] Xiao Yang, Lilong Chai, Ramesh Bahadur Bist, Sachin Subedi, and Zihao Wu. A deep learning model for detecting cage-free hens on the litter floor. *Animals*, 12(15), 2022. 2

[43] Syed Sahil Abbas Zaidi, Mohammad Samar Ansari, Asra Aslam, Nadia Kanwal, Mamoona Naveed Asghar, and Brian Lee. A survey of modern deep learning based object detection models. *Digit. Signal Process.*, 126:103514, 2022. 2

[44] Libo Zhang, Junyuan Gao, Zhen Xiao, and Heng Fan. Animaltrack: A benchmark for multi-animal tracking in the wild. *Int. J. Comput. Vis.*, 131(2):496–513, 2023. 2