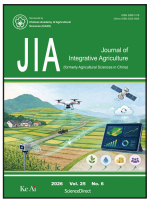
Journal of Integrative Agriculture 2026, 25(6): 2534–2544

KeAi CHINESE ROOTS GLOBAL IMPACT

Available online at ScienceDirect

Journal of Integrative Agriculture



RESEARCH ARTICLE

Promote computer vision applications in pig farming scenarios: High-quality dataset, fundamental models, and comparable performance

Jiangong Li[1, 2*#], Xiaodan Hu[3*], Ana Lucic[4], Yiqi Wu[1], Isabella C.F.S. Condotta[5], Ryan N. Dilger[5], Narendra Ahuja[3], Angela R. Green-Miller[6]

[1] *State Key Laboratory of Animal Nutrition and Feeding, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China*
[2] *Key Laboratory of Equipment and Informatization in Environment Controlled Agriculture, Ministry of Agriculture and Rural Affairs/College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China*
[3] *Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana 61801, USA*
[4] *Applied Research Institute, University of Illinois at Urbana-Champaign, Champaign 61820, USA*
[5] *Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana 61801, USA*
[6] *Department of Agricultural and Biological Engineering, University of Illinois at Urbana-Champaign, Urbana 61801, USA*

Highlights

- Proposed benchmark dataset creation protocols for animal recognition tasks.
- Evaluated the robustness of state-of-the-art (SOTA) computer vision models in pig production scene.
- Directions to zero-shot computer vision modeling opportunities.

Abstract

Computer vision is widely recognized as an influential technology in the field of precision management of animals. Emerging studies have demonstrated the potential to improve pig health and welfare through animal surveillance systems and computer vision (CV) algorithms. However, the lack of benchmark datasets and robust fundamental algorithms restrict CV applications for the commercial use. This study aims to bridge the gap between technology development and commercial applications in pig farming scenarios by introducing a general-purpose dataset (PigLife), comparing benchmark performances of foundational CV algorithms and model development workflows. The PigLife dataset contains video clips and images (38 short video clips, 2K image frames, 22K pig instances) across most pig production phases in a typical commercial pig farm: Breeding and Gestation, Farrow to Wean, Weaning & Nursery, and Growth to Finish. Three detection algorithms (Faster R-CNN, RetinaNet, TridentNet) and three segmentation algorithms (Mask R-CNN, MViTv2, PointRend) were trained on the PigLife dataset from scratch. Fine-tuning of pre-trained models (YOLO8-m, Faster-RCNN-r50) and no-training from zero-shot models (CLIP-SAM, Grouddino-HQSAM) were also evaluated to suggest faster CV development workflows for commercial applications in pig farming. This study emphasizes the necessity of a benchmark dataset for evaluating the robustness of algorithms and identifying the remaining difficulties and challenges across various algorithms. Furthermore, developing CV models from pre-trained algorithms or zero-shot models showed better performance and a faster process, which could reduce barriers when developing high-performance CV products in pig production industry.

Keywords: PigLife, pig recognition, benchmark performance, transfer learning, zero-shot algorithm

1. Introduction

Modern pig production systems confine large numbers of pigs that are overseen by a small number of relatively low-skilled caregivers (Li J *et al.* 2022). According to the United States Department of Agriculture (USDA), the loss and death rate of pig production in the United States (>13 million head) was around 10% in 2020. Therefore, automatic pig monitoring and precise care of individual animals are the next-generation

Received 27 February 2024; Received in revised form 6 June 2024; Accepted 26 June 2024; Available online 22 August 2024
Xiaodan Hu, E-mail: xiaodan8@illinois.edu; #Correspondence Jiangong Li, Tel: +86-10-62732763, E-mail: jli153@cau.edu.cn
* These authors contributed equally to this study.

© 2026 CAAS. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). Peer review under responsibility of Editorial Board of *Journal of Integrative Agriculture*.
doi: 10.1016/j.jia.2024.08.014

concepts for the success of the modern livestock industry, known as precision livestock farming (Berckmans 2017). As an affordable and non-invasive system, digital camera and computer vision are widely recognized as influential technologies in precision livestock farming (Yang and Xiao 2020; Neethirajan and Kemp 2021).

Vision-based pig recognition is the foundational technology that builds animals' "digital representations" and links all relevant growth variables to animal management and commercial strategies (Norton *et al.* 2019). Integrating information about pigs from vision systems will assist farm workers with many routine tasks and enable precise care for each animal, such as stress management (temperature, abusive handling, overcrowding, etc.) (Fernandes *et al.* 2020; Wu *et al.* 2022), welfare improvement (space, mobility, preferences) (Benjamin and Yik 2019; Chen *et al.* 2020; Leonard *et al.* 2021), medicine control (sickness detection, recovery) (Matthews *et al.* 2017), and growth evaluation (weight, body measurement) (He *et al.* 2021; Qiao *et al.* 2021). The relevant fundamental computer vision tasks with respect to pig recognition can be divided into three categories:

Detection, recognition of the pig, and its identification.

Detection or segmentation of the pig's geometry.

Recognition or classification of a pig's visual body characteristics.

Current pig recognition studies tend to focus on a single growth stage (e.g., gestation, farrowing and lactation, grow-out, etc.) and unique housing type (Yang and Xiao 2020). Although several research studies have already introduced vision augmentation and dataset variation into algorithm development, we are not aware of any studies that have evaluated the robustness of any foundational models (detection, segmentation, etc.) to different environments, various vision augmentation, and various animals. The PigLife dataset incorporates scenes that span different growth stages and different housing conditions, which represents a more challenging scenario for computer vision tasks and applications.

With the rapid development of computer vision algorithms, large generalized models had great success with various tasks and data distributions beyond the training dataset (Kirillov *et al.* 2023). These models were built on a large open-source dataset (COCO dataset, ImageNet, etc.) that includes millions of images and billions of annotations. Earlier studies have suggested the importance of establishing standard open-source and generalized datasets in pig farming that enable the development of robust basic algorithms and applications for research purposes and commercial uses (Li J *et al.* 2022). As shown in Table 1, only a few swine-focused datasets are publicly available. These datasets are either too task-specific or prepared for a single application, which prevents computer vision development for researchers without dataset access (Bergamini *et al.* 2021). Meanwhile, image datasets and visual scenes in pig housing environments are not incorporated into large datasets, such as ImageNet, COCO, and SA-1B. The general visual issues in pig farms, such as pig overlapping, illumination, and occlusion, need to be brought to the forefront and solved by each iteration of generalized foundational models (Tian *et al.* 2019; Riekert *et al.* 2020).

The objective of this study was to fill the gaps between technologies in research and actual applications in pig farming scenarios by introducing a general-purpose dataset, comparable benchmark performances of up-to-date algorithms, and appropriate development workflows for pig-specific computer vision applications. We employed a standard workflow to prepare the PigLife with sufficient dataset description and precise annotation regarding abundant vision augmentation, diverse housing environments, and various animal conditions (growth stage, size, sex). Several well-known detection and segmentation models were trained from initial random model weights on the PigLife dataset, and the performance was compared with the performance of transfer learning models and zero-shot models to establish the benchmark performances of detection and segmentation models on pig recognition tasks. This study established the foundation for designing & evaluating CV applications in pig farm scenarios and would advance the research on pig recognition algorithms toward actual application and commercial use.

2. PigLife dataset

PigLife is a public dataset that contains video clips and images across the pig production cycle in different systems: breeding, gestation, farrow, wean, nursery, and growth. To promote the development of robust computer vision models and algorithms in pig farms, we document the dataset construction workflow in this paper and publish the first subset (Sub-1) in PigLife (https://data.aifarms.org/view/piglife) to introduce and illustrate the use of a benchmark dataset. The PigLife dataset has the following features:

Encoded filename to indicate different pig production stages, vision augmentation (view angles, illuminations, etc.), and environment diversity.

Table 1 Pig-related dataset with public access

Reference	Annotation	Images/videos	Resolution	Dataset description
Tian *et al.* (2019)	Dots	2,984 images from internet and self-capture	72×72	Group-housed pigs; variation in number of pigs, illumination, view angles, overlapping
Psota *et al.* (2020)	Keypoint, ID	15 videos, 30 mins each, 5 frames per second (fps)	2,688×1,520	Group-housed pig; variation in environment, number of pigs, age of pigs
Riekert *et al.* (2020)	Boxes, 2 behaviors	7,277 annotated frames, 10 fps	1,280×720; 640×480	Group-housed pig; variation in view angles, environment, lighting, number of pigs
Bergamini *et al.* (2021)	Boxes, IDs, 5 behaviors	7,200 annotated frames, 6 fps	1,280×720	Group-housed pig; single environment, single group
Shirke *et al.* (2021)	Key points, action	20 5.5-min videos, 30 fps	1,024×1,024	Laboratory setup, single pig and environment

38 video files (1–1.5 minutes), 2,146 frames, 374 individually different pigs, 22K animal instances in total; manual annotation every 15 frames.

A complete commercial pig production line, including Breeding and Gestation, Farrow to Wean, Weaning & Nursery, and Growth to Finish.

Consistent bounding box and segmentation to the same animal identification in video files.

2.1. Dataset collection & annotation

The commercial breed pigs of (PIC Camborough®) were continuously recorded for various farms and angles through 4K resolution (3,840×2,160) and 15 frames per second (fps) using surveillance cameras (IP8M-T2599E, Amcrest) at the University of Illinois, Urbana-Champaign agricultural animal research farms. Video contents then selected by an animal expert to include different animal behaviors across pig production stages. All video clips were converted to MP4 format with 15 fps. All images were converted to JPEG format. Images extracted from video clips were named sequentially to the standard format. As shown below, the filename is encrypted by the combination of 4-digit codes and separators (sub-category appendant [s]; sequence or number [-]) to demonstrate the content of a video or an image. Filename contains five parts: growth stage (SID, 1,000–1,100), image description (IID, 1,101–2,000), housing condition (EID, 2,001–5,000), animal description (AID, 5,001–6,000), frame number (FN, 0–INF).

Video: SIDsIIDsEIDsAID.mp4

Image: SIDsIIDsEIDsAID(FN).jpg

Example:

Encoded filename:

1000s1110s4001-2s5301-1(0).png

Decoded filename:

Image was about pigs In PreWean stage, took from RGB camera (front video).

The scene is about Mix housing, research type-0001 (2 separated pens).

Pigs are in active.

This is the frame zero in a video sequence-one.

The first 4-digit code (SID) indicates which pig growth stage the video demonstrates. The second 4-digit code (IID) explains how video is recorded, such as camera type and view angle. The third 4-digit code (EID) with the sequential number demonstrates the housing type and the number of pens or stalls within the video. The last 4-digit code (AID) briefly describes the animal status in the video, such as active, aggressive, etc. The meaning of all four-digit codes is summarized in Appendix A for post-analysis. Annotation files are converted into JSON files in COCO format.

All videos and images were manually masked in black to remove the background with human and irrelevant information, as shown in Fig. 1. Images were extracted from the masked video clips with the sampling rate of 1 image (frame) per second. Pig segmentation annotation was processed manually by five volunteer students through VGG Image Annotator (https://www.robots.ox.ac.uk/~vgg/software/via/). Each pig instance polygon has an average of 43 vertices and takes an average of 2 mins (zoom-in/out; create a polygon, assign ID) for drawing. A total of 2 rounds of quality checks were performed separately by two authors to uphold the data quality. Pig annotation across images from the same video clip was merged into a video annotation file, and pig identification was consistent across frames, which was confirmed by the authors. Video annotation was stored in a JSON file following the COCO format. A detail

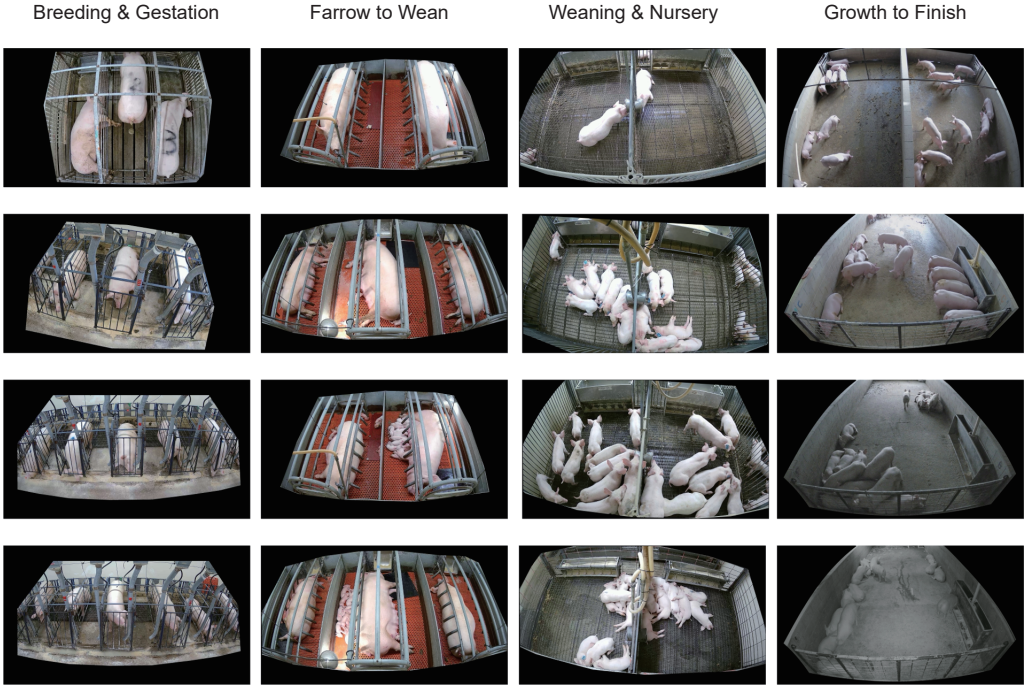


Fig. 1 Brief scenes of PigLife Sub-1 dataset.

description of data preparation was described in "Dataset Datasheet" in Appendix A following the Microsoft datasheet format (Gebru *et al.* 2021).

2.2. Dataset variation

As a benchmark dataset, PigLife is designed to embrace variation and diversity that can reflect the real application scene in pig farms. The commercial pig production line raises pigs in different housing types at different ages, as shown in Fig. 1. The PigLife dataset contains individual stalls housed female pigs for breeding and gestation. Female pigs and their offspring stay together in the same pen (mix-housed) during the farrowing and breast-feeding period. Then, weaned pigs are raised as a group until finished. Therefore, vision conditions and computer vision difficulties differ significantly among pig production stages. As shown in Fig. 2-A, we balance the number of images in different production stages. Number of pigs in the image varies from 2–25. In general, images of the gestation stage have the least number of pigs (3–6) per image and the significant occlusions from structure, pipes, drinkers, etc. The image of the farrowing stage has the highest contrast of pig sizes and pig overlapping. Images of the nursery stage have the greatest number of pigs and the highest pig density. The camera usually cannot record an entire pen for pigs in the growth stage and is generally deployed on the side view to expand the view. Therefore, videos during the grow-finish stage may have distorted shapes, with individual pigs also going out and coming into disappearing from the camera view.

As shown in Fig. 2-B, the number of pigs with different

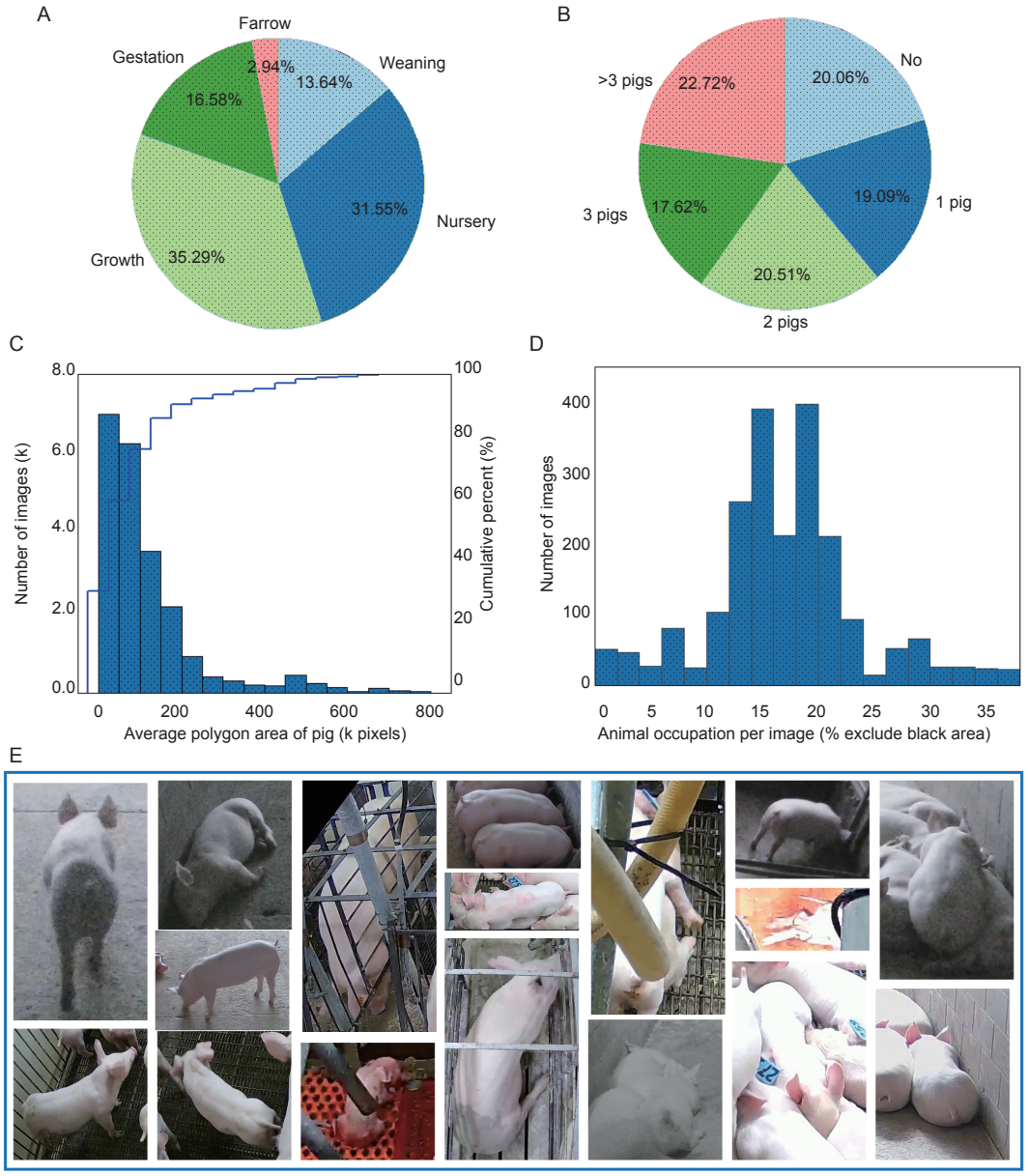


Fig. 2 Demonstration of PigLife Sub-1 dataset variation. A, images at different pig growth stages (%). B, pig instances overlapping with different numbers of pigs (%). C, number of pig instances with different sizes. D, number of images with different animal occupation areas. E, sample images.

overlapping scenarios is well-balanced. Pigs without occlusion and overlapping took around 20% of the annotated pig instances. Another 22% of pig instances overlap with >3 pigs, which are densely annotated to split each of them with shared contour boundaries. Around 68% of the dataset have vision difficulties (occlusion, pig overlapping, lightning problems), which are typical scenes in pig production farms. As shown in Fig. 2-C and D, the variation of pig occupation within the image is made from pig size, density, and overlapping. To record as many pigs as possible over the field of view, the typical pixel sizes of pigs are less than 200K pixels over 8M-pixel images. Pigs with larger sizes are also selected to increase the variation. Pig size, ground size, and the field of view lead to the difference in pig density over the background. The proposed dataset includes extreme cases with 1% occupation and up to 38% occupation.

3. Training models from scratch on PigLife dataset

In this section, we assess the performance of various well-known detection and segmentation models through the proposed PigLife dataset. Images and annotations in PigLife dataset were randomly pre-split into one training-testing sets (80%, 20%). We assess the Faster R-CNN model (Massa and Girshick 2018), RetinaNet model (Lin *et al*. 2017), TridentNet model (Li *et al.* 2019) for the pig detection task, and Mask R-CNN model (Massa and Girshick 2018), MViTv2 model (Li Y *et al.* 2022) and PointRend model (Kirillov *et al.* 2020) for the pig segmentation task. All models underwent 300 iterations of training on the PigLife training set, utilizing the default settings of their respective projects within the Detectron2 framework. Importantly, we trained these models from scratch without any additional parameter tuning. Subsequently, we tested these models on the test set.

The training process for the detection models followed a uniform configuration. These models were built upon the ResNet-50 architecture with a Feature Pyramid Network (FPN) backbone. Standard convolutional layers were employed for mask prediction, while fully connected heads were used for bounding box predictions. The training was conducted on four NVIDIA RTX A4000 GPUs, with a batch size of 4. To align with our specific task, we set the number of classes to 1, representing the target class 'pig'. The algorithm performance is evaluated by the standard evaluation metrics: standard box Average-Precision (AP) for pig detection task and mask Average-Precision for pig segmentation task. The AP reported here is the same as mean average precision, which is averaged over all categories. AP_{50} and AP_{75} are computed at a single Intersection Over Union (IoU) of .50 and .75, respectively. AP_M is the AP for medium objects where 322<area<962. AP_L is the AP for large objects where area>962.

3.1. Pig detection

The pig detection task is to classify and localize individual pigs from images or videos accurately. The detection task establishes the basis for efficient pig counting, tracking, and identification, having a high impact on evaluating disease processes and monitoring pig health. Object detection has been widely explored in many studies (Liu *et al.* 2023; Wang C Y *et al.* 2023). Region Proposal Network (RPN) is a fully convolutional network that predicts object boxes/proposals and scores at each position simultaneously. The faster R-CNN model uses RPN for generating region proposals and merges into a single network during training to generate object proposals. The RetinaNet model proposes a novel focal loss that focuses on a sparse set of hard examples to address the class imbalance between vast numbers of easy negatives and less positive examples. TridentNet model creates scale-specific feature maps through parallel branches where convolutions share the same parameters but with different dilation rates, to detect objects of varying scales. These models have achieved good performances on several benchmark datasets for object detection.

Table 2 shows the quantitative comparison of these methods on the PigLife dataset. The PigLife dataset, with its unique composition of different pig sizes from various production stages, presents a complex challenge. The presence of female pigs alongside their offspring in the same pen introduces significant variability in object scale and density, which is compounded by the possibility of occlusions and overlapping objects. The Faster R-CNN and RetinaNet models perform poorly on medium pigs than large ones, indicating how critical context is for the recognition and detection of smaller objects. The scale of the object is a critical factor in detection tasks, and it is apparent that the receptive fields of these networks are not optimally attuned to the smaller objects present in the PigLife dataset. In contrast, the TridentNet model achieves the best performance on all of the AP scores and also obtains significant improvement over the AP_M score, which may primarily benefit from their proposed scale-specific feature maps. These maps are specifically designed to cater to objects of varying sizes by adjusting the receptive field accordingly, resulting in improved accuracy for medium-sized objects. However, the TridentNet model does not achieve the same level of performance on detection task over PigLife scene than the average performance in a generalized scene, such as COCO dataset.

Fig. 3 also shows consistent qualitative results as that in Table 2. Group-housed pigs are very likely to rest next to each other. The difficulty of the pig detection task is mainly about the distinction of overlapping boundaries of similar pigs. The raw detection results of the RetinaNet model in Fig. 3 shows RetinaNet model generates multiple boxes around the same pig. Non-Maximum Suppression (NMS) filter may be able to remove these redundant boxes. And also, the pig occluded by the pipes with only part of the body available. Unlike Faster R-CNN and RetinaNet models that

Table 2 Quantitative model comparison on pig detection tasks

Models	AP	AP_{50}	AP_{75}	AP_M	AP_L	Inference time
Faster R-CNN	29.1	40.5	35.4	2.9	30.1	0.2 s
RetinaNet	29.8	40.6	33.8	3.3	31.2	0.1 s
TridentNet	**37.2**	**41.5**	**40.1**	**35.5**	**37.3**	**0.1 s**
TridentNet*	42.0	63.5	45.5	47.0	56.9	–

*, performance of TridentNet model on the COCO minival-set. Bold, achieve the best performance in pig detection tasks.

miss the detection of the occluded pig, the TridentNet model can detect this pig with a clean and accurate box.

3.2. Pig instance segmentation

The pig segmentation task is partitioning a digital image into multiple pig segments. It traces the contours of the pigs at the pixel level, preserving richer information. The estimated pig segmentation masks can be used in many downstream tasks, such as pig weight estimation, key points estimation, and 3D reconstruction. Significant progress has been made in the instance segmentation methods (Fang *et al.* 2023). The Mask R-CNN model is simple to train and extends the Faster R-CNN model by adding a branch to predict an object mask in parallel with the existing branch for bounding box recognition (Massa and Girshick 2018). The MViTv2 model is an improved version of the multiscale transformer model (MViT model), which upgrades the performance by optimizing the pooling attention along two axes: decomposed relative positional embeddings and residual pooling connections (Li Y *et al.* 2022). The PointRend model proposes a module that performs point-based segmentation predictions at adaptively selected locations based on an iterative subdivision algorithm (Kirillov *et al.* 2020).

In the task of pig segmentation, Table 3 shows that the Mask R-CNN and MViTv2 models vastly outperform the PointRend model on most of the scores. The Mask R-CNN model has a higher AP that indicates superior instance mask segmentation, but relatively lower performance in other scores shows that the Mask R-CNN model has difficulties in overall object detection, object localization, and classification. In Fig. 4, the Mask R-CNN model performs well in precisely delineating object boundaries and generating accurate instance masks. MViTv2 model can detect the top-center pig (with the yellow mask) occluded by the tube in the group-housed scene and the leftmost pig (with the orange mask) in the individual-housed scene with only a small part of its body visible.

3.3. Error analysis

In modern computer vision applications for automated animal monitoring, pig detection can determine a pig's specific position, and instance segmentation can anticipate a relatively comprehensive mask of each animal. Although target detection methods such as Faster R-CNN, RetinaNet, and TridentNet models, instance segmentation approaches like Mask R-CNN, MViTv2, and PointRend models have been shown the ability in computer vision recognition of pigs, several obstacles remain in the actual breeding setting.

Pig detection in complex environment In the Faster R-CNN and RetinaNet models, overlapping is one of the major drivers of a substantial proportion of false negative detections of pigs. In an intense feeding scenario, the target detection algorithm incorrectly recognizes numerous pigs as one pig or confuses one pig as multiple objects, as shown in Fig. 5. The Faster R-CNN and RetinaNet models generate several candidate regions, most of which do not contain the target. The uncleared boundary of pig instances leads to similar

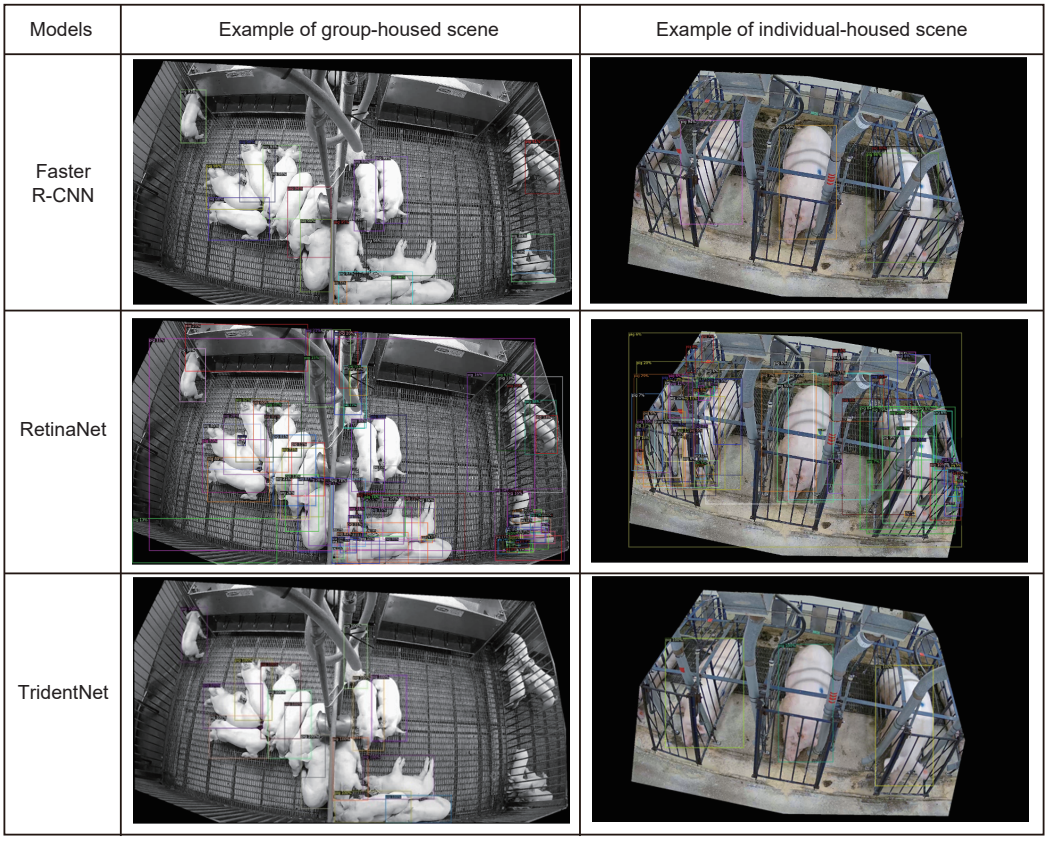


Fig. 3 Model comparison on the pig detection task.

spatially constrained contexts, exacerbating the difficulty of recognizing the overlapped pigs. NMS is a critical feature in established detection models, including Faster R-CNN and RetinaNet, utilized to minimize the redundancy of overlapping detection boxes. Despite its standard application, further optimization of the NMS approach might be proposed to overcome the overlapping problem of candidate frames and increase the detection accuracy of pig-overlapped scenarios.

Occlusion was another primary cause of misdetections in individual housed scenarios. As depicted in Fig. 6, the accuracy of object detection models is compromised when parts of a pig are obscured by other elements within the scene. These elements, such as pipes, fences, and troughs, do not serve as direct detection targets but disrupt the models' ability to detect and segment the pigs accurately. This is especially true for smaller targets, which are more likely to be conflated with or hidden by the surrounding foreground objects. The target pigs could be defined as "middle contexts" compared with background (floors or walls). Most benchmark algorithms are well-designed and trained for foreground elements but could be better for objects in the middle layer.

Table 3 Quantitative model comparison on segmentation tasks

Models	AP	AP_{50}	AP_{75}	AP_M	AP_L	Inference time
Mask R-CNN	**39.7**	40.4	36.8	0.9	30.9	0.1 s
MViTv2	32.7	**41.6**	**40.3**	**3.0**	**34.0**	0.1 s
PointRend	15.6	36.1	10.9	0.7	16.2	1.7 s

Bold, achieve the best performance in pig detection tasks.

Pig instance segmentation in complex environment The incorrect segmentation examples using Mask R-CNN, MViTv2, and PointRend models are illustrated in Figs. 7 and 8. One of the typical errors in pig segmentation among all three models is the invalid masks targeted to the partial pig instance. As shown in Fig. 7, the Mask R-CNN model randomly generates multiple masks for a single pig without occlusion or overlap. The segmentation quality of the MViTv2 model objects is marginally better than the Mask R-CNN and PointRend models but still generates several masks that do not represent pig instances. Minimizing the number of masks or scoring the confidence of masks should be able to address and correct this error.

Poor segmentation performance is most noticeable when the pig is at the image boundary, and the fence shields the pigs. The integrity of the target object influences the segmentation performance of these three approaches. Compared to the grouped-housed pigs (Fig. 7), the foreground occlusion in the individual-housed scenario (Fig. 8) is more severe. In the individual-housed scene, compared to pigs in the center of the view, pigs on the side are relatively small in area but more occluded by stall bars from the side view, leading to worse segmentation performance. The diverse camera perspectives and facility characteristics directly impact segmentation quality. Lower and pan-tilt camera angles result in more foreground occlusions but more context details at lower levels (e.g., feet, joints, feed). A camera

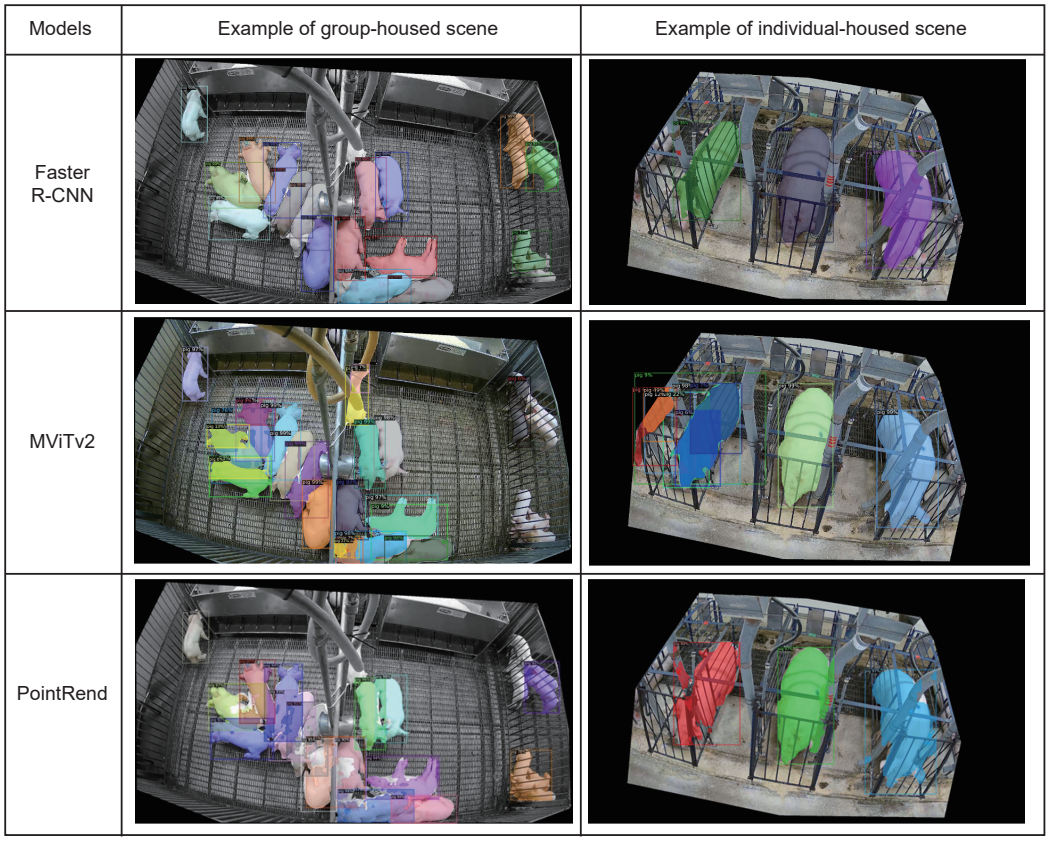


Fig. 4 Methods comparison on the pig segmentation task.

closer to the animals will expose more details for a single animal but cover fewer animals. The segmentation effect of the three approaches for the target with large pixel size in the center of the view is superior to that of the target with small pixel size at the picture's edge. There is a trade-off between algorithm performance, camera economy, and the designed visual purpose.

4. Transfer learning and zero-shot models

4.1. Algorithm development

The foreground occlusion of edge objects was the pig production scene's primary detection and segmentation error source. Due to the substantial similarities between overlapped pigs and foreground occlusions, detection and segmentation methods typically cannot handle misjudged split-animal instances in the local image region. TridentNet uses dilated convolution with the best detection results that is a potential step in this direction, as it provides an extended field of view to capture the full extent of partially visible objects. The detection effect and segmentation quality are also affected by visual circumstances. Varying degrees of erroneous detection and mis-segmentation occurred due to diverse monitoring views and target size caused by fixed viewing angles. The detection model's overall effect can be improved by designing the receptive field based on the size of the target scale, and the future iterations of the model could benefit from integrating scale-adaptive strategies and camera perspective adjustments to further refine detection accuracy in these complex scenarios. Furthermore, the test results showed that an overhead view of the image was more efficient in spotting pigs and distinguishing them from the background when using these methods. In addition to focusing on recognizing smaller objects, the assessment of incomplete individuals at the image's edge should also be considered.

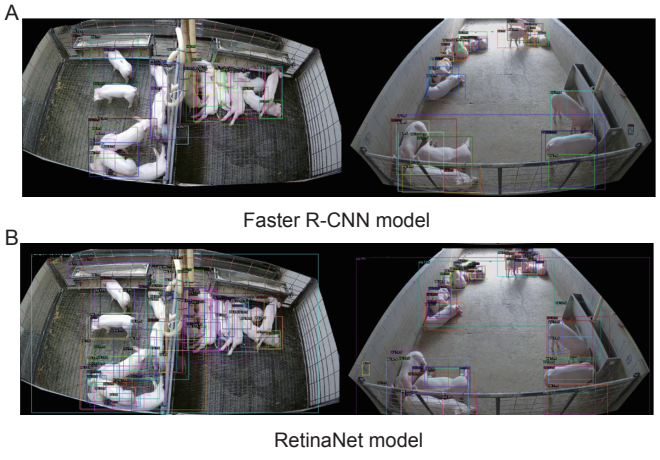


Fig. 5 Example of false detection in group-housed scene. A, Faster R-CNN model. B, RetinaNet model.

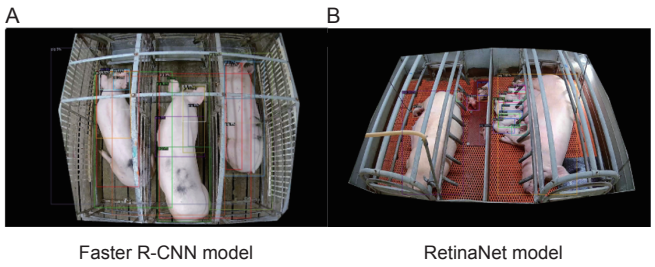


Fig. 6 Example of false detection in individual-housed scene. A, Faster R-CNN model. B, RetinaNet model.

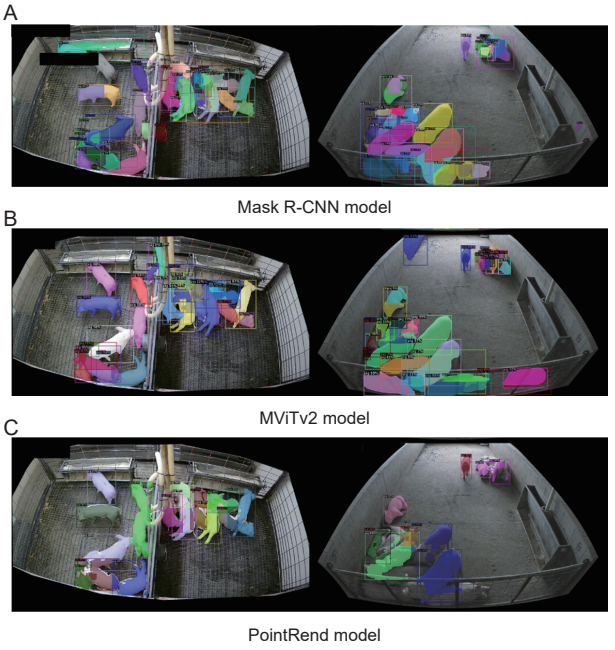


Fig. 7 Example of false segmentation in group-housed scene. A, Mask R-CNN model. B, MViTv2 model. C, PointRend model.

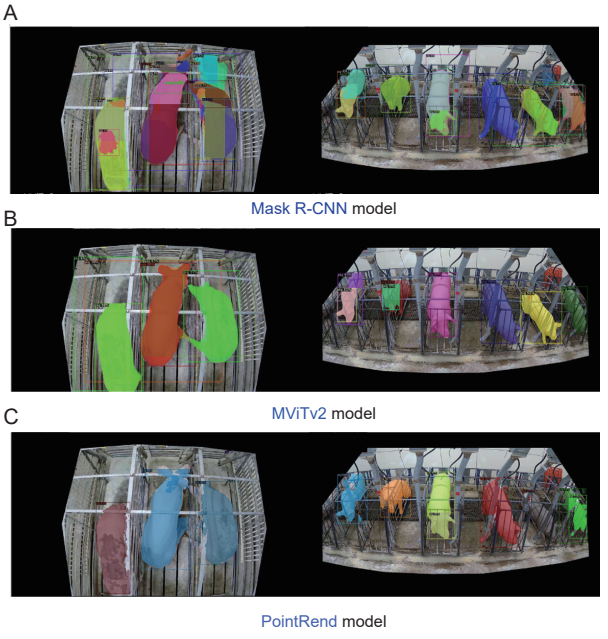


Fig. 8 Example of false segmentation in individual-housed scene. A, Mask R-CNN model. B, MViTv2 model. C, PointRend model.

Another important use of PigLife dataset is to develop algorithms from pre-trained model. Pre-trained models were initially well developed on a large and representative dataset (such as COCO, ImageNet), and could be further improved for a specific recognition scenario by a smaller dataset. There are many researches documented the use of pre-trained model on AI applications in livestock production scenarios (Koger *et al.* 2023; Pan *et al.* 2023; Wang M *et al.* 2023). We also used the training dataset of PigLife Sub-1 to fine-tune several popular object detection models and evaluated it with the testing dataset as shown in Table 4. The detection performances are significantly improved compared with the performance of the model developed from scratch. Training RetinaNet model from scratch with this small dataset only achieved 29.8 of average precision (Table 2) but was significantly improved to 78.6 of average precision with pretrained models. The use of pre-trained weight can speed up the training process, improve network convergence and enhance the specificity on a scenario that was demonstrated on smaller dataset. However, the failure of these algorithms is mostly about the recognizing small-medium pixel pigs, split animal instances, and pig on the edges. Recognition of dynamic-static characteristics of animals in video sequences may rebuild the animal integrity between foreground and background contexts.

4.2. Zero-shot algorithms

The rapid development of foundational or benchmark models will change and promote artificial intelligent applications in our society. In 2023, the Segment-Anything model (SAM) was firstly introduced by Meta AI and showed a stunning and general ability to segment various instances (Kirillov *et al.* 2023). SAM accepts various prompts that refers what to segment without additional training or dataset and enables the zero-shot grounding of unfamiliar objects on images. Although SAM does not need new data to address new segmentation tasks it still achieves a relatively high accuracy. As shown in Fig. 9, the SAM can segment the region of interest in a general pig production scene without any training. Although SAM still has challenges in segmenting pig instances in the occlusion scenario, it overcomes the problem of multiple masks for a single instance by introducing confidence scoring mechanism.

The SAM is able to estimate a valid mask when given a prompt, such as foreground/background points, box, and free-form text, that indicates what to segment. To achieve the particular segmentation of pig instance, the “pig” prompt is a prerequisite for the segmentation task. The fact that we are dealing with monitoring of one species in this instance helps us with prompt identification. The performance of pig segmentation is the combined effects of two processes: prompting quality and segmentation ability. Therefore, we tested two benchmark approaches. We applied the Contrastive Language-Image Pre-Training (CLIP) model to enable natural language instruction to reference visual concepts and zero-shot ability of the SAM (Radford *et al.* 2021). We also applied the BERT-styled object detection algorithm called Grounding DINO model that takes arbitrary text input as model input prompts to obtain bounding box information corresponding to the text (Liu *et al.* 2023). The bounding boxes obtained in this way were then feed into the High Quality Segment-Anything model (HQ-SAM), which achieves more fine-tuned segmentation results while retaining the SAM’s excellent zero-shot capability of SAM (Ke *et al.* 2024). As shown in Table 5, SAM-based algorithms achieved the same level of performance as the MViT2 model. With the stronger prompting ability of bounding boxes and showed the advanced version of the SAM-style model, the segmentation performance has tremendous improvement compared to the most up-to-date benchmark models.

Table 4 Quantitative model comparison of fine-tuned pretrained models

Models	AP	AP_{50}	AP_{75}	AP_{M}	AP_{L}	Source
YOLO8-m	**91.3**	**98.0**	**96.9**	**77.2**	**92.0**	Nazarkevych *et al.* (2024)
Faster-RCNN-r50	78.5	97.6	91.1	62.6	79.1	Bergamini *et al.* (2021)
RetinaNet-r50	78.6	97.9	89.1	64.9	79.3	Lin *et al*. (2017)

Bold, achieve the best performance in pig detection tasks.

5. Limitations and future works

5.1. High-quality and large-scale dataset

Developing the PigLife dataset was a very time-consuming and costly activity, involving meticulous data collection, annotation, and balance processes to ensure the data is representative, correct, consistent, and includes the necessary level of detail. This study demonstrates the process of development of a high-quality benchmark dataset. We recognize that the ability to evaluate, test, and compare algorithms is very critical only possible with access to high-quality datasets. However, this dataset is not large enough to comprise enough vision augmentation, animal variation, and environmental differences, which should be continuously expanded in the future (Li J *et al.* 2022). Zero-shot algorithms significantly improve the robust recognition ability to pigs. Although the performance is still not applicable in for commercial application, there is a huge potential to implement it as data engine that pre-annotates various data to reduce the heavy labor works. This approach has the potential to exponentially enlarge the livestock dataset in the future.

Meanwhile, comprehensive animal behavior recognition and animal health evaluation still require expert input on the visual contexts of images or video sequences. In the

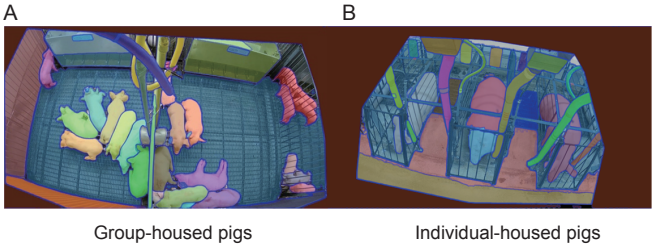


Fig. 9 Dense segmentation using the Segment-Anything model (SAM) on PigLife Sub-1 dataset. A, group-housed scenes. B, individual-housed scenes.

future, there will be a shortage of professional multi-modal data annotation that covers multiple instances with rich semantic information about the housing environment, animal status, and even more comprehensive context of the pig. Professional labeling of animal behaviors and actions should be highlighted to promote advanced AI applications, such as automatic animal analysis (Siegford *et al.* 2023). There is an urgent need to create a tool and standard that is easy to use for experienced contributors and scalable to provide the ability for comments and labels. We also propose a demonstration of the Animal Video Analysis Tool (https://github.com/AIFARMS/AVAT) for text description inputs to images or video annotations in an accessible and user-friendly layout. This open-source annotation tool is supposed to provide an initial demo that links professional inputs to visual annotations for professional AI model development in the future.

5.2. Importance of fundamental CV models in commercial use applications

Continuous animal monitoring and long-term behavioral analysis are two commercial needs in intensive pig production scenes. However, fundamental CV models (detection, segmentation, identification, etc.) are still the bottleneck preventing commercial applications. Existing detection and segmentation approaches for pigs are still facing challenges, particularly in various commercial farms with overlapping and foreground occlusion problems. With the advancement of computer vision technologies, there is still room for algorithmic enhancement to overcome current constraints and support the achievement of long-term animal monitoring and identification of high-level behaviors in real-life scenarios. Fine-tuning pre-trained algorithms with a small scenario dataset has shown a great step forward to achieving high-performance detection and segmentation, which allows us to see a bright future for AI applications in the pig production industry.

5.3. Behavioral analysis using results from fundamental CV models

Computer vision technology has been widely recognized as an efficient method (e.g., DeepLabCut) to track and analyze animal behaviors in laboratory experiments (Lauer *et al.* 2022). Many studies have demonstrated the successful use of simple and basic CV models (detection, segmentation, identification, tracking) in the applications of feeding behavior recording, locomotion analysis, recording of the access to a region of interest, analyzing the behavioral sequences in livestock farm settings (Schmidt *et al.* 2022; Melfsen *et al.* 2023). Recognizing behaviors based on feature differences in locations, shapes, and key points of animal graphic instances is more versatile compared with specific behavior detection algorithms (e.g., lying, lameness, ear biting, and sucking). The explosive growth of animal surveillance data will allow us to discover new understanding and digital indexes about animal welfare and health. Animal-related practitioners and students could benefit from automatic behavioral analysis derived from the results of basic and open CV models in practice.

Table 5 Quantitative methods comparison on the Segment-Anything model (SAM)-based algorithms

Models	AP	AP_{50}	AP_{75}	AP_M	AP_L
CLIP/SAM	34.7	43.3	39.11	2.0	37.0
Grouddino/HQSAM	**54.6**	**81.7**	**58.5**	**7.8**	**56.3**

Bold, achieve the best performance in pig detection tasks.

6. Conclusion

High-quality datasets provide sufficient variability, diversity, and representativeness of the natural world, enabling computer vision models to learn and generalize intelligence ability. The PigLife dataset is designed to evaluate and compare the performance of various computer vision algorithms. In the experiment of training models from scratch, the TridentNet model achieved the best performance in the pig detection task with a mean AP of 37.2%. The Mask R-CNN model achieved the best performance in the pig segmentation task, with a mean AP of 39.7%. In the experiment of fine-tuning pre-trained models, YOLO8-m model achieved significant improvement in performance with a mean AP of 91.3%. In addition, the combination of two zero-shot models (Grouddino model, HQSAM model) achieved a competitive performance of mean AP of 54.6% on pig instance segmentation without any training process. The reported benchmark performances could represent the current artificial intelligence ability in pig recognition-related tasks, such as identification, tracking, animal counting, etc. Implementing transfer learning with a small, high-quality dataset could achieve a promising performance and enable faster development of computer vision applications. Zero-shot models have irreplaceable advantages of robustness on pig recognition tasks, which could be further improved through enhancing referring and grounding ability. This study established the foundation for designing & evaluating CV applications in pig farm scenarios and would advance the research on pig recognition algorithms toward actual application and commercial use.

Acknowledgements

This research received support from the Key Laboratory of Equipment and Informatization in Environment Controlled Agriculture, Ministry of Agriculture and Rural Affairs, China (2011NYZD2204); the Chinese Universities Scientific Fund, (2024TC097); the Agriculture and Food Research Initiative of United States (AFRI, 2020-67021-32799); and the USDA National Institute of Food and Agriculture (project accession 1024178).

Declaration of competing interest

The authors declare that they have no conflict of interest.

Declaration of generative AI and AI-assisted technologies in the writing process

The authors declare that they did not use AI in the preparation and writing of this manuscript.

Ethical statements

Only recorded animal videos were used in this study. The

research protocol was approved by the Ethics Committee of China Agricultural University (Beijing, China).

Appendix associated with this paper is available at https://doi.org/10.1016/j.jia.2024.08.014

References

Benjamin M, Yik S. 2019. Precision livestock farming in swine welfare: A review for swine practitioners. *Animals*, **9**, 133.

Berckmans D. 2017. General introduction to precision livestock farming. *Animal Frontiers*, **7**, 6–11.

Bergamini L, Pini S, Simoni A, Vezzani R, Calderara S, Eath R B, Fisher R B. 2021. Extracting accurate long-term behavior changes from a large pig dataset. In: *16th International Joint Conference on Computer Vision*, *Imaging and Computer Graphics Theory and Applications*, *VISIGRAPP 2021*. SciTePress, Austria. pp. 524–533.

Chen C, Zhu W, Oczak M, Maschat K, Baumgartner J, Larsen M L V, Norton T. 2020. A computer vision approach for recognition of the engagement of pigs with different enrichment objects. *Computers and Electronics in Agriculture*, **175**, 105580.

Fang Y, Wang W, Xie B, Sun Q, Wu L, Wang X, Huang T, Wang X, Cao Y. 2023. EVA: Exploring the limits of masked visual representation learning at scale. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. CVF, Canada. pp. 19358–19369.

Fernandes A F A, Dórea J R R, Rosa G J D M. 2020. Image analysis and computer vision applications in animal sciences: An overview. *Frontiers in Veterinary Science*, **7**, 551269.

Gebru T, Morgenstern J, Vecchione B, Vaughan J W, Wallach H, Iii H D, Crawford K. 2021. Datasheets for datasets. *Communications of the ACM*, **64**, 86–92.

He Y, Tiezzi F, Howard J, Maltecca C. 2021. Predicting body weight in growing pigs from feeding behavior data using machine learning algorithms. *Computers and Electronics in Agriculture*, **184**, 106085.

Ke L, Ye M, Danelljan M, Tai Y W, Tang C K, Yu F. 2024. Segment anything in high quality. *Advances in Neural Information Processing Systems*, **36**, 29914–29934.

Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg A C, Lo W Y. 2023. Segment anything. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. CVF, France. pp. 4015–4026.

Kirillov A, Wu Y, He K, Girshick R. 2020. Pointrend: Image segmentation as rendering. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. CVF, USA. pp. 9799–9808.

Koger B, Deshpande A, Kerby J T, Graving J M, Costelloe B R, Couzin I D. 2023. Quantifying the movement, behaviour and environmental context of groupliving animals using drones and computer vision. *Journal of Animal Ecology*, **92**, 1357–1371.

Lauer J, Zhou M, Ye S, Menegas W, Schneider S, Nath T, Rahman M M, Di Santo V, Soberanes D, Feng G. 2022. Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nature Methods*, **19**, 496–504.

Leonard S M, Xin H, Ramirez B C, Stinn J P, Dutta S, Liu K, Brown-Brandl T M. 2021. Static and dynamic space usage of late-gestation sows. *Transactions of the ASABE*, **64**, 151–159.

Li J, Green-Miller A R, Hu X, Lucic A, Mohan M M, Dilger R N, Condotta I C, Aldridge B, Hart J M, Ahuja N. 2022. Barriers to computer vision applications in pig production facilities. *Computers and Electronics in Agriculture*, **200**, 107227.

Li Y, Chen Y, Wang N, Zhang Z. 2019. Scale-aware trident networks for object detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. CVF, South Korea. pp. 6054–6063.

Li Y, Wu C Y, Fan H, Mangalam K, Xiong B, Malik J, Feichtenhofer C. 2022. MViTv2: Improved multiscale vision transformers for classification and detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. CVF, USA. pp. 4804–4814.

Lin T Y, Goyal P, Girshick R, He K, Dollár P. 2017. Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*. CVF, USA. pp. 2980–2988.

Liu S, Zeng Z, Ren T, Li F, Zhang H, Yang J, Li C, Yang J, Su H, Zhu J. 2023. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: *European Conference on Computer Vision*. Cham: Springer Nature. Switzerland. pp. 38–55.

Massa F, Girshick R. 2018. Maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. https://github.com/facebookresearch/maskrcnn-benchmark

Matthews S G, Miller A L, PlÖtz T, Kyriazakis I. 2017. Automated tracking to measure behavioural changes in pigs for health and welfare monitoring. *Scientific Reports*, **7**, 17582.

Melfsen A, Lepsien A, Bosselmann J, Koschmider A, Hartung E. 2023. Describing behavior sequences of fattening pigs using process mining on video data and automated pig behavior recognition. *Agriculture*, **13**, 1639.

Nazarkevych M, Kostiak M, Oleksiv N, Vysotska V, Shvahuliak A T. 2024. A YOLO-based method for object contour detection and recognition in video sequences. In: *Proceedings of CPITS-2024*. CEUR-WS, Ukraine. pp. 49–58.

Neethirajan S, Kemp B. 2021. Digital livestock farming. *Sensing and Bio-Sensing Research*, **32**, 100408.

Norton T, Chen C, Larsen M L V, Berckmans D. 2019. Precision livestock farming: Building 'digital representations' to bring the animals closer to the farmer. *Animal*, **13**, 3009–3017.

Pan Y, Zhang Y, Wang X, Gao X X, Hou Z. 2023. Low-cost livestock sorting information management system based on deep learning. *Artificial Intelligence in Agriculture*, **9**, 110–126.

Qiao Y, Kong H, Clark C, Lomax S, Su D, Eiffert S, Sukkarieh S. 2021. Intelligent perception for cattle monitoring: A review for cattle identification, body condition score evaluation, and weight estimation. *Computers and Electronics in Agriculture*, **185**, 106143.

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J. 2021. Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning*. ICVM, USA. pp. 8748–8763.

Riekert M, Klein A, Adrion F, Hoffmann C, Gallmann E. 2020. Automatically detecting pig position and posture by 2D camera imaging and deep learning. *Computers and Electronics in Agriculture*, **174**, 105391.

Schmidt T B, Lancaster J M, Psota E, Mote B E, Hulbert L E, Holliday A, Woiwode R, Pérez L C. 2022. Evaluation of a novel computer vision-based livestock monitoring system to identify and track specific behaviors of individual nursery pigs within a group-housed environment. *Translational Animal Science*, **6**, txac082.

Shirke A, Golden R, Gautam M, Green-Miller A, Caesar M, Dilger R N. 2021. Vision-based behavioral recognition of novelty preference in pigs. *arXiv*, **2106**, 12181.

Siegford J M, Steibel J P, Han J, Benjamin M, Brown-Brandl T, Dórea J R, Morris D, Norton T, Psota E, Rosa G J. 2023. The quest to develop automated systems for monitoring animal behavior. *Applied Animal Behaviour Science*, **265**, 106000.

Psota E T, Schmidt T, Mote B, C. Pérez L. 2020. Long-term tracking of group-housed livestock using keypoint detection and map estimation for individual animal identification. *Sensors*, **20**, 3670.

Tian M, Guo H, Chen H, Wang Q, Long C, Ma Y. 2019. Automated pig counting using deep learning. *Computers and Electronics in Agriculture*, **163**, 104840.

Wang C Y, Bochkovskiy A, Liao H Y M. 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. CVF, Canada. pp. 7464–7475.

Wang M, Li X, Larsen M L, Liu D, Rault J L, Norton T. 2023. A computer vision-based approach for respiration rate monitoring of group housed pigs. *Computers and Electronics in Agriculture*, **210**, 107899.

Wu D, Cui D, Zhou M, Ying Y. 2022. Information perception in modern poultry farming: A review. *Computers and Electronics in Agriculture*, **199**, 107131.

Yang Q, Xiao D. 2020. A review of video-based pig behavior recognition. *Applied Animal Behaviour Science*, **233**, 105146.

Executive Editor-in-Chief Xugang Luo
Managing Editor Yanyan Sun