RAFE-DETR: AN RT-DETR-BASED ALGORITHM FOR MULTI-BEHAVIOR DETECTION IN GROUP-HOUSED PIGS
/
RAFE-DETR：一种基于 RT-DETR 的群养猪多行为检测算法

Lihong RONG[1)], Fang SUN[1)], Xiusong LI[2)], Weilong ZHANG[1)], Chengguo HAN[1)], Zhimin TONG*[1)]
[1)] College of Mechanical and Electrical Engineering, Qingdao Agricultural University, Qingdao / China
[2)] Qingdao Big Herdsman Machinery Co., Ltd., Qingdao / China
Tel: +86 15845982569; E-mail: leicahit@qau.edu.cn
Corresponding author: Zhimin Tong
DOI: https://doi.org/10.35633/inmateh-78-113

Keywords: *group-housed pigs, multi-behavior detection, RT-DETR, intelligent livestock monitoring, edge deployment*

ABSTRACT
Accurate detection of multiple behaviors in group-housed pigs was important for precision livestock farming and intelligent farm management. This study proposed RAFE-DETR, an improved detector based on RT-DETR, for recognizing standing, lying, feeding, drinking, and fighting in overhead surveillance images. RFAConv was embedded into RepViT blocks to construct the RFA-RepViT backbone for stronger local feature extraction. The original intra-scale interaction module was replaced with BiFormer to improve contextual modeling. The neck was redesigned with ASF-CSA to enhance adaptive multi-scale fusion, and Focaler-Shape-IoU was introduced to refine box regression. Experiments were conducted on a five-class dataset reconstructed from public surveillance videos. The proposed model achieved 93.9% precision, 92.7% recall, and 94.2% mean average precision at an intersection-over-union threshold of 0.5. Compared with RT-DETR-L, these values increased by 1.4, 2.8, and 3.0 percentage points, respectively. At the same time, the number of parameters decreased from 32.0 M to 21.9 M, and GFLOPs decreased from 103.5 to 77.0. Supplementary experiments on a second public dataset supported the robustness of the method. Deployment on Jetson Orin NX Super reached 13.8 and 19.1 frames per second under PyTorch and TensorRT, respectively, indicating good edge-deployment potential.

摘要
群养猪多行为检测是精准养殖和智能化猪场管理中的重要环节。针对俯视监控图像中的站立、躺卧、采食、饮水和打斗五类行为识别任务，本研究提出了一种基于 RT-DETR 的改进检测模型 RAFE-DETR。通过在 RepViT 模块中嵌入 RFAConv，构建了 RFA-RepViT 主干网络，以增强局部特征提取能力。原始同尺度特征交互模块被替换为 BiFormer，以提升上下文建模能力。Neck 结构被重构为 ASF-CSA，以增强自适应多尺度特征融合，并引入 Focaler-Shape-IoU 以优化边界框回归质量。实验基于由公开监控视频重建的五类行为数据集开展。结果表明，RAFE-DETR 在主数据集上的精确率、召回率和 mAP@0.5 分别达到 93.9%、92.7% 和 94.2%。与 RT-DETR-L 相比，这三项指标分别提高了 1.4、2.8 和 3.0 个百分点。同时，参数量由 32.0 M 降至 21.9 M，GFLOPs 由 103.5 降至 77.0。第二公共数据集上的补充实验进一步验证了该方法的稳定性。部署结果表明，模型在 Jetson Orin NX Super 平台上采用 PyTorch 和 TensorRT 后端时，推理速度分别达到 13.8 FPS 和 19.1 FPS，表明该模型具有较好的边缘部署潜力。

INTRODUCTION
In group-housed pig production, behaviors such as standing, lying, feeding, drinking, and fighting provide important information for health assessment, welfare evaluation, and routine farm management (*Berckmans, 2014*). Accurate and continuous monitoring of these behaviors has therefore become an important component of precision livestock farming. In commercial farms, however, behavior assessment still relies largely on manual observation and offline video review, which are labor-intensive, subjective, and difficult to maintain at scale (*Xu et al., 2025*). Wearable sensing tools have also been explored, but their practical use is still constrained by hardware cost, wearing comfort, and maintenance requirements (*Pandey et al., 2021*). By contrast, camera-based computer vision provides a non-contact and scalable solution, and has become an important technical route for intelligent animal farming (*Wurtz et al., 2019; Bao and Xie, 2022*).

Early studies on pig behavior recognition mainly relied on traditional image processing and machine-learning pipelines. Image segmentation, contour extraction, and handcrafted geometric or morphological descriptors were used to represent posture features, and conventional classifiers were then applied for recognition. *Nasirahmadi et al. (2019)* used image processing and support vector machines to score lying postures in grouped pigs. *Bonneau et al. (2021)* compared convolutional neural networks with segmentation combined with support vector machines for sow posture prediction from video images. *Xu et al. (2022)* further introduced depth images and a convolutional neural network–support vector machine framework for grouped-pig posture scoring. Although these methods improved automatic posture analysis to some extent, their performance remained strongly affected by image quality, foreground extraction, and body contour integrity, which limited their robustness in routine multi-behavior monitoring.

With the development of deep learning, pig behavior analysis gradually shifted toward visual detection frameworks. Two-stage detectors represented by Faster R-CNN usually favor detection accuracy, whereas one-stage detectors represented by YOLO provide higher inference speed (*Ren et al., 2017; Redmon et al., 2016*). Recent reviews have shown that camera-based deep learning has become one of the main technical routes in precision pig farming and pig behavior analysis (*Arulmozhi et al., 2021; Xu et al., 2025*). Compared with traditional methods, these detectors reduce dependence on explicit contour extraction and improve feature learning directly from images. Even so, YOLO-style detectors still rely on non-maximum suppression, which may suppress valid predictions when pigs are distributed closely in the same pen.

End-to-end detectors provide another option for this task. DETR formulates object detection as a set prediction problem and reduces dependence on hand-crafted post-processing. RT-DETR further improves practical efficiency through a hybrid encoder and a query selection strategy, which makes Transformer-based detection more suitable for agricultural vision tasks (*Zhao et al., 2024*). In pig-related applications, *Shi et al. (2024)* developed an improved RT-DETR model for pig behavior detection, and *Guo et al. (2025)* further enhanced RT-DETR for abnormal behavior recognition in group-housed pigs. These studies confirmed the value of RT-DETR in pig-house scenes. However, five-class behavior detection still required further improvement when feature representation, fusion quality, localization accuracy, and model complexity were considered together.

To address this issue, this study proposes RAFE-DETR, an improved RT-DETR framework for detecting standing, lying, feeding, drinking, and fighting in group-housed pigs. Compared with previous RT-DETR-based studies on pig behavior analysis, the specific novelty of RAFE-DETR lies in its coordinated adaptation of RT-DETR to complex pig-house scenes. Specifically, RFA-RepViT is constructed by embedding RFAConv into RepViT for lightweight local representation; BiFormer replaces the original AIFI module to improve intra-scale feature interaction; ASF-CSA is designed to refine adaptive multi-scale fusion with channel–spatial attention; and Focaler-Shape-IoU is integrated to stabilize box regression under posture variation and occlusion. A five-class dataset reconstructed from public surveillance videos released by *Bergamini et al. (2021)* was used for training and evaluation. A second public dataset was added for supplementary validation. Comparative experiments, ablation analysis, and edge deployment tests were conducted to assess the effectiveness and practical feasibility of the proposed method.

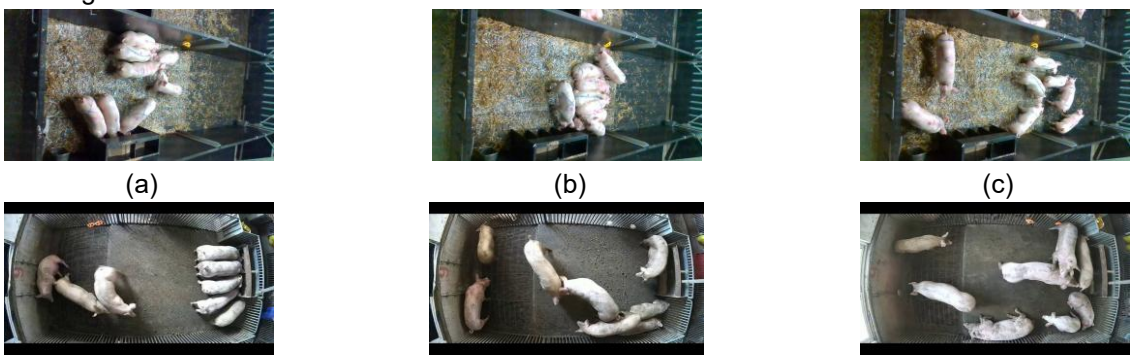
MATERIALS AND METHODS

Public Video Sources and Dataset Construction

This study used publicly available pig-behavior surveillance videos rather than an existing pre-annotated image dataset. Two overhead-view public video sources were used. The primary source was released by *Bergamini et al. (2021)* and recorded pigs, feeders, and drinkers in a group-housed pen. A second public pig-behavior video source reported by *Ji et al. (2023)* was further introduced for supplementary evaluation under related but different pen conditions. Both sources were recorded from overhead viewpoints and were suitable for multi-behavior detection in grouped pigs.

Five behavior categories were defined in this study: standing, lying, feeding, drinking, and fighting. For clips containing fighting behavior, a frame-difference strategy was used to screen candidate frames with obvious motion variation, so that attack-related samples could be extracted more efficiently. For clips without fighting behavior, frames were sampled at 5 s intervals to reduce redundancy among adjacent images and to avoid excessive repetition during training. After frame extraction, all images were manually checked, and blurred, hazy, or highly repetitive frames were removed. As a result, 3200 original images were retained for annotation. Bounding boxes were then labeled in LabelImg, and each box tightly enclosed the corresponding behavior instance.

The dataset was divided into training, validation, and test subsets at a ratio of 8:1:1 at the clip level. Data augmentation was applied only to the training subset. After augmentation, the primary dataset contained 5760 images. Representative raw surveillance frames from the two public video sources, before data augmentation, are shown in Figure 1.



(a) (b) (c)

(d) (e) (f)

Fig. 1 - Representative raw surveillance frames from the two public video sources used in this study:
(a)–(c) primary public source; (d)–(f) second public source

An additional dataset was then constructed from the second public video source using the same behavior categories and annotation protocol. About 600 representative images were selected and manually labeled. The dataset was also divided into training, validation, and test subsets at a ratio of 8:1:1 at the clip level, and data augmentation was applied only to the training subset. After augmentation, the second dataset contained about 1800 images. In this study, the second dataset was used for supplementary near-domain validation rather than strict cross-domain evaluation. Following data augmentation, this study further analyzed the class distribution of the annotated behavior instances. As shown in Table 1, the primary dataset comprises 45,414 annotated behavior instances derived from 5,760 images, while the supplementary dataset contains 14,190 annotated behavior instances derived from 1,800 images.

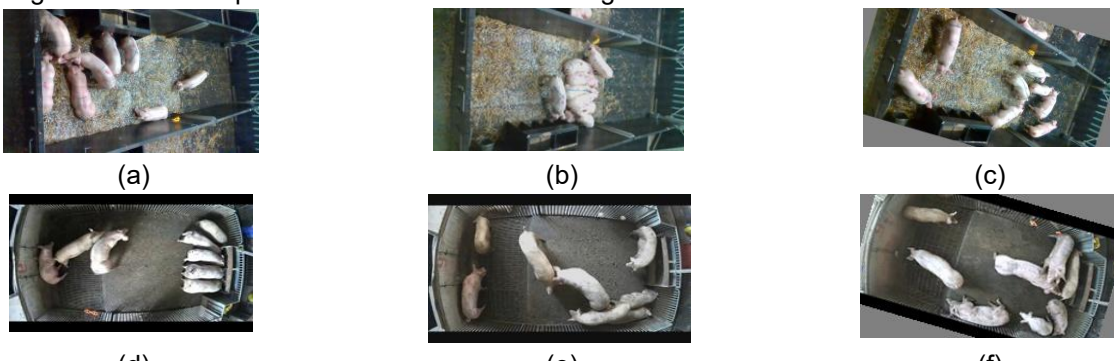
Table 1

Pig behavior annotation data after augmentation

Behavior	Primary dataset	Supplementary dataset
Standing	12,262	4,115
Lying	11,808	3,406
Feeding	9,083	2,554
Drinking	5,450	1,561
Fighting	6,811	2,554

Note: The values indicate labeled behavior instances rather than image numbers. For standing, lying, feeding, and drinking, each visible pig was annotated with one behavior label. For fighting, one interaction box was used to annotate the two pigs involved in the fighting behavior. Data augmentation was applied only to the training subset, while the validation and test subsets were kept unchanged.

Standing and lying accounted for a large proportion of the annotations because these behaviors usually last longer in group-housed pig scenes. Feeding, drinking, and fighting had fewer natural occurrences in continuous surveillance videos, but sufficient samples were retained through targeted frame selection and training-set augmentation. This distribution helped reduce excessive class imbalance while preserving the main behavioral characteristics of group-housed pigs. Examples of the augmentation strategies applied to the training images from the two public datasets are shown in Figure 2.



(a) (b) (c)

(d) (e) (f)

Fig. 2 - Examples of data augmentation applied to the training images from the two public datasets:
(a) and (d) vertical flip; (b) and (e) color jitter; (c) and (f) random rotation

Proposed RAFE-DETR Algorithm

RT-DETR, proposed by Zhao et al. (2024), is an end-to-end object detector built on a hybrid CNN-Transformer framework. It mainly consists of a backbone, a hybrid encoder, a Transformer decoder, and a prediction head. RT-DETR direct application to multi-behavior recognition in group-housed pigs remains limited. In overhead pig-pen images, several pigs often appear in close proximity, and behavior discrimination depends strongly on subtle structural cues, such as head position, trunk posture, and inter-pig contact regions. These cues may be weakened during feature extraction and fusion, which reduces the stability of behavior classification and localization.

To improve five-behavior detection in this setting, an enhanced model named RAFE-DETR was developed on the RT-DETR architecture, and its overall structure is shown in Figure 3. The proposed framework refines the detector at four key stages. First, the original backbone is redesigned as RFA-RepViT to strengthen early feature extraction and preserve fine-grained behavior-related responses. Second, the original AIFI module in the hybrid encoder is replaced with BiFormer to improve selective intra-scale contextual interaction. Third, the neck is reconstructed as ASF-CSA to enhance adaptive multi-scale fusion and strengthen discriminative feature aggregation across adjacent targets. Finally, Focaler-Shape-IoU is introduced as the regression loss to improve bounding-box localization quality. These modifications are not independent replacements, but coordinated adjustments to feature representation, context modeling, feature fusion, and regression supervision. As a result, RAFE-DETR provides a more suitable detection framework for group-housed pig multi-behavior recognition in overhead surveillance images.

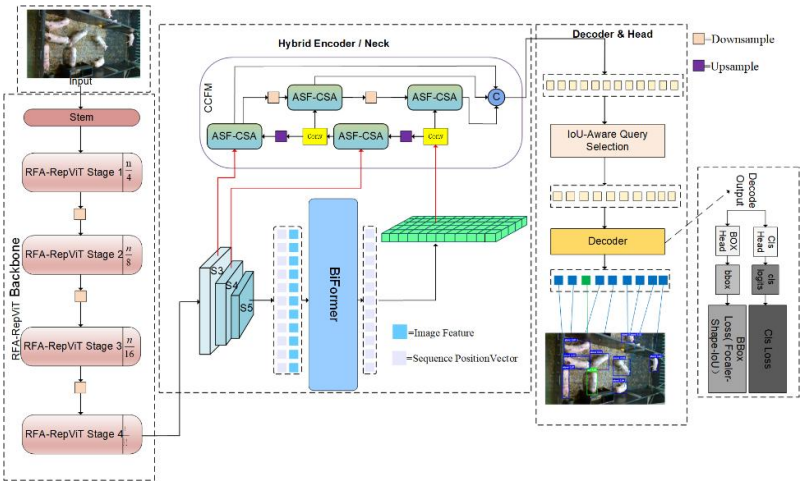


Fig. 3 – Overall architecture of RAFE-DETR

RFA-RepViT Backbone

RepViT is a lightweight backbone that combines efficient convolutional design with transformer-inspired architectural organization, and has shown strong representation capability in visual recognition tasks (*Wang et al., 2024*). In this study, RepViT was not directly adopted as a finished replacement for the original backbone. Instead, it was used as the structural basis for further redesign. RFAConv was embedded into selected RepViT blocks to construct the RFA-RepViT backbone (*Zhang et al., 2026*), as illustrated in Figures 4 and 5. This modification extends the original local convolutional modeling of RepViT by introducing receptive-field attention and multi-branch feature extraction.

For an input feature map, the output of the receptive-field attention fusion can be expressed as:

$$X_{rf} = \sum_{i=1}^{N} \alpha_i B_i(X), \qquad \sum_{i=0}^{N} \alpha_i = 1 \tag{1}$$

where $B_i(X)$ denotes the feature response extracted by the i-th receptive-field branch; α_i denotes the adaptive fusion weight assigned to that branch, and X_{rf} denotes the fused output feature. This design enables the backbone to preserve finer structural responses from head position, trunk posture, and inter-pig contact regions, which are critical for behavior discrimination in overhead pig-pen images. Compared with the original backbone, RFA-RepViT provides a more discriminative early-stage feature representation without causing a substantial increase in model complexity.

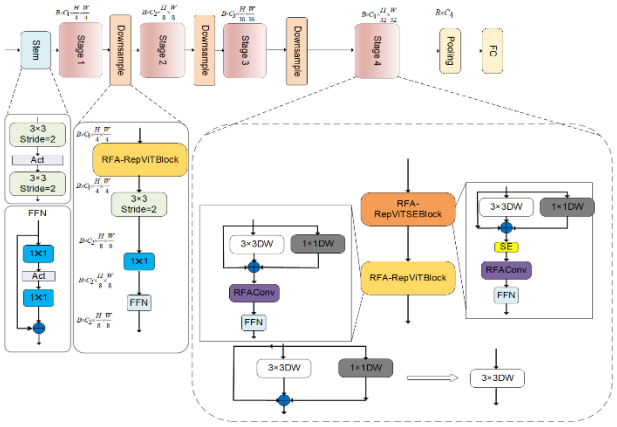


Fig. 4 – Schematic diagram of the RFA-RepViT backbone architecture

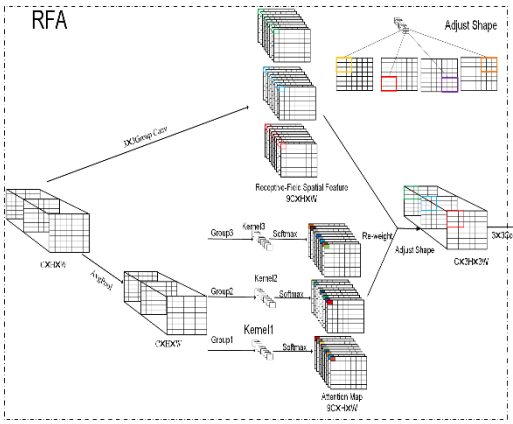


Fig. 5 – RFAConv receptive-field attention module

BiFormer for Intra-Scale Feature Interaction

In the original RT-DETR, the Attention-based Intra-scale Feature Interaction (AIFI) module is used to exchange contextual information within the same scale. In this study, AIFI is replaced with BiFormer in the hybrid encoder (*Zhu et al., 2023*), and the corresponding routing attention mechanism is illustrated in Figure 6. This replacement changes the mechanism of same-scale contextual modeling. Instead of performing uniformly dense interaction over the full feature map, BiFormer introduces Bi-Level Routing Attention, which first identifies relevant coarse regions and then applies fine-grained attention only to the most informative neighboring areas. This design reduces weakly related interactions and makes intra-scale aggregation more selective.

In overhead pig-pen images, the response of one pig is often influenced by adjacent pigs, feeders, drinkers, floor texture, and pen boundaries. This effect becomes stronger when pigs are in close contact or partially overlap. Under such conditions, behavior recognition depends not only on local appearance, but also on context from spatially related regions. The role of BiFormer lies in retaining relevant context while suppressing unrelated responses. Compared with the original AIFI-based interaction, the modified encoder provides more discriminative same-scale features for dense group scenes and offers a more reliable feature basis for the subsequent fusion and prediction stages.

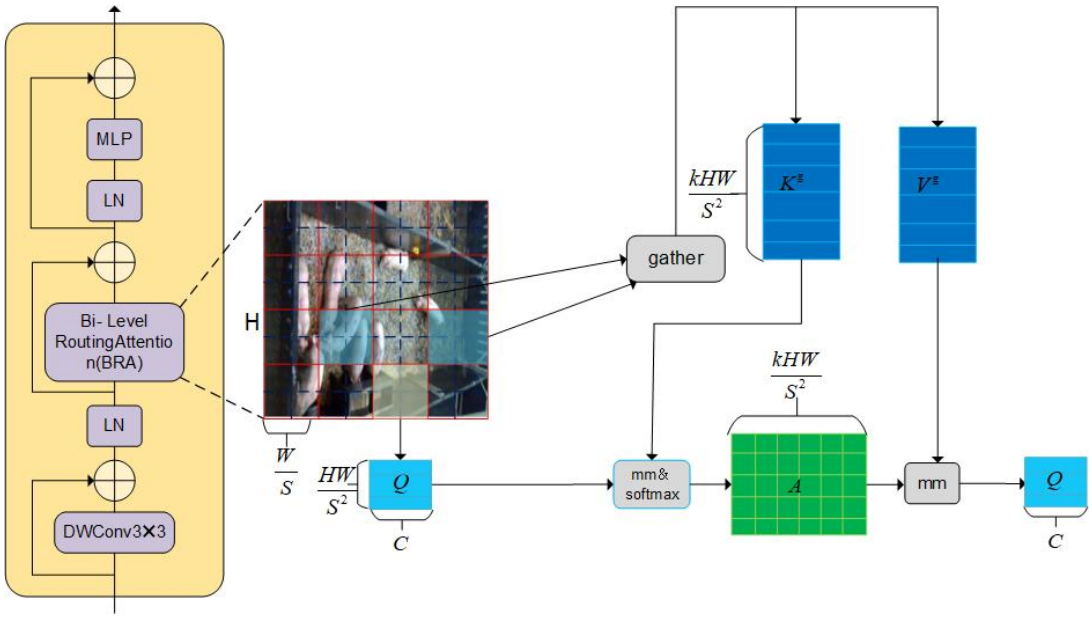


Fig. 6 – Schematic of BRA in BiFormer

ASF-CSA for Cross-Scale Feature Fusion

In the original RT-DETR, the neck is responsible for cross-scale feature transmission, but its fusion process remains relatively fixed when features from different levels are combined. In this study, the original neck was redesigned as ASF-CSA, and the corresponding structures are shown in Figures 7 and 8.

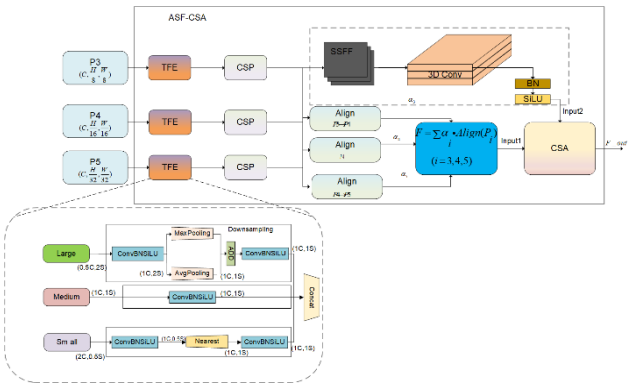


Fig. 7 – Architecture of the proposed ASF-CSA neck

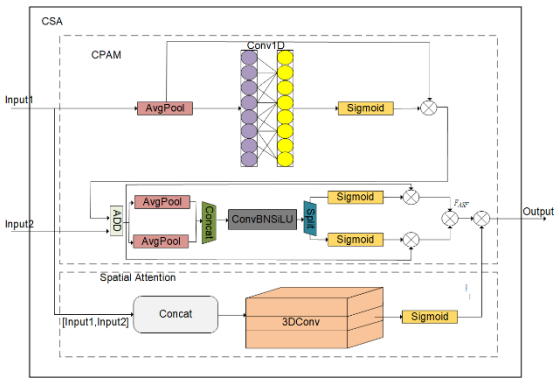


Fig. 8 – CSA module structure

The ASF stage performs adaptive integration of features from different scales. The CSA stage then refines the fused representation through channel and spatial reweighting. The adaptive fusion process in ASF can be written as:

$$F_{asf} = \sum_{l=1}^{L} \omega_l \phi_l(F_l), \qquad \sum_{l=0}^{L} \omega_l = 1 \tag{2}$$

where F_l denotes the input feature at scale level l, ϕ_l denotes the corresponding alignment or transformation operation; ω_l denotes the adaptive fusion weight, and F_{asf} denotes the fused multi-scale feature.

In overhead pig-pen images, behavior cues are often distributed across different feature levels. Whole-body posture is usually reflected in higher-level semantics, whereas head orientation, neck extension, and inter-pig contact regions depend more strongly on lower-level detail. When adjacent pigs are close to each other, these cues may also appear in intertwined body regions. ASF-CSA was designed for this condition. Its role is to improve cross-scale continuity and to strengthen behavior-related responses after fusion, rather than relying on a fixed aggregation pathway. Compared with the original neck, the modified design provides more discriminative multi-scale representations for dense group scenes and supports more stable behavior prediction in the subsequent decoder.

Focaler-Shape-IoU Loss

In the original RT-DETR, bounding-box regression mainly depends on overlap-based supervision, which is effective for general object localization but less stable when target geometry varies markedly. In this study, the regression loss is replaced with Focaler-Shape-IoU, and its principle is illustrated in Figure 9. This modification shifts the emphasis of regression supervision. Focaler-IoU introduces a quality-aware calibration mechanism (*Zhang and Zhang, 2024*), whereas Shape-IoU adds explicit geometric constraints related to box shape and scale (*Zhang and Zhang, 2023*). Their combination makes the regression process responsive not only to overlap quality, but also to the structural consistency of the predicted box.

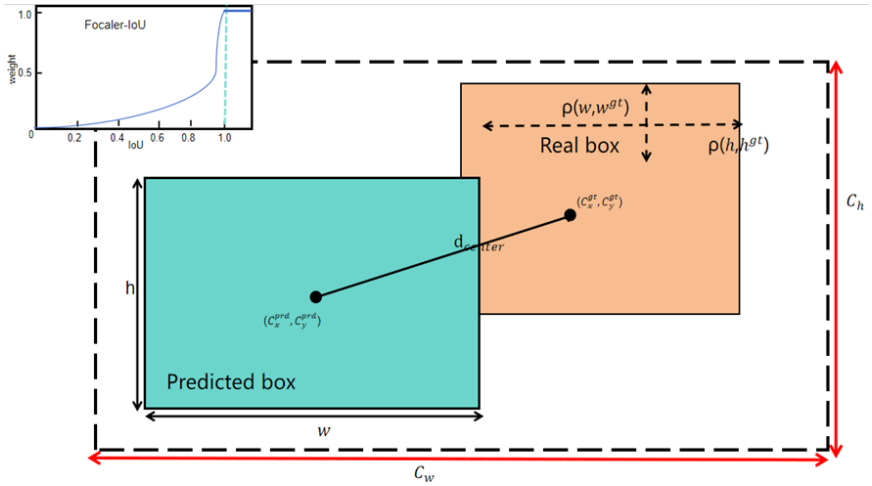


Fig. 9 – Diagram illustrating the Focaler-Shape-IoU principle

The final regression objective is expressed as:

$$L_{Focaler\text{-}Shape\text{-}IoU} = 1 - IoU_f + d_{shape} + \lambda S_{shape} \quad (3)$$

where IoU_f denotes the calibrated overlap term, d_{shape} denotes the distance penalty associated with center and geometry mismatch, S_{shape} denotes the shape-consistency term, and λ is the balancing coefficient. In overhead pig-pen images, the geometry of behavior instances changes frequently with posture variation, body contact, and partial visibility. Under such conditions, a box with acceptable overlap may still show noticeable deviation in center position, width–height proportion, or overall shape. Focaler-Shape-IoU provides more stable regression supervision at the prediction stage and improves localization quality in dense group scenes.

Model Performance Evaluation

During the experiment, all models were trained using the same set of parameters. The input size was fixed at 640 × 640 pixels, and all models were trained for 300 epochs with a batch size of 16 using identical training, validation, and test splits. AdamW was used as the optimizer, with an initial learning rate of 0.001, a momentum of 0.937, and a weight decay of 0.0001. The same training-set augmentation strategy, including vertical flip, color jitter, and random rotation, was applied to all models, while the validation and test sets were kept unchanged. The same learning-rate schedule was used for all models.

To comprehensively evaluate the detection performance of the proposed RAFE-DETR model, Precision (P), Recall (R), average precision at an IoU threshold of 0.5 (AP@0.5), and mean average precision at an IoU threshold of 0.5 (mAP@0.5) were used as accuracy metrics. In addition, parameter count and GFLOPs were reported to evaluate model complexity.

The metrics are defined as follows:

$$P = \frac{TP}{TP + FP} \times 100\% \quad (4)$$

$$R = \frac{TP}{TP + FN} \times 100\% \quad (5)$$

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{n} \times 100\% \quad (6)$$

where: TP is the number of true positives; FP is the number of false positives; FN is the number of false negatives; n is the number of classes.

RESULTS

RT-DETR Baselines and Settings

To identify a suitable baseline for subsequent improvement, five RT-DETR models with different scales were compared under the same training and evaluation conditions. The results are listed in Table 2.

Table 2

Results of different RT-DETR models

Model	Backbone	P%	R%	mAP@0.5/%	Params/M	FLOPs/G
RT-DETR-R18	ResNet-18	90.7	89.1	89.9	19.9	57.2
RT-DETR-R50	ResNet-50	91.4	89.9	89.7	41.9	136
RT-DETR-R101	ResNet-101	92.2	89.0	89.8	76.4	259
RT-DETR-L	HGNetv2-L	92.5	89.9	91.2	32.0	103.5
RT-DETR-X	HGNetv2-X	92.1	89.3	89.4	65.5	234

RT-DETR-R18 showed the lowest complexity, with 19.9 M parameters and 57.2 GFLOPs, but its precision, recall, and mAP@0.5 were only 90.7%, 89.1%, and 89.9%, respectively. RT-DETR-R50 increased precision and recall to 91.4% and 89.9%, but its mAP@0.5 remained at 89.7% and its complexity rose to 41.9 M parameters and 136 GFLOPs. RT-DETR-R101 and RT-DETR-X further increased model size, yet neither produced a clear gain in overall detection performance. By comparison, RT-DETR-L achieved the best balance between detection accuracy and computational cost, reaching 92.5% precision, 89.9% recall, and 91.2% mAP@0.5 with 32.0 M parameters and 103.5 GFLOPs. Therefore, RT-DETR-L was selected as the baseline model for the following experiments.

Backbone Comparison and Validation

To validate the performance of the proposed model, this study conducted comparative experiments within the same RT-DETR framework using five representative backbone networks. The compared backbones were GhostNetV2 (*Tang et al., 2022*), ConvNeXt V2 (*Woo et al., 2023*), MobileNetV4 (*Qin et al., 2024*), Swin Transformer (*Liu et al., 2021*), and RepViT (*Wang et al., 2024*). The results are shown in Table 3.

Table 3

Comparison of different backbone networks

Backbone	P%	R%	mAP@0.5/%	Params/M	FLOPs/G
HGNetv2-L	92.5	89.9	91.2	32.0	103.5
GhostNetV2	91.0	90.1	90.4	21.5	59.4
ConvNeXtV2	93.0	88.8	90.2	21.4	65.4
MobileNetV4	92.8	91.0	89.7	20.4	73.0
RepViT	93.2	90.9	91.7	22.4	75.3
Swin Transformer	92.2	90.6	89.9	45.4	130.5
RFA-RepViT	93.3	91.6	92.7	25.5	77.6

Lower backbone complexity did not consistently translate into better detection performance. GhostNetV2 had the lowest computational cost, with 59.4 GFLOPs and 21.5 M parameters, but its mAP@0.5 was only 90.4%. ConvNeXtV2 and MobileNetV4 likewise showed no clear advantage over HGNetv2-L when accuracy and efficiency were considered together. Swin Transformer further increased model complexity to 45.4 M parameters and 130.5 GFLOPs, whereas its mAP@0.5 remained at 89.9%.

Among the candidate backbones, RepViT achieved the best balance between detection accuracy and computational cost. It reached 93.2% precision, 90.9% recall, and 91.7% mAP@0.5 with only 22.4 M parameters and 75.3 GFLOPs. Compared with HGNetv2-L, RepViT improved mAP@0.5 by 0.5 percentage points while reducing parameters and GFLOPs by 9.6 M and 28.2, respectively. Based on RepViT, RFAConv was further introduced to construct the RFA-RepViT backbone. This modification strengthened local feature extraction with only a limited increase in complexity.

Ablation Experiments

To verify the contribution of each proposed module, ablation experiments were conducted on the RT-DETR baseline model. The tested configurations sequentially introduced RFA-RepViT, BiFormer, ASF-CSA, and Focaler-Shape-IoU into the RT-DETR-L baseline. The comparative results are summarized in Table 4.

Table 4

Ablation experiment results: performance comparison of different models

Baseline Model	A	B	C	D	P/%	R/%	mAP@0.5/%	Params/M	FLOPs/G
					92.5	89.9	91.2	32.0	103.5
	√				93.3	91.6	92.7	25.5	77.6
		√			93.7	91.6	93.5	31.5	103.1
			√		92.0	91.7	93.5	29.7	109.5
RT-DETR	√	√			93.4	91.3	93.0	24.9	77.2
	√		√		93.0	91.9	94.1	22.4	77.4
		√	√		93.7	91.7	93.6	29.2	109.1
	√	√	√		93.8	92.5	93.6	21.9	77
	√	√	√	√	93.9	92.7	94.2	21.9	77

Note: A: RFA-RepViT, B: BiFormer, C: ASF-CSA, D: Focaler-Shape-IoU. "√" indicates that the corresponding module is included

As shown in Table 4, each proposed module improved the detector when introduced individually, but the gain pattern differed across modules. Replacing the backbone with RFA-RepViT increased precision, recall, and mAP@0.5 from 92.5%, 89.9%, and 91.2% to 93.3%, 91.6%, and 92.7%, while parameters and GFLOPs decreased from 32.0 M and 103.5 to 25.5 M and 77.6. This result indicates that the redesigned backbone preserves local behavior cues more effectively and also reduces computational cost. BiFormer produced a different effect.

Precision and recall reached 93.7% and 91.6%, and mAP@0.5 rose to 93.5%, with only a minor change in complexity. This suggests that selective intra-scale interaction is beneficial in crowded scenes. ASF-CSA also increased recall and mAP@0.5 to 91.7% and 93.5%, mainly by strengthening multi-scale feature aggregation, although GFLOPs increased to 109.5. Overall, the three structural modules improved the detector in different ways: RFA-RepViT enhanced local detail representation, BiFormer improved contextual discrimination, and ASF-CSA strengthened feature fusion across adjacent targets.

The RFA-RepViT and ASF-CSA combination achieved 94.1% mAP@0.5 with only 22.4 M parameters and 77.4 GFLOPs, showing the strongest accuracy–complexity balance among the structural combinations. When all three structural modules were used together, precision, recall, and mAP@0.5 reached 93.8%, 92.5%, and 93.6%, respectively, with 21.9 M parameters and 77.0 GFLOPs. Although this configuration did not yield the highest single mAP@0.5, it produced a more balanced improvement across the three detection metrics, suggesting that the modules were partly complementary rather than linearly cumulative.

Focaler-Shape-IoU was then introduced on top of the full structural model. Precision, recall, and mAP@0.5 further increased to 93.9%, 92.7%, and 94.2%, while the number of parameters and GFLOPs remained unchanged. Because this loss only modifies the regression objective, its contribution is better understood as a refinement of localization quality rather than a change in feature representation. Overall, the ablation results indicate that the final model benefits from coordinated improvements in local detail preservation, contextual interaction, multi-scale fusion, and box regression.

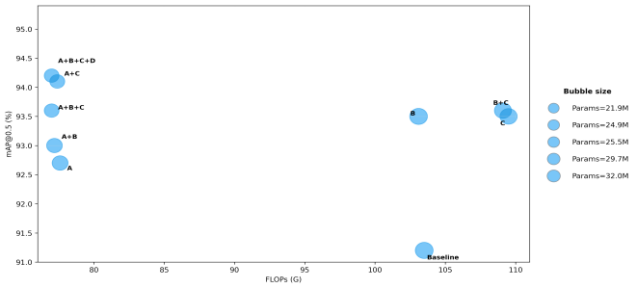


Fig. 10 – Comparison of mAP@0.5 and GFLOPs under different ablation settings

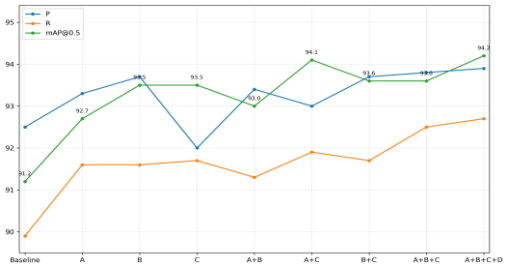


Fig. 11 – Performance trends of different ablation settings in terms of precision, recall, and mAP@0.5

Detection Performance by Behavior Category

To further examine model performance across behavior categories, RT-DETR-L and RAFE-DETR were compared on the five pig behaviors shown in Table 5. Overall, RAFE-DETR achieved 93.9% precision, 92.7% recall, and 94.2% mAP@0.5, indicating stable and accurate recognition of pig behaviors.

Table 5

Experimental results of behavior recognition

Models	AP@0.5/%					P%	R%	mAP@0.5/%
	Standing	Lying	Drinking	Feeding	Fighting			
RT-DETR	92.1	96.1	90.3	91.9	85.5	92.5	89.9	91.2
RAFE-DETR	94.3	97.5	93.5	94.7	90.9	93.9	92.7	94.2

Table 5 shows the category-wise AP@0.5 results. RAFE-DETR outperformed RT-DETR in all five behaviors, with the largest gain observed for fighting, where AP@0.5 increased from 85.5% to 90.9%. Feeding and drinking also improved by 2.8 and 3.2 percentage points, respectively. Overall, RAFE-DETR increased mAP@0.5 from 91.2% to 94.2%, indicating better behavior-level detection performance in group-housed pig scenes.

Comparison with Different Detection Models

To evaluate the proposed method more comprehensively, RAFE-DETR was compared with representative CNN-based, YOLO-based, and Transformer-based detectors. The comparative results are listed in Table 6.

RAFE-DETR achieved the best overall performance among all evaluated models, with 93.9% precision, 92.7% recall, and 94.2% mAP@0.5, while requiring only 21.9 M parameters and 77 GFLOPs. This result indicates that the proposed modifications improve detection accuracy without increasing model burden. Compared with conventional CNN-based detectors, the advantage of RAFE-DETR remained clear. Faster R-CNN and TOOD achieved mAP@0.5 values of 91.6% and 91.3%, respectively, but both models showed much lower recall and substantially higher computational cost.

Among the YOLO-based models, YOLOv12L produced the highest mAP@0.5 at 92.7%, while YOLOv10L reached the highest recall of 91.6%. Even so, none of the YOLO variants surpassed RAFE-DETR in overall performance.

Compared with RT-DETR-L, RAFE-DETR improved precision, recall, and mAP@0.5 by 1.4, 2.8, and 3.0 percentage points, respectively. At the same time, the number of parameters decreased from 32.0 M to 21.9 M, and GFLOPs decreased from 103.5 to 77.0. This confirms the superiority of the proposed method for multi-behavior recognition in group-housed pigs.

Table 6

Comparison results of different target detection models

Models	P/%	R/%	mAP@0.5/%	Params/M	FLOPs/G
Faster-R-CNN	75.6	76.3	91.6	41.4	208
TOOD	70.1	79.0	91.3	32.1	199
YOLOv8L	92.4	88.7	92.4	43.6	164.8
YOLOv10L	92.3	91.6	90.6	24.3	120.0
YOLOv11L	92.5	89.4	91.3	25.3	86.6
YOLOv12L	92.2	89.2	92.7	26.3	88.6
DETR	78.8	79.2	89.8	41.6	96.5
RT-DETR-R34	91.0	91.6	90.4	31.3	89.1
RT-DETR-L	92.5	89.9	91.2	32.0	103.5
RAFE-DETR	93.9	92.7	94.2	21.9	77

Qualitative Visualization

Figure 12 presents qualitative comparisons of YOLOv11L, RT-DETR-L, and RAFE-DETR in representative pig-pen scenes. All three models detected the major pigs and behavior instances. The differences became evident in crowded regions, close-contact areas, and partially visible targets near image boundaries.

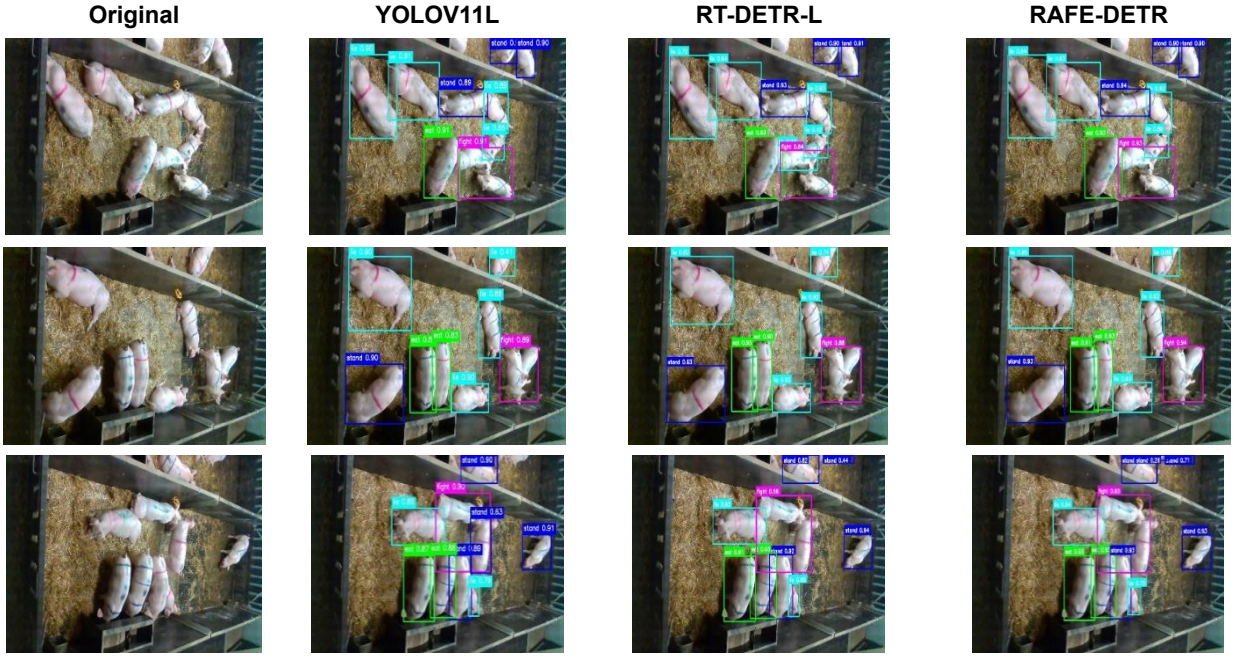


Fig. 12 – Comparison of three detection models for group-housed pig behavior recognition

In these challenging cases, RAFE-DETR produced more stable predictions, clearer behavior assignment, and more consistent confidence scores than the two comparison models. False suppression and category ambiguity were also reduced.

The visual patterns were consistent with the functional roles of the proposed modules. RFA-RepViT improved the preservation of local details in small or partially visible pigs. BiFormer enhanced selective contextual interaction in dense scenes. ASF-CSA strengthened feature aggregation for adjacent targets under close contact. Focaler-Shape-IoU further improved localization stability when posture variation and spatial overlap occurred simultaneously. Even so, severely truncated boundary targets remained difficult for all compared detectors. Overall, the qualitative results were in good agreement with the quantitative evaluation and supported the advantage of the proposed method in complex surveillance images.

Additional Evaluation on a Second Public Dataset

To further examine robustness beyond the primary source, supplementary experiments were conducted on a second public pig-behavior dataset. In the validation design of this study, the primary dataset served as the same-source evaluation set, whereas the second dataset was used for near-domain supplementary evaluation under related but different local scene conditions. The corresponding results are listed in Table 7.

Table 7

Comparison of different models on the second public pig-behavior dataset

Models	P/%	R/%	mAP@0.5/%
YOLOv8L	89.6	82.7	87.4
YOLOv11L	90.4	88.2	88.2
YOLOv12L	91.4	83.4	86.3
RT-DETR-L	92.4	87.5	88.7
RAFE-DETR	91.4	89.9	90.4

RAFE-DETR achieved 91.4% precision, 89.9% recall, and 90.4% mAP@0.5 on this dataset, ranking first in recall and mAP@0.5 among all compared models. Relative to RT-DETR-L, precision decreased slightly by 1.0 percentage point, whereas recall and mAP@0.5 increased by 2.4 and 1.7 percentage points, respectively. The proposed model also remained superior to the YOLO-based detectors in overall detection performance, with mAP@0.5 gains of 3.0 and 4.1 percentage points over YOLOv11L and YOLOv12L, respectively. These results suggest that RAFE-DETR retained stable effectiveness under related scene variation and was not restricted to a single public source.

Edge Deployment Validation

To verify practical deployment feasibility, on-device validation was conducted on a Jetson Orin NX Super platform. The model was successfully executed under both PyTorch and TensorRT backends. The complete workflow of model loading, inference, and result output was maintained on the edge device. The deployment results are summarized in Table 8.

Table 8

Deployment feasibility and accelerated inference validation on an edge platform.

Model	Inference backend	Input size	FPS	Improvement over RT-DETR (%)
RT-DETR	PyTorch	640×640	9.1	
RAFE-DETR	PyTorch	640×640	13.8	51.65
RAFE-DETR	TensorRT	640×640	19.1	109.89

Under the PyTorch backend, RT-DETR reached 9.1 FPS, whereas RAFE-DETR achieved 13.8 FPS. After TensorRT acceleration, the inference speed of RAFE-DETR further increased to 19.1 FPS. Relative to RT-DETR under PyTorch, the corresponding speed gains were 51.65% and 109.89%, respectively. Detection boxes and class labels were generated normally after deployment, and the saved visual results are shown in Figure 13. Taken together, these results confirm that the proposed detector is not only effective in offline evaluation but also executable on embedded hardware for edge-side pig behavior monitoring.

Fig. 13 – Deployment and inference verification of RAFE-DETR on the NVIDIA Jetson Orin NX Super platform

CONCLUSIONS

Accurate recognition of standing, lying, feeding, drinking, and fighting in group-housed pigs is important for intelligent livestock monitoring and precision farm management. To address the limitations of the original RT-DETR in multi-behavior detection under overhead pig-pen surveillance, this study developed RAFE-DETR by improving the backbone, intra-scale interaction, multi-scale fusion, and box regression.

The following conclusions can be drawn:

(1) In the proposed framework, RFAConv was embedded into RepViT blocks to construct the RFA-RepViT backbone, BiFormer was introduced to strengthen selective intra-scale interaction, ASF-CSA was designed to improve adaptive multi-scale feature fusion, and Focaler-Shape-IoU was adopted to refine bounding-box regression. These coordinated modifications improved local feature preservation, contextual discrimination, feature aggregation, and localization stability in dense pig-pen scenes.

(2) Experimental results on the primary dataset showed that RAFE-DETR achieved 93.9% precision, 92.7% recall, and 94.2% mAP@0.5. Compared with RT-DETR-L, precision, recall, and mAP@0.5 increased by 1.4, 2.8, and 3.0 percentage points, respectively, while parameters decreased from 32.0 M to 21.9 M and GFLOPs decreased from 103.5 to 77.0. These results indicate that the proposed model improved detection accuracy while maintaining a more favorable balance between performance and computational cost.

(3) Supplementary experiments on a second public pig-behavior dataset further supported the robustness of the proposed method under related scene variation. In addition, deployment tests on the Jetson Orin NX Super platform showed that the model remained executable on embedded hardware, reaching 13.8 FPS under PyTorch and 19.1 FPS under TensorRT. These results demonstrate that RAFE-DETR has practical potential for intelligent livestock monitoring and edge-side behavior analysis in group-housed pig production.

In summary, RAFE-DETR provided a practical combination of detection accuracy, computational efficiency, and deployment feasibility for multi-behavior recognition in group-housed pigs.

ACKNOWLEDGMENTS

This work was supported by the Special Fund for Leading Talent in Mount Tai Industry of Shandong Province (TSCX202507039).

REFERENCES

[1] Arulmozhi E., Bhujel A., Moon B.-E., Kim H.-T., (2021). The application of cameras in precision pig farming: an overview for swine-keeping professionals. *Animals*, vol.11, article 2343. DOI: 10.3390/ani11082343

[2] Bao J., Xie Q., (2022). Artificial intelligence in animal farming: a systematic literature review. *Journal of Cleaner Production*, vol.331, article 129956. DOI: 10.1016/j.jclepro.2021.129956

[3] Berckmans D., (2014). Precision livestock farming technologies for welfare management in intensive livestock systems. *Revue Scientifique et Technique de l'Office International des Epizooties*, vol.33, pp.189-196. DOI: 10.20506/rst.33.1.2273

[4] Bergamini L., Pini S., Simoni A., Vezzani R., Calderara S., D'Eath R.B., Fisher R.B., (2021). Extracting accurate long-term behavior changes from a large pig dataset. *Proceedings of the 16th International Conference on Computer Vision Theory and Applications (VISAPP),* Vienna/Austria, pp.524-533.

[5] Bonneau M., Benet B., Labrune Y., Bailly J., Ricard E., Canario L., (2021). Predicting sow postures from video images: comparison of convolutional neural networks and segmentation combined with support vector machines under various training and testing setups. *Biosystems Engineering*, vol.212, pp.19-29. DOI: 10.1016/j.biosystemseng.2021.09.014

[6] Guo Y., Wang X., Liu J., Peng B., Ran X., Mao R., (2025). Enhancing abnormal behavior recognition in group-housed pigs through the MS-FA-DETR: a multi-scale and frequency-aware fusion method. *Proceedings of the 2025 6th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, Ningbo/China. DOI: 10.1109/CVIDL65390.2025.11085581

[7] Ji H., Teng G., Yu J., Wen Y., Deng H., Zhuang Y., (2023). Efficient aggressive behavior recognition of pigs based on temporal shift module. *Animals*, vol.13, article 2078. DOI: 10.3390/ani13132078

[8] Liu Z., Lin Y., Cao Y., Hu H., Wei Y., Zhang Z., Lin S., Guo B., (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC/Canada, pp.10012-10022.

[9] Nasirahmadi A., Sturm B., Olsson A.-C., Jeppsson K.-H., Müller S., Edwards S., Hensel O., (2019). Automatic scoring of lateral and sternal lying posture in grouped pigs using image processing and support vector machine. *Computers and Electronics in Agriculture*, vol.156, pp.475-481. DOI: 10.1016/j.compag.2018.12.009

[10] Pandey S., Kalwa U., Kong T., Guo B., Gauger P.C., Peters D.J., Yoon K.-J., (2021). Behavioral monitoring tool for pig farmers: ear tag sensors, machine intelligence, and technology adoption roadmap. *Animals*, vol.11, article 2665. DOI: 10.3390/ani11092665

[11] Qin D., Leichner C., Delakis M., Fornoni M., Luo S., Yang F., Wang W., Banbury C., Ye C., Akin B., (2024). MobileNetV4: Universal models for the mobile ecosystem. *Proceedings of the European Conference on Computer Vision (ECCV)*, Milan/Italy, pp.78-96.

[12] Redmon J., Divvala S., Girshick R., Farhadi A., (2016). You only look once: unified, real-time object detection. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV/USA, pp.779-788.

[13] Ren S., He K., Girshick R., Sun J., (2017). Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.39, no.6, pp.1137-1149. DOI: 10.1109/TPAMI.2016.2577031

[14] Shi L., Ying H., Yang P., Yang C., (2024). Pig behavior detection model based on improved RT-DETR. *Proceedings of the 2024 6th International Conference on Electronic Engineering and Informatics (EEI),* Chongqing/China. DOI: 10.1109/EEI63073.2024.10696250

[15] Tang Y., Han K., Guo J., Xu C., Xu C., Wang Y., (2022). GhostNetV2: Enhance cheap operation with long-range attention. *Advances in Neural Information Processing Systems*, vol.35, pp.9969-9982.

[16] Wang A., Chen H., Lin Z., Han J., Ding G., (2024). RepViT: revisiting mobile CNN from ViT perspective. *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* Seattle, WA/USA, pp.15909-15920.

[17] Woo S., Debnath S., Hu R., Chen X., Liu Z., Kweon I.S., Xie S., (2023). ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC/Canada, pp.16133-16142.

[18] Wurtz K., Camerlink I., D'Eath R.B., Fernández A.P., Norton T., Steibel J., Siegford J., (2019). Recording behaviour of indoor-housed farm animals automatically using machine vision technology: a systematic review. *PLOS ONE*, vol.14, no.12, article e0226669. DOI: 10.1371/journal.pone.0226669

[19] Xu J., Zhou S., Xu A., Ye J., Zhao A., (2022). Automatic scoring of postures in grouped pigs using depth image and CNN-SVM. *Computers and Electronics in Agriculture*, vol.194, article 106746. DOI: 10.1016/j.compag.2022.106746

[20] Xu J., Ying Y., Wu D., Hu Y., Cui D., (2025). Recent advances in pig behavior detection based on information perception technology. *Computers and Electronics in Agriculture*, vol.235, article 110327. DOI: 10.1016/j.compag.2025.110327

[21] Zhang H., Zhang S., (2023). Shape-IoU: more accurate metric considering bounding box shape and scale. arXiv, *arXiv:2312.17663*. DOI: 10.48550/arXiv.2312.17663

[22] Zhang H., Zhang S., (2024). Focaler-IoU: more focused intersection over union loss. arXiv, arXiv:2401.10525. DOI: 10.48550/arXiv.2401.10525

[23] Zhang X., Liu C., Yang D., Song T., Ye Y., Li K., (2026). RFAConv: receptive-field attention convolution for improving convolutional neural networks. *Pattern Recognition*, vol.176, article 113208. DOI: 10.1016/j.patcog.2026.113208

[24] Zhao Y., Lv W., Xu S., Wei J., Wang G., Dang Q., Liu Y., Chen J., (2024). DETRs beat YOLOs on real-time object detection. *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA/USA, pp.16965-16974.

[25] Zhu L., Wang X., Ke Z., Zhang W., Lau R.W., (2023). BiFormer: vision transformer with bi-level routing attention. *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC/Canada, pp.10323-10333.