Smart Agricultural Technology 12 (2025) 101414

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

CD-ViT-YOLO: A lightweight Hybrid ViT-YOLO model for caged duck behaviour recognition under varying lighting conditions

Yujin Gong [a,b], Gen Zhang [a,b], Chuntao Wang [a,b,c,*], Deqin Xiao [a,b,c]

[a] Collage of Mathematics and Informatics, South China Agricultural University, 510642, Guangdong, People's Republic of China
[b] Key Laboratory of Smart Agricultural Technology in Tropical South China, Ministry of Agriculture and Rural Affairs, Guangzhou, 510642, People's Republic of China
[c] Guangdong Engineering Research Center of Agricultural Big Data, Guangzhou, 510642, People's Republic of China

ARTICLE INFO

Keywords:
Caged duck farming
Behaviour recognition
Lightweight
YOLO
ViT
Dynamic feature extraction

ABSTRACT

The daily behaviours of caged ducks directly reflect their physiological health, making behaviour monitoring crucial for welfare assessment. Currently, many behaviour monitoring methods rely on manual observation, which is labour-intensive and inefficient for large-scale production. In addition, only a limited number of studies have explored intelligent behaviour monitoring, and existing approaches still require further accuracy improvement. In particular, most current models lack robustness under real farm lighting variations and struggle to reliably detect occluded targets within densely grouped animals. To tackle these problems, this study proposes a lightweight hybrid ViT-YOLO model for caged duck behaviour recognition under varying lighting conditions, which is coined as CD-ViT-YOLO,. Specifically, the YOLOv5s is selected as the baseline model for its high stability and lightweight structure. To fuse effectively features of densely packed and partially occluded ducks in cage farming environments, a novel Dynamic *Re*-parameterised Enhancement (DRE) block is designed to replace the C3 block of YOLOv5s. Additionally, to better adapt to irregular duck shapes and improve inference efficiency, deformable convolution and RepConv are integrated. Moreover, to reduce computational cost while enhancing global perception capability, a stage-tailored EfficientViT-m0 is exploited to substitute the backbone of the improved YOLOv5s. Besides, to improve the model's ability to identify challenging samples, a difficulty-aware detection head is also incorporated. Combining these optimization strategies gives rise to the proposed CD-ViT-YOLO. Experimental simulation on the self-built dataset DuckBehaviour shows that the proposed CD-ViT-YOLO achieves 97.4 % mAP@0.5 and 88.6 % mAP@0.5:0.95. These are 1.1 and 2.3 percentage points higher than those of YOLOv5s, respectively, at 32 % reduced parameters and 45 % decreased computational complexity, which outperforms the state of the art. This indicates that the proposed CD-ViT-YOLO offers strong support for intelligent and precise management in caged duck farming.

1. Introduction

As a major agricultural nation, China has consistently prioritised the sustainable development of agriculture. Livestock farming, a crucial component of agriculture, has been rapidly advancing towards automation and intelligent systems in recent years. Throughout the livestock farming process, the health status of the animals remains a primary concern. Since there is a strong correlation between animals' daily behaviours and their health conditions, real-time monitoring of animal behaviours is essential for the early detection of abnormalities, enabling timely interventions to ensure animal health.

Historically, animal behaviour monitoring relies heavily on manual observation and recording [6,31]. However, this approach is time-consuming and labour-intensive, resulting in low monitoring efficiency. Moreover, manual monitoring is inadequate for large-scale animal populations, generally leading to missed or incorrect animal behaviour recognition, which fails to meet the demands of modern large-scale farming. Consequently, intelligent perception and automatic recognition of animal behaviours has emerged as evident challenges in the field of smart farming.

Over the past decade, extensive researches have been conducted on animal behaviour recognition. Sensing technologies facilitated to animal behaviour recognition can be broadly categorised into two types based on whether the sensors are contact-based: contact sensor-based recognition methods [16] and non-contact computer vision-based recognition

* Corresponding author.
E-mail address: wangct@scau.edu.cn (C. Wang).

https://doi.org/10.1016/j.atech.2025.101414
Received 6 June 2025; Received in revised form 2 September 2025; Accepted 2 September 2025
Available online 4 September 2025

Nomenclature

Abbreviations	Full name
AP	Average Precision
CD-ViT-YOLO	Cage Duck-ViT-YOLO
C3	CSPDarkNet53
CSP	Cross Stage Partial
DAD	Difficulty-Aware Detection
DCNv4	Deformable Convolutional Networks v4
DRE	Dynamic *Re*-parameterised Enhancement
DWconv	Depthwise Convolution
FFN	Feed Forward Network
FN	False Negative
FP	False Positive
F1	F1 score
GFLOPs	Giga Floating Point Operations per Second
IoU	Intersection over Union
IRRep-Bottleneck	Inverted Residual *Re*-parameterisable Bottleneck
mAP	mean Average Precision
P	Precision
R	Recall
TP	Ture Positive
ViT	Vision Transformer
YOLO	You Only Look Once

methods. The contact-based sensor technology is an Internet of Things-based technique for monitoring livestock and poultry behaviour that relies on biological data acquisition [35]. This technology involves attaching sensor devices (such as RFID tags, accelerometers, and heart rate sensors) to individual animals to collect behavioural data. These sensors capture motion characteristics, physiological data, and other information, which are then transmitted wirelessly to a data processing centre. This method demonstrates high accuracy and real-time performance in tracking individual animal behaviour and monitoring health status [10,28,40]. For instance, Brown-Brandl et al. [3] used RFID to monitor changes in pig's weight and analyse their feeding behaviour, Barker et al. [2] classified dairy cow's exit and milking behaviours using location data recorded by sensors, Zehner et al. [60] collected dietary data from cows through sensors to predict calving times. However, the contact-based sensor technology has notable drawbacks. First, the wearables may interfere with the natural behaviour of the animals. For example, wearing a sensor may induce stress reactions in animals, thereby affecting the authenticity of the behavioural data. Second, the maintenance cost of wearable devices is relatively high, especially in large-scale farms, in which regular battery replacements or repairs are necessary. Additionally, the installation process is complex and requires highly skilled personnel to operate the equipment. Thus, although the contact-based sensor technology is applied in certain high-value farms [39], its widespread adoption remains limited.

In recent years, with the rapid development of deep learning and computer vision technologies, non-contact behaviour monitoring has become the mainstream approach in livestock and poultry behaviour research. This type of approach uses cameras to capture video or image data and then recognise and evaluate the animal's behaviours through image processing [1,51,65]. Compared with manual observation and contact-based technologies, the non-contact technology offers several advantages. This technology is non-intrusive, as it captures behavioural data remotely without interfering with the animal's activities, providing a more accurate reflection of their natural behaviours. Additionally, it is highly efficient, as computer vision can rapidly process and analyse large data volumes in real-time, making it ideal for large-scale farming applications. Furthermore, the non-contact technology enables multi-dimensional data collection and processing, as multi-modal fusion allows for the simultaneous monitoring of various behavioural parameters such as trajectories and posture changes. For instance, Chen et al. [7] used an LSTM framework and employed a Softmax prediction function to identify pig feeding behaviours within pens, achieving an accuracy of 98.4 %. Similarly, Yin et al. [59] proposed an efficient feature extraction method combining a bidirectional fusion pyramid with EfficientNet, using a bidirectional long short-term memory module to recognise basic behaviours in cows, obtaining an accuracy of 97.8 %. Later, Wu et al. [53] fused convolutional neural networks with long short-term memory networks to identify behaviours such as drinking, ruminating, walking, standing, and lying, achieving a detection accuracy exceeding 95.8 %. Hao et al. [15] developed a method for detecting laying hen's feeding behaviour based on an improved Faster R-CNN, enhancing the model's accuracy, recall, and F1 score from 84.40 %, 72.67 %, and 78.1 % to 90.12 %, 79.14 %, and 84.3 %, respectively. Fang et al. [11] utilised deep neural networks to construct posture skeletons for broilers and combined them with a naive Bayes model to classify behaviours, achieving over 90 % accuracy in testing various behaviours. Despite the high accuracy of these two-stage recognition methods, their efficiency is generally insufficient for real-time applications in certain farming tasks.

To meet the requirements of real-time performance, many researchers have adopted the You Only Look Once(YOLO) single-stage object detection and recognition method for animal behaviour recognition tasks [57,61]. YOLO's real-time processing capabilities allow for efficient animal detection and behaviour recognition, making it well-suited for continuous monitoring in large-scale animal husbandry. YOLO's inference speed and accuracy enable the effective analysis of animal behaviour without direct interference with animals, thereby preserving their natural behaviour. Additionally, YOLO's ability to handle multiple objects and classify them simultaneously makes it highly effective for monitoring behaviour of multiple animals [5]. For instance, Subedi et al. [38] developed YOLOv5s-pecking and YOLOv5x-pecking models based on YOLOv5s to detect feather pecking behaviour in broilers. The two models achieved the precision of 85.2 % and 88.3 %, respectively, with YOLOv5s-pecking performing the best, providing an effective solution for real-time automated monitoring of feather pecking behaviour and feather damage in chickens. Cho et al. [9] combined YOLOv5 with the DeepSORT algorithm to propose a real-time tracking method for black cattle, achieving an mAP@0.5 of 99.5 % and tracking accuracy of 99.4 %, thus realizing efficient tracking of black cattle. Zheng & Qin [64] introduced an efficient cow behaviour recognition model named PrunedYOLO-Tracker, which reduced the model size of YOLOv5l through a channel pruning algorithm and constructed a cascade buffer Intersection over Union (IoU)to expand the detection and trajectory matching space. By combining behaviour information with trajectory data, the model achieved multi-cow behaviour recognition and tracking. This method shows excellent performance in higher-order tracking accuracy and multi-object tracking precision, reaching 72.4 % and 86.1 %, respectively. Tran & Thanh [42] used YOLOv7 and an improved DeepSORT algorithm to detect and track pig behaviour, establishing behavioural patterns for healthy pigs at different times of the day and effectively identifying both normal and abnormal pig behaviours. Tu et al. [43] designed the YOLOv5-Byte model for multi-object tracking of pig behaviour, achieving a tracking accuracy of 80 %, effectively addressing the problem of missed detection in complex environments.

Prior studies have developed effective approaches for recognising behaviour in both large domestic animals (e.g., pigs and cows) and small poultry (e.g., broilers) [32,50]. However, studies on the behaviour recognition of caged waterfowl, particularly caged ducks, remain relatively underexplored. Furthermore, although non-contact computer

vision technology has been widely applied in monitoring broiler and laying hen behaviour, their application in waterfowl is still limited. For instance, Merenda et al. [30] developed a YOLOv5-based framework for detecting and classifying individual behaviours such as feeding and drinking in broilers. Paneru et al. [33] proposed optimised YOLOv8 models to accurately monitor dustbathing behaviour across different growth stages in cage-free laying hens. Jensen et al. [19] utilised pre-trained CNNs combined with temporal models to detect piling behaviour in laying hens. Recent research has increasingly focused on non-intrusive techniques for poultry behaviour analysis. However, recognising the complex and often overlapping behaviour of cage-raised ducks remains challenging, and related investigations are still relatively limited.

With the continuous expansion of large-scale duck farming, intelligent monitoring of duck behaviour has become increasingly necessary. Behaviour of caged ducks is influenced by environmental factors, particularly in confined spaces and under inadequate lighting conditions. Such environments generally induce stress responses, such as fright, wing spreading, or preening behaviours, which may lead to health issues and reduced productivity. Therefore, real-time behaviour recognition and response to these behaviours are essential for effective monitoring in duck farming. By recognising behavioural patterns intelligently, farms can collect practical data to inform better management decisions, helping to improve living conditions and reduce the frequency of abnormal activity among duck populations.

Recognising behaviour of caged ducks through intelligent systems poses considerable difficulties due to factors like shading, limited lighting, and spatial constraints. These environmental limitations may contribute to abnormal actions such as reduced feeding, trampling, and aggression. Research on the intelligent behaviour recognition of caged ducks is scarce. For example, Gu et al. [12] utilised the YOLOv5 model to detect and recognise specific behaviours of caged ducks, such as stepping, spreading, and preening. While intelligent behaviour recognition using the YOLOv5 model has shown promise, there is a need for further optimisation to balance complexity and accuracy. In light of the increasing integration of intelligent technologies into livestock farming, improving the management of caged duck production requires a better understanding of behavioural patterns. Addressing this need, the present work introduces a refined behaviour detection approach tailored for caged ducks. To address this need, the present work proposes a lightweight hybrid Vision Transformer (ViT)-YOLO model for caged duck behaviour recognition under varying lighting conditions, which is coined as CD-ViT-YOLO.

The main contributions and highlights of this study are as follows:

(1) Dataset construction: A new dataset, *DuckBehaviour*, was constructed. It contains 5700 images of caged ducks with six representative behaviours of drinking, lying, standing, preening, spreading, and eating, providing a reliable benchmark for behaviour recognition research.
(2) Model design: A lightweight and effective caged duck's behaviour recognition framework, CD-ViT-YOLO, was developed by integrating the tailored EfficientViT-m0 backbone, the designed Dynamic *Re*-parameterised Enhancement (DRE) block, and the Difficulty-Aware Detection (DAD) head. These improved strategies together enhance global feature extraction, boost detection of irregular targets, and address sample imbalance.
(3) Performance validation: Extensive experiments on DuckBehaviour and cross-dataset tests demonstrated that the proposed CD-ViT-YOLO achieves a favourable balance between accuracy and complexity. Specifically, compared with the baseline model of YOLOv5s, the CD-ViT-YOLO reduces parameters by nearly 50 % while achieving a 2.3 % improvement in mAP@0.5:0.95; and it outperforms advanced models such as YOLOv12s. These confirm its effectiveness and practicality for intelligent livestock farming applications.

2. Materials and methods

2.1. Data acquisition and processing

In this study, the daily behavioural data of caged ducks were collected by a video camera. The data collection site was a self-built caged duck house at the South China Agricultural University Observatory of Animal Nutrition and Feed Science (Guangzhou, China), and the external environment of the duck house is shown in Fig. 1(a). Inside the duck house, there were four rows of duck cages, and the structure of each row of cages is shown in Fig. 1(b). Each row of duck cages was divided into three tiers, and each tier consisted of ten groups of duck cages with two cages in each group. The dimensions of each duck cage were 700 mm × 700 mm × 550 mm ($L \times W \times H$) as shown in Fig. 1(c). Three ducks were housed in each cage.

For data acquisition, the camera was placed one metre away from the duck cages on one side of a row of duck cages so that the camera mainly targeted the ducks in the two adjacent cages (see Fig. 1(b)) in the second tier of each row of cages. The recording resolution was 1920 × 1080, the frame rate was set to 30 fps, and the recording time was from 12:00 pm to 9:00 pm. Using these settings, videos of the daily behaviour of caged ducks in different light conditions over a sustained period of time were obtained for 90-day-old caged ducks filmed for a continuous period of two days.

After capturing videos of the daily behaviours of caged ducks, all videos were extracted at a rate of 1 frame/second from the corresponding frames to form 32,400 images of caged ducks. For the obtained images, the blurred and highly similar images were further removed to yield 5700 caged duck images. One of the caged duck images is shown in Fig. 1(d), from which it can be seen that a single image contains a variety of daily behaviours of caged ducks, such as lying down and drinking. Other images are similar.

Based on the relationship between the behaviour and health of broiler ducks, six typical daily behaviours were considered in this study: drinking, lying, standing, preening, spreading, and eating. The specific definitions of these behaviours are presented in Table 1. Among them, lying and standing are the most common daily activities, drinking and eating are essential for maintaining duck health. Preening, however, may lead to feather damage, potentially causing animal welfare concerns. Additionally, in limited spaces, spreading may lead to ducks interfering with each other, which in turn may bring about fights, resulting in damage to feathers and other problems detrimental to the health of the ducks.

The LabelImg labelling tool and the bounding box annotation method were used to label the duck behaviours in each image. For each meat duck in the image, if its behaviour belongs to the meat duck behaviours listed in Table 1, each duck is marked with a box and given a corresponding behaviour label. Otherwise, it is not annotated and no behaviour label is given. After the images were labelled, the labelling information regarding coordinates of labelled boxes and the label ID corresponding to each box was saved in two separate *.txt* files. The resulting dataset is named DuckBehaviour.

Enhancing the credibility of the training strategy involved adopting a five-fold cross-validation approach for partitioning and evaluating the dataset DuckBehaviour. Specifically, the entire dataset was evenly divided into five mutually exclusive subsets of equal or nearly equal size. In each iteration, four subsets were combined to form the training set, from which 10 % of the samples were randomly selected as the validation set, and the remaining 90 % served as the training subset. The model was trained on the training subset, and its performance was monitored on the validation set to select the best-performing model (e.g., the one with the lowest loss or highest evaluation metric). This selected model was then evaluated on the remaining fifth subset, which served as the test set. The entire process was repeated five times, with a different subset used as the test set in each iteration. The final generalisation error was estimated by averaging the five test results. This

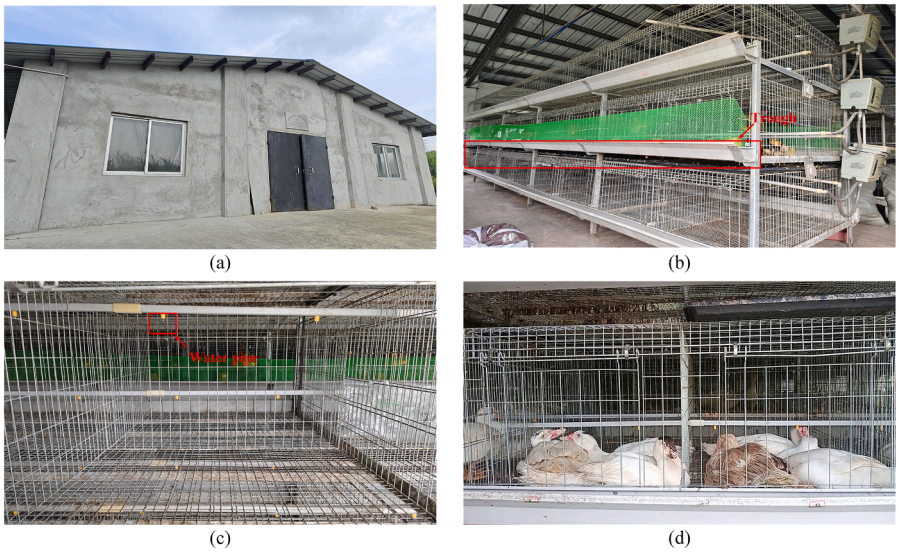

Fig. 1. Schematic diagram of the data acquisition scenario and setup for this study (a)External environment of duck house, (b) Structural diagram of duck cages, (c) Environment of individual duck cages, (d) Illustration of an acquired caged duck image.

Table 1
Definitions of six typical duck behaviours.

Type of behaviour	Define	Labels
Drinking	The body is close to the water, with the bill near or submerged in the water [62].	drinking
Lying	The body is relaxed, with feet retracted, in a resting or sleeping position [54].	lying
Standing	The body is erect with the feet flat on the ground. The neck is usually erect or arched, with wings close to the body. Legs are upright [62].	standing
Eating	The duck inserts its head into the trough to obtain food [54].	eating
Preening	The beak is near the breast feathers or wings, with the neck curved. The legs are either slightly bent or erect, and the wings are slightly spread [62].	preening
Spreading	The duck stands up, tilts its head, stretches its neck, and spreads its wings [62].	spreading

combined use of cross-validation and internal validation ensures a robust assessment of the model's generalisation ability and mitigates the bias introduced by a single arbitrary data split. Furthermore, data augmentation techniques were applied online to the training set to increase data diversity. Specifically, random horizontal flipping, rotation (±45 degrees), and random brightness adjustment were performed on the training images during training. These transformations not only expanded the diversity of the training data in real-time but also helped the model learn invariant features of duck behaviours under different visual conditions, enhancing the model's robustness in complex environmental scenarios. As a result, the effective size of the training dataset was substantially increased through these online augmentations. The details of the dataset DuckBehaviour are listed in Table 2.

Table 2
Statistical counts of duck behaviours in the dataset DuckBehaviour.

Type of behaviour	Number of duck behaviours	Proportion (%)
Drinking	606	1.77
Lying	22,380	65.55
Standing	5199	15.23
Eating	1326	3.88
Preening	3956	11.59
Spreading	676	1.98

2.2. Behaviour recognition method for caged ducks

Intelligent behaviour detection and recognition of caged ducks is a crucial component of smart farming, as highlighted in the Section Introduction. However, there are few works in the literature on this research topic and the performance needs further improvement. To address this problem, this study develops a novel behaviour recognition model, CD-ViT-YOLO, which leverages non-contact computer vision technology. The YOLO series, particularly YOLOv5, are known for their exceptional detection performance and robustness. Additionally, YOLOv5 has relatively fewer parameters and thus becomes an ideal choice for this study as the baseline model.

Despite its strength, the C3 block in YOLOv5 struggles with efficient feature extraction for irregularly shaped targets. To overcome this limitation, a block denoted DRE was developed by incorporating DCNv4, IRRep-Bottleneck and hierarchical aggregation structure to replace the block C3 in YOLOv5, enabling more effective aggregation of multi-scale information through the design of multi-branch and cross-layer paths, thereby enhancing the model's performance. Furthermore, as the CSPDarknet backbone network in YOLOv5 has limitations in capturing global features and contains a large number of parameters, this study develops a more lightweight EfficientViT-m0 network to serve as a new backbone network. This backbone network employs a lightweight multi-scale linear attention mechanism, enabling the model to capture global dependencies and expand the receptive field, thus achieving superior feature extraction results while maintaining a lightweight model structure. In practice, there exists a marked imbalance in the number of samples across different behaviour categories in the dataset DuckBehaviour, which generally hinders detection accuracy. To address this problem, a DAD Head is introduced. It estimates difficulty scores for each behaviour category and dynamically adjusts feature weights accordingly, thereby enhancing the model's ability to detect challenging samples.

Combining these optimisation strategies results in the proposed CD-ViT-YOLO for caged ducks. It consists of Modules Backbone, Neck, and Head, as illustrated in Fig. 2. Module Backbone performs feature extraction at different scales on the input duck images. Module Neck fuses the multi-scale features and inputs them to the Module Head. The module Head has three detection heads at different scales to predict the location (i.e., bounding box), category, and confidence level of the target in each feature map. The main improvements of CD-ViT-YOLO over the baseline YOLOv5 include the DRE block in the Module Neck,

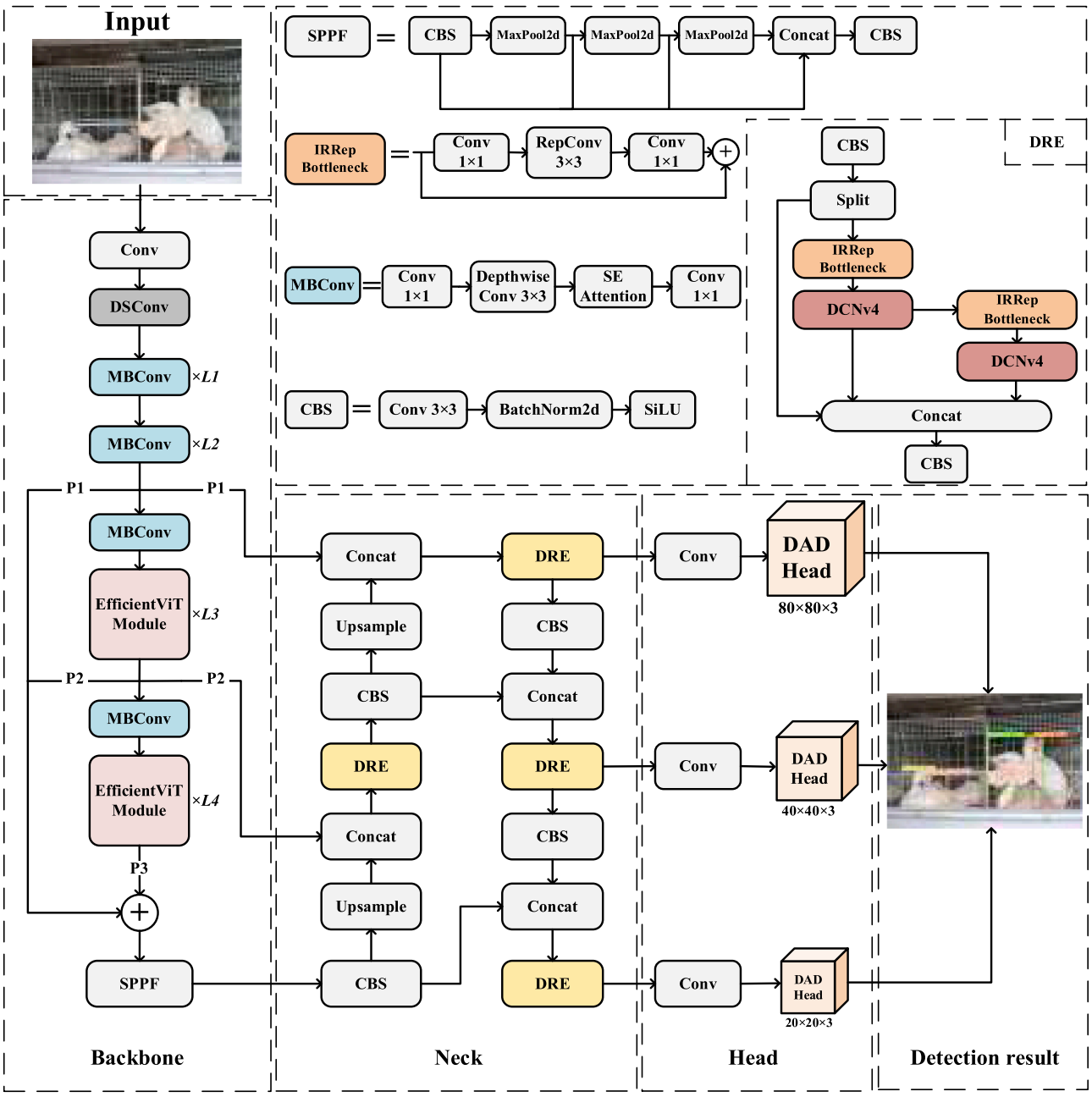

Fig. 2. Structure diagram of CD-ViT-YOLO.

the EfficientViT-m0 network in the Module Backbone, and the DAD Head. The DRE block enhances dynamic feature fusion. The EfficientViT-m0 introduces lightweight global feature modelling with reduced computational cost. The DAD head can alleviate the problem of class imbalance. These components are described in detail in Sections 2.2.1, 2.2.2, and 2.2.3, respectively.

2.2.1. *DRE block*

In the behaviour recognition task for cage-raised ducks, the high target density, complex behaviour, and diverse background interference (e.g., cages) demand strong expressive power and robustness from the feature extraction module. The C3 block constructs a single residual path by stacking Bottleneck layers, resulting in narrow feature flow and ineffective propagation of shallow features to deeper layers, which weakens multi-level feature fusion. Additionally, diverse duck postures and dense aggregation cause partial occlusion, leading to weakened or lost shallow features and reduced detection accuracy, especially for subtle behaviours like preening and lying. The fixed 3×3 convolution kernel in C3 offers limited receptive field and adaptability, insufficient to capture the overall contours of large postures (e.g., preening, lying) or the weak texture, and blurred boundaries of small targets such as drinking tubes. This limitation results in incomplete feature extraction, localization errors, and missed detections. Overall, the C3 block reveals structural bottlenecks in feature propagation and multi-scale perception, restricting detection accuracy and robustness in complex cage-raised duck behaviour scenarios.

This study aims to clarify the design motivation of the proposed DRE block by visualising attention distributions under different module configurations in YOLOv5s. As shown in Fig. 3, with the overall backbone architecture kept unchanged, replacing the C3 block with the DRE block resulted in a distinct difference in feature attention. When using the C3 block (Fig. 3(b)), the attention was relatively dispersed, with

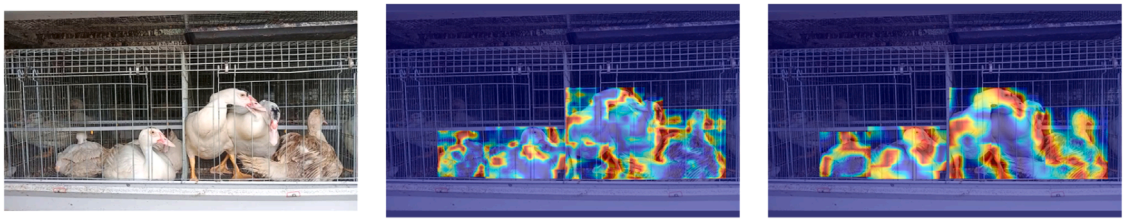

Fig. 3. Comparison of heatmaps in YOLOv5s with different block.

some activated regions falling on background or irrelevant areas. In contrast, after replacement with the DRE block (Fig. 3(c)), the model's attention became more focused on key parts of the duck, particularly the head and torso, which are crucial for behaviour recognition. This comparison reveals the limitations of the C3 block in complex scenarios involving dense aggregation and occlusion of caged ducks, where its feature representation capacity is insufficient and prone to attention drift and missed detections. Drawing from these findings, specific optimisations were applied to the original block with the aim of improving detection accuracy and adapting to complex scenarios.

To address these challenges, the DRE block incorporates DCNv4 for optimisation [56]. The specific configurations of the C3 and DRE blocks are illustrated in Fig. 4(a) and (b), respectively. Unlike traditional rectangular convolution kernels, DCNv4 recognises that optimal kernel shapes can vary across different feature map stages. To facilitate this adaptability, DCNv4 introduces positional offsets at the sampling grid points, enabling the convolution kernel to better conform to the shape and size of the object. Additionally, each sampling point in DCNv4 is assigned a weight, which regulates the offset range of newly sampled points, thereby mitigating the influence of irrelevant factors. Both the positional offsets and weights are refined through gradient descent. This variable convolution design allows DRE to dynamically adjust the convolution kernel according to the actual conditions of duck targets, thereby enhancing the accuracy of feature extraction.

Fig. 5(a) illustrates the sampling flow of DCNv4. Offsets refer to the displacement of each sampling grid point on the feature map, enabling the convolution kernel to adapt more flexibly to the object's shape and size. N is calculated based on the kernel size. For example, with a typical 3×3 convolution, there are 9 parameters, so N equals 9. The offsets and weights of each sampling point in the input feature map are computed by a separate ordinary convolutional layer, as shown in the orange box at the top of Fig. 5(a). The channel dimension of the offset field is 2 N, which includes the offsets in both the x and y directions. Then, new sample point locations are computed based on these offsets and weights. Finally, sampling is performed to obtain the output feature map.

For any point in the sampling window, the expression for the DCNv4 is: Given an input feature map $x \in \mathbb{R}^{H\times W\times C}$(where H, W, and C denote the height, width, and number of channels, respectively), the output feature map y at position p_0 is computed by:

$$y_g(p_0) = \sum_{k=1}^{K} m_{gk} x_g\left(p_0 + p_k + \Delta p_{gk}\right) \tag{1}$$

$$y = concat\left(\left[y_1, y_2, ..., y_G\right], axis = -1\right) \tag{2}$$

where G is the number of spatial aggregation groups, partitioning the input and output channels into $C' = C/G$ channels per group; x_g and y_g represent the sliced input and output feature maps for group g, each with dimensions $\mathbb{R}^{H\times W\times C'}$; K is the number of sampling points; p_k denotes the predefined grid sampling position (e.g., a 3×3 grid for standard convolution); $\Delta p_{gk} \in \mathbb{R}^2$ is the learnable offset for the k-th sampling point in group g, enabling adaptive spatial sampling; $m_{gk} \in \mathbb{R}$ is the modulation scalar (spatial aggregation weight) for the k-th sampling point in group g, computed without softmax normalization (unlike DCNv3), allowing unbounded dynamic weights to enhance model flexibility and expressive power.

DCNv4 has achieved remarkable breakthroughs in both dynamic modelling capabilities and computational efficiency. Removing the softmax normalization in spatial aggregation, optimises the weight constraint mechanism of the traditional DCNv3. In DCNv3, softmax

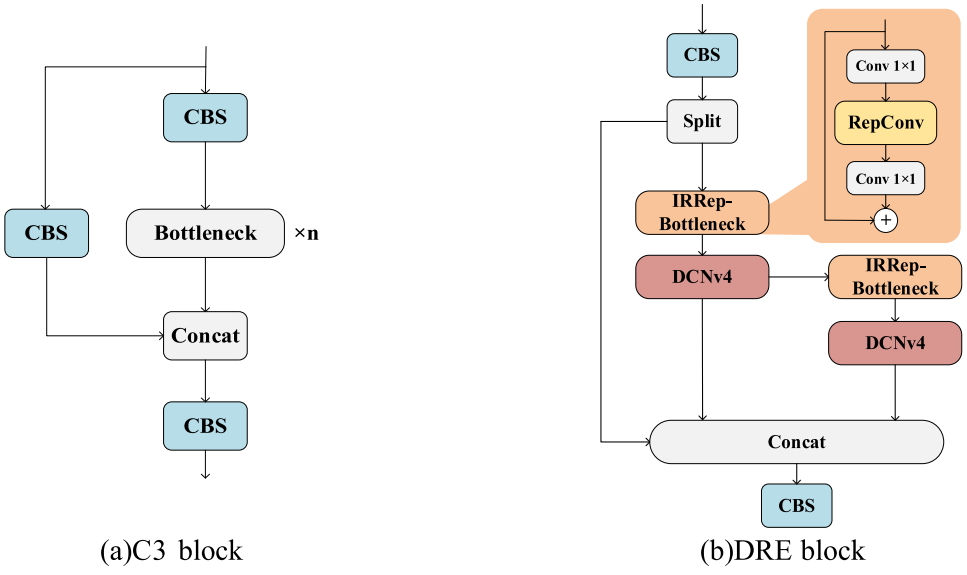

Fig. 4. Diagrams of C3 and DRE blocks: (a) C3 block, (b) DRE block.

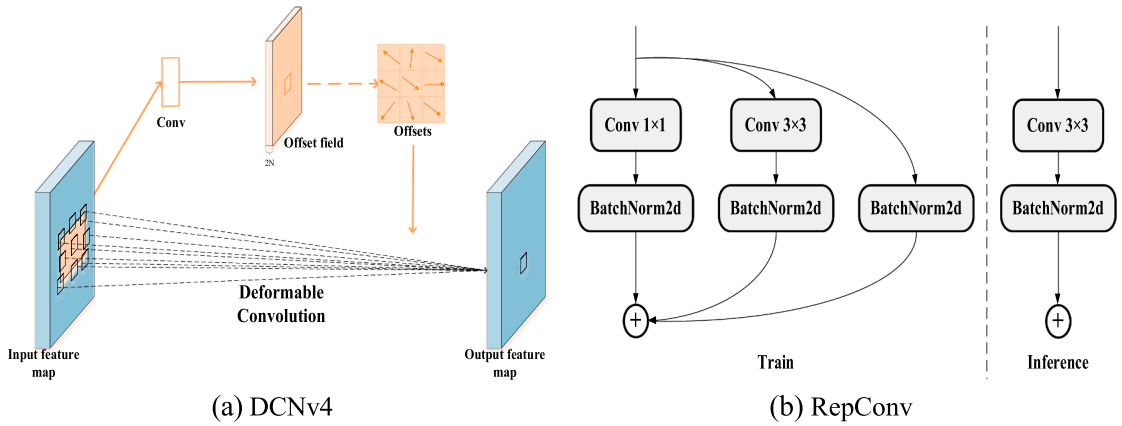

Fig. 5. Schematic diagram of DCNv4 and RepConv structure.

restricts the spatial aggregation weights within the range of 0–1. Although this conforms to the design logic of the dense attention mechanism, it limits the expressive power of the convolutional operator for irregular targets. In contrast, DCNv4 removes this constraint, changing the range of the modulation factor from [0,1] to unbounded dynamic weights, which significantly enhances the model's ability to capture the features of complex targets such as deformed objects. In terms of optimising computational efficiency, DCNv4 focuses on the memory access bottleneck. When reading data, it combines the operations of reading single values multiple times into one operation of reading multiple values at once, dramatically reducing the number of memory accesses. Meanwhile, by using the channel-group parallel processing method, a single thread can process multiple channels within the same group simultaneously, avoiding the repeated reading of the same sampling offsets and aggregation weight values by different threads, and effectively reducing the bandwidth consumption. Additionally, DCNv4 introduces half-precision computing to balance accuracy and computational overhead.

In the task of detecting the behaviours of caged ducks, these improvements play a crucial role. By leveraging its dynamically changing convolution kernel, DCNv4 enables the convolution kernel to better adapt to target outline (such as small targets like water pipes or large targets like the behaviour of spreading), thereby significantly enhancing the ability to extract target features, thereby reducing missed detections and localization biases. Additionally, the computational efficiency of the model is improved through optimised methods including vectorised load, channel-group parallel processing, and half-precision computing. This allows it to meet the requirement of real-time processing of a large amount of image data of caged ducks, providing timely information for breeding management, thereby enabling more scientific and efficient practices.

As previously discussed, the C3 block exhibits structural limitations in the context of complex background caged duck behaviour detection, resulting in insufficient feature fusion capabilities. To address this, this study proposes an Inverted Residual *Re*-parameterised Bottleneck (IRRep-Bottleneck) within the DRE Block. The design builds upon the principles of *Re*-parameterised convolution (RepConv), which merges multiple branches, typically including 3 × 3 convolution, 1 × 1 convolution, and identity mappings(as shown in Fig. 5(b)) into a single equivalent convolution during inference. This enables the network to benefit from the representational richness of multi-branch structures during training while maintaining a compact and efficient single-branch form at deployment. As a result, RepConv enhances spatial modelling and feature diversity. Building on this, an inverted residual structure is adopted, where the input feature is first projected into a higher-dimensional space through an expansion layer, enhancing the capacity for intermediate feature representation. The expanded feature is then processed by a RepConv block, effectively capturing both local and contextual spatial features. Finally, a projection layer compresses the feature back to its original dimension. A skip connection is introduced between the input and output to facilitate gradient flow and feature reuse. By integrating RepConv into the inverted residual framework, the proposed IRRep-Bottleneck not only preserves spatial details and semantic richness but also improves the information flow within the network. This design is particularly well-suited for dense scenes such as cage-raised duck behaviour recognition, where fine-grained spatial features and efficient computation are both critical.

Conventional C3 block in YOLOv5 suffer from structurally identical branch inputs, leading to limited gradient diversity and redundant feature learning. To overcome these limitations, the proposed DRE block adopts a hierarchical feature aggregation approach combined with partial cross-layer connections to enhance both feature representation and computational efficiency. Specifically, the input feature map is first split along the channel dimension into two branches: one branch is retained as a shortcut for later fusion, while the other is processed through two sequential IRRep-Bottleneck for deep feature extraction, enabling the branches to learn diverse features and promoting more diversified gradient flow during training. To further strengthen spatial modelling capabilities and object alignment perception, each IRRep-Bottleneck is followed by a DCNv4. The outputs from the original shortcut path and the two deep branches are then selectively concatenated along the channel dimension and fused via a 1 × 1 convolution for information integration and dimensionality reduction. By integrating channel splitting with more efficient convolutional operators, the proposed structure enhances the diversity of feature representations across branches and enables dynamic receptive fields tailored to varying object geometries. This design facilitates more robust spatial modelling and improves the network's ability to capture subtle behavioural cues under complex scenarios. In the task of cage-raised duck behaviour recognition, where challenges such as high object density, complex postures, and cluttered backgrounds are present, the DRE block enables precise behavioural feature extraction without increasing inference overhead, thereby improving detection robustness and accuracy in complex environments.

2.2.2. *YOLOv5 backbone network optimisation*

The YOLOv5's backbone network uses the CSPDarknet53 structure, which incorporates the Cross Stage Partial (CSP) mechanism. This mechanism divides the input feature map into two parts for processing, enabling efficient extraction of multi-scale features and expanding the receptive field. Despite these advantages, CSPDarknet53 primarily relies on local convolution operations, which are inadequate for capturing global features and adapting to the global dependencies of irregular targets in complex scenes. Furthermore, the network's parameter number is relatively high, indicating potential for improvement in balancing lightweight design and performance enhancement.

To optimise the YOLOv5's backbone network, lightweight convolutional neural networks are generally employed as replacements for the original backbone [13,58]. Although these approaches have achieved some success in reducing the model's complexity, they have also resulted in significant performance losses. Consequently, a more effective strategy is needed to optimise the backbone network of YOLOv5.

In recent developments, ViT has demonstrated strong performance in various recognition tasks. Unlike CNNs, ViT can directly model global contextual information through a self-attention mechanism, which allows each image block to interact with all others to capture global dependencies, thereby improving feature extraction. However, the self-attention module in ViT presents a computational bottleneck, as its complexity increases quadratically with the input resolution. To mitigate this problem, previous studies have suggested reducing the input resolution [52,55] and limiting attention to a fixed-size localised window (e.g., 7 × 7) [25,26]. Although these methods alleviate computational complexity to some extent, they compromise ViT's core strength in global feature extraction.

To address the computational inefficiency of ViT's self-attention module, Cai et al. [4] introduced the Efficient Vision Transformer (EfficientViT). As illustrated in Fig. 6, EfficientViT's network structure comprises four stages. Initially, the Stem operation extracts features by downsampling the input in the first stage. Subsequently, Stages 1 to 3 undergo downsampling using a series of MBconv with a stride of 2, The MBConv contains two 1 × 1 convolutions for channel adjustment, a Depthwise (DW) convolution, and a squeeze-and-excitation (SE) attention. Several EfficientViT modules are integrated into Stages 2 and 3. The outputs from Stages 1, 2, and 3, labelled as P1, P2, and P3, respectively, are then subjected to pyramid feature mapping. This involves matching their channels and sizes through 1 × 1 convolution and merging the three outputs via addition. Finally, the output feature maps of EfficientViT are obtained through the SPPF module.

The EfficientViT module is the core of the EfficientViT network structure. It consists of a multi-scale linear attention module, a Feed Forward Neural network(FFN), and DWconv, as shown in Fig. 7. Specifically, the input feature map is converted into three sets of features, Query(Q), Key(K), and Value(V), by a linear operation (Linear). Then, Q, K, and V are fed into three branches for processing to enable multi-scale attention computation. Attention employs ReLU linear attention [20] to enable a global receptive field, instead of the computationally expensive softmax attention. The first branch uses linear attention computation for direct attention computation, and the other two branches first process these features using depth-separable convolutions at different scales to generate multi-scale Q, K, and V. The computation of the other two branches is briefly described below.

The attention calculation process for the second and third branches is as follows. Initially, features are processed using depthwise separable convolutions of sizes 3 × 3 and 5 × 5, along with 1 × 1 group convolutions. These features are then input into linear attention for further computation. Specifically, in linear attention, the Q and K features undergo ReLU activation to enhance nonlinear expression. Subsequently, the transpose of K is multiplied by V through matrix multiplication to obtain a weighted feature representation (*KTV*), which is then multiplied by the Q matrix to produce the final attention output for the branch.

Upon completing the calculation of attention for the three branches, the features generated by each branch are fused at multiple scales. This fusion process results in an enhanced feature map, which is subsequently projected back to its original dimension through a linear operation. The enhanced feature map is then processed by the FFN + DWconv module. Initially, this module employs a 1 × 1 convolution to expand the number of channels, thereby improving feature expressiveness. Following this, the module utilises DWconv to capture local features, enhancing model performance while maintaining computational efficiency. Finally, a 1 × 1 convolution compresses the number of channels to match the input channel number, ensuring consistency in the output feature map. Throughout this process, the residual connection plays a crucial role in maintaining a stable information flow and effectively enhancing the feature map. As a result, the output feature map exhibits enhancement at both global and local levels, demonstrating a stronger feature expression capability.

Through the above branching path attention calculation and effective fusion, the EfficientViT module greatly improves the feature extraction ability while ensuring computational efficiency, which provides an effective way to optimise the performance of the visual transformer model.

Previous studies have shown that allocating more functional blocks to deeper stages of a network can enhance model performance, particularly by concentrating blocks in the third stage to achieve a balance between accuracy and model complexity [36,37]. Following this trend, recent lightweight ViT architectures such as Conv2former-S and RepViT have adopted aggressive stage ratios like 1:1:8:1 and 1:1:7:1, respectively [17,45]. Similarly, EfficientViT provides a series of scalable models with varied stage distributions (i.e., L1–L4) to address different application requirements, as illustrated in Fig. 6. Drawing on this design principle, the present study proposes a modified version named EfficientViT-m0, which employs a three-stage configuration and a revised block ratio of 2:2:4:3. This compromise design increases the number of blocks in the later stages, enabling the network to capture more comprehensive global features and thereby enhance detection performance.

2.2.3. *Difficulty-Aware detection(dad) head*

In the task of cage-raised duck behaviour detection, a notable class imbalance exists, with behaviours such as drinking and spreading being markedly underrepresented compared to frequent actions like lying (see Table 2). This imbalance can lead the model to favour dominant classes, thereby impairing its ability to detect rare or difficult behaviours. To address this challenge, this paper introduces the DAD Head, which incorporates class-level difficulty information into the classification branch of the detection head.

In the multi-label detection framework of YOLOv5, each target class is predicted independently using a sigmoid activation function, and training is optimised with binary cross-entropy (BCE) loss. The proposed DAD Head introduces learnable class difficulty embedding vectors to perform additive modulation on the logits of difficult classes, enabling

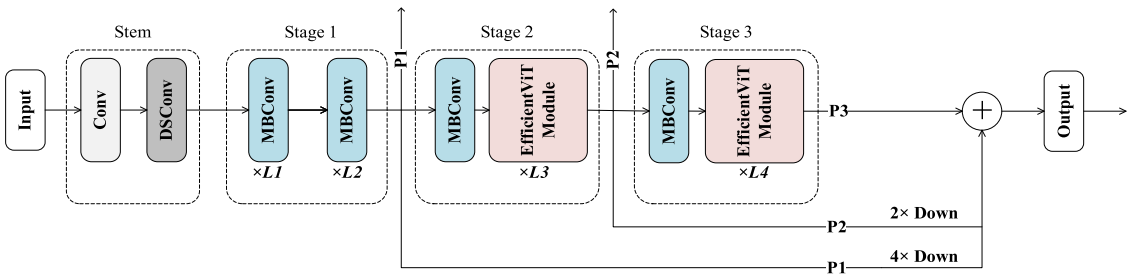

Fig. 6. EfficientViT network structure.

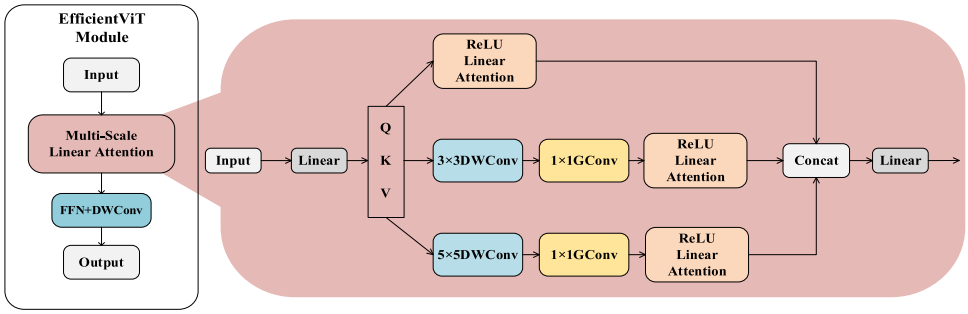

Fig. 7. EfficientViT module and Multi-scale linear attention.

the predicted probability for the corresponding class to better approximate the ground truth. When a sample is misclassified, the BCE loss is particularly sensitive to low-confidence predictions and produces larger gradients for the adjusted logits, thereby increasing the model's focus on the corresponding class. The proposed approach consists of the following key steps:

(1) Class Difficulty score Encoding: For each class c, a difficulty score $d_c$is calculated as the inverse of its sample count as shown in Eq (3):

$$d_c = \frac{1}{N_c + \varepsilon}, \tag{3}$$

where N_c is the number of training samples for class c, and ε is a small constant to avoid division by zero.

To maintain numerical stability and enable learning, the difficulty scores are normalised:

$$\widetilde{d_c} = \frac{d_c}{\sum_{i=1}^{C} d_i} \tag{4}$$

(2) Difficulty score Embedding Generation: As shown in Eq (5), the normalised difficulty score $\widetilde{d_c}$ is passed through a lightweight MLP to generate a learnable embedding vector $e_c \in \mathbb{R}^D$:

$$e_c = \mathrm{MLP}(\widetilde{d_c}) \tag{5}$$

The MLP consists of 1–2 fully connected layers with ReLU activations. The resulting vector e_c captures class-specific difficulty information in a compact form and serves as a conditioning input to the classification branch.

(3) Embedding Injection: x denote the shared feature representation input to the classification head. For each class c, the raw classification logit l_c is computed using a convolutional layer, and then inject the difficulty embedding e_c via additive modulation, as shown in Eq (6):

$$l_c = \mathrm{Conv}_{cls}(x), \quad l'_c = l_c + \mathrm{W}_p \cdot e_c \tag{6}$$

where $\mathrm{W}_p \in \mathbb{R}^{K \times D}$ is a learnable projection matrix that maps the embedding vector $e_c \in \mathbb{R}^D$ to the same dimension as the classification logits. This design can be viewed as injecting a class-specific "modulation bias" into the classification head, enabling additive correction to the raw logits. It helps boost the response of long-tail categories in the lower score range, thus improving the detection sensitivity for underrepresented behaviours.

The proposed DAD Head addresses class imbalance in caged duck behaviour detection by dynamically modulating classification outputs based on the varying difficulty of different behaviours. Through the introduction of learnable difficulty embeddings, the head enhances the model's sensitivity to rare or hard-to-distinguish behaviours without introducing excessive computational overhead. This is because the embedding injection is lightweight and only affects the classification logits, without altering the backbone or increasing feature map dimensions. This design is particularly beneficial for improving the model's response to difficult samples, such as behaviours with visual ambiguity or low inter-class variance, thereby enhancing detection robustness and precision in imbalanced and complex scenarios.

2.3. Training settings

All experiments in this study were conducted on a single server. The server used a 64-bit Windows 10 operating system, 16 GB of random access memory, and an NVIDIA RTX3070 GPU with 8 GB of video memory. Additionally, the deep learning framework used was PyTorch 2.0.1. The CUDA version was 11.7. the development software used was Pycharm version 2021.2.1, and the programming language Python version is 3.9.0.

The dataset DuckBehaviour was used for model training and evaluation, where the training, validation and test sets were divided as described in Section 2.1. The same parameters were used for all model training, i.e., batch size was set to 4, the maximum number of iterations epoch was 300, the optimiser was SGD, and the initial learning rate was set to 0.01.

2.4. Model evaluation metrics

In this study, the precision (P), recall (R), F1 score, and the mean of the average precision (mAP) of all categories were used to assess the effectiveness of CD-ViT-YOLO for behaviour detection in caged ducks. Metric precision is the proportion of correct predictions among all samples predicted to be positive. Metric recall refers to the proportion of samples predicted to be positive among those that were positive. Metric F1 is introduced to solve the conflicting evaluation metrics of P and R to comprehensively measure the performance of the model. Metric mAP is a commonly used evaluation metric in target detection models, especially in multi-classification problems, and is obtained by calculating the mean of the Average Precision(AP) of all classes. When the IoU threshold is 0.5, the calculated mAP is denoted as mAP@0.5. mAP@0.5:0.95 metric averages all mAP across IoU thresholds from 0.5 to 0.95 with an interval of 0.05, which is more stringent than mAP@0.5. As shown in Eqs. (7) – (10) all these metrics are defined as:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{7}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{8}$$

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \times 100\% \tag{9}$$

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{n} \times 100\% \tag{10}$$

where TP is the number of positive samples when the model correctly predicts positive, FP denotes the number of negative samples when the model incorrectly predicts positive, FN is the number of positive samples when the model incorrectly predicts negative, and n represents the

number of categories.

$$Params = \sum (K \times K \times C_{in} \times C_{out}) \tag{11}$$

$$FLOPs = \sum (K \times K \times C_{in} \times C_{out} \times H \times W) \tag{12}$$

In addition, the number of parameters, #Params, and floating point operations, GFLOPs, were used to assess model complexity, as shown in Eqs. (11) and (12). Where K is the size of the convolution kernel, C_{in} is the number of input channels, C_{out} is the number of output channels, and H and W represent the height and width of the output feature map, respectively. The higher #Params and GFLOPs of the model, the higher the model complexity is, and the more the computational resources are required; and vice versa. Besides, the FPS is chosen to measure the real-time performance of the model. FPS indicates the number of image frames processed in one second.

3. Experimental results

In this study, a series of experiments were conducted to comprehensively evaluate the performance of CD-ViT-YOLO in detecting caged duck behaviours. Section 3.1 presents ablation experiments based on the YOLOv5s model to analyse the impact of various modifications. Section 3.2 presents the detection capabilities of the proposed CD-ViT-YOLO model for various behaviours under different lighting conditions, along with visualisation of the detection results and Section 3.3 discusses the comparative results between the proposed CD-ViT-YOLO model and other target detection models, highlighting its effectiveness and potential advantages.

3.1. Ablation experiments and analysis

3.1.1. Performance comparison on different backbone networks

The CD-ViT-YOLO is obtained by modifying the YOLOv5s model, which has fewer parameters and better performance in the YOLOv5 series and exhibits a better balance between detection performance and speed. Therefore, choosing YOLOv5s as the base model can better meet the research objectives of this study.

To verify the effectiveness of the adopted new backbone EfficientViT-m0, this study compares its performance with that of different backbone networks. For the experimental simulation, the mainstream and excellent lightweight backbone networks such as Fasternet [8], MobilenetV3 [18], ShufflenetV2 [27], Ghostnet [14], FastViT-t8 [44], MobileViTv2 [29] and EfficientViT-m0 were used to replace the backbone network of YOLOv5s. To ensure a fair comparison, all models were trained and evaluated under identical conditions, including the same training epochs, learning rate, data augmentation strategies, and evaluation metrics. This setup highlights the role of the backbone network, enabling a more objective assessment of its impact on detection results. In addition, the selected backbone networks cover both CNN-based and Transformer-based architectures, enabling a comprehensive evaluation of different networks in the context of caged duck behaviour detection. Table 3 shows the caged duck behaviour detection performance of these variants.

According to the results presented in Table 3, substituting the original YOLOv5s backbone with a lightweight convolutional neural network markedly decreases the number of model parameters but incurs a certain reduction in accuracy. Compared with the baseline YOLOv5s, using EfficientViT-m0 as the backbone greatly decreases the number of parameters and GFLOPs by 43 % and 53 %, respectively, while maintaining comparable mAP performance. Specifically, mAP@0.5:0.95 slightly improves due to the lightweight linear attention mechanism in EfficientViT-m0, which extracts multi-scale global features with fewer parameters, and the DWConv block, which further reduces parameters.

Further analysis of Table 3 reveals that taking EfficientViT-m0 as the backbone achieves superior behaviour detection performance with fewer parameters than other methods. For instance, EfficientViT-m0-YOLOv5s achieves the highest mAP@0.5 and mAP@0.5:0.95. Compared with GhostNet-YOLOv5s, which performs the second best, EfficientViT-m0-YOLOv5s decreases parameters by 18.3 % and improves mAP by 2.1 %. Additionally, several lightweight Vision Transformer networks (e.g., MobileViT) are included in the comparison. Compared with them, EfficientViT-m0 achieves a better balance between model size and detection performance, demonstrating superior adaptability when integrated with YOLO models. This suggests that EfficientViT-m0 is a more suitable lightweight network choice for the task at hand. Therefore, this study selects EfficientViT-m0 as the backbone network.

3.1.2. Ablation on different stage ratios in the EfficientViT network

Table 4 presents the performance of EfficientViT networks with different stage ratios. The five stage ratios (2:2:2:2, 2:2:2:4, 2:2:3:3, 2:2:3:4, and 2:2:4:3) were chosen to explore how varying the distribution of blocks across different stages of the EfficientViT network influences performance. The motivation behind this experiment is the current trend in ViT architectures, where increasing the number of blocks in the later stages has been proven to enhance model performance, particularly in capturing more complex, high-level features.

Specifically, it is hypothesised that increasing the number of blocks in the later stages of the network can refine feature extraction, which is critical for tasks involving high-resolution or detailed image data. The 2:2:4:3 ratio is selected as the primary candidate based on this intuition, as allocating more blocks to the later stages may improve performance on complex visual features.

By testing these five ratios, the objective is to identify a more efficient ratio that optimises performance while maintaining a reasonable computational cost. The comparative results will provide valuable insights into the impact of block distribution on the overall performance of EfficientViT networks. The experimental results are summarised in Table 4.

As shown in Table 4, compared with the original stage ratio of 2:2:2:2, the improved stage ratio (2:2:2:2 → 2:2:4:3) significantly enhances performance metrics with a slight increase in parameter number,

Table 3
Performance comparison on behaviour recognition of caged ducks with different backbone networks.

Model	P (%)	R (%)	F1 (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	#Params (M)	GFLOPs	FPS
YOLOv5s	93.2	95.8	94.4	96.3	86.3	7.02	15.8	76.468
Fasternet-YOLOv5s	94.1	92.6	93.4	95.2	84.3	3.33	6.7	88.163
ShufflenetV2-YOLOv5s	92.9	92.8	92.8	94.8	84.8	3.34	7.0	87.942
MobilenetV3-YOLOv5s	91.9	94.9	93.3	95.8	84.8	3.46	6.6	80.648
Ghostnet-YOLOv5s	94.3	93.0	93.6	95.1	85.3	4.87	6.6	79.546
FastViT-t8 YOLOv5s	92.5	92.2	92.3	94.5	86.6	6.61	14.6	62.338
MobileViTv2-YOLOv5s	91.6	94.1	92.5	94.9	86.0	4.38	11.5	70.157
EfficientViT-m0-YOLOv5s	93.4	94.2	93.7	95.8	87.4	3.98	7.4	77.695

Table 4
Comparison of EfficientViT network test results with different stage ratios.

Stage ratios	P (%)	R (%)	F1 (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	#Params (M)	GFLOPs
2:2:2:2 (Original)	95.1	91.9	93.5	95.0	84.9	3.62	6.9
2:2:2:4	93.5	90.0	91.7	95.5	85.4	4.09	7.2
2:2:3:3	92.2	93.0	92.6	95.2	85.0	3.92	7.3
2:2:3:4	93.9	90.5	92.2	95.3	85.5	4.15	7.4
2:2:4:3 (EfficientViT-m0)	93.4	94.2	93.7	95.8	87.4	3.98	7.4

achieving a 2.5 % improvement in mAP@0.5:0.95. Moreover, among the tested stage ratios, EfficientViT-m0 with a ratio of 2:2:4:3 demonstrates superior performance in metrics such as mAP@0.5:0.95, achieving the highest detection accuracy with a relatively small parameter number.

3.1.3. Ablation on model design strategies

As described in Section 2.2, the backbone of YOLOv5s was replaced by EfficientViT-m0 with lightweight and high-performance, and a DRE block was developed to replace the C3 block in Module Neck of YOLOv5s. This modification overcomes the challenge of the C3 block in efficiently extracting features from irregularly shaped objects. The novel DAD head can mitigate the issue of sample imbalance to a certain degree. The results of the ablation experiments are presented in Table 5.

As shown in the ablation Table 5, each improvement strategy proposed in this study contributes distinct benefits in terms of detection accuracy and model complexity. Incorporating only EfficientViT-m0, a lightweight ViT network employing linear attention, leads to a notable increase in mAP@0.5:0.95 (87.4 %). Moreover, the number of parameters and GFLOPs considerably drop to 3.98 M and 7.4, respectively. These results indicate that EfficientViT-m0 helps improve fine-grained recognition performance under high-IoU conditions while effectively reducing model complexity. The improvement stems from its ability to model long-range dependencies with minimal overhead, making it better suited for subtle behavioural distinctions. When adding the DRE block, the mAP@0.5 and mAP@0.5:0.95 rise to 96.6 % and 87.8 %, respectively, with minimal changes in #Params and GFLOPs. This performance gain can be attributed to DRE's enhanced spatial adaptability and dynamic receptive fields, which are particularly beneficial for handling irregular object contours, occlusions, and background interference in cage-raised duck behaviour detection. Integrating the DAD head that embeds a difficulty score vector into the prediction process, raises the mAP@0.5 to 96.4 % and mAP@0.5:0.95 to 87.2 %, without increasing model complexity. This result demonstrates that DAD improves the model's sensitivity to hard samples and alleviates the class imbalance problem by dynamically adjusting the detection confidence based on sample difficulty. Combining multiple modules yields further performance improvements. For example, using both EfficientViT-m0 and DRE increases mAP@0.5 to 96.8 % and mAP@0.5:0.95 to 88.1 %, while reducing the parameter count and computational cost.

Finally, when all three novel strategies(EfficientViT-m0, DRE, and DAD head) are integrated into the baseline YOLOv5s model, the resulting CD-ViT-YOLO achieves notable progress in overall accuracy, with mAP@0.5 reaching 97.4 % and mAP@0.5:0.95 reaching 88.6 %, while still maintaining low #Params and GFLOPs. These results demonstrate that the proposed methods complement each other and jointly enable a better trade-off between model efficiency and detection precision.

3.2. Comparison between CD-ViT-YOLO and the baseline YOLOv5s model

3.2.1. Performance comparison with YOLOv5s on different behaviours

A comprehensive evaluation of the model's performance was conducted using five-fold cross-validation, with the final result of each fold presented using box plots to more intuitively reflect performance variation. To reduce the impact of fluctuations during training, the metric from the final epoch of each fold was selected as the representative result, as it better reflects the model's final convergence state. Benefiting from the introduction of a Lightweight Vision Transformer backbone that enhances global context modelling and optimisations to the feature extraction structure, CD-ViT-YOLO demonstrates improved performance in detecting behaviours in the dense and occluded environments typical of caged duck farming. As shown in Fig. 8, CD-ViT-YOLO significantly outperforms YOLOv5s in terms of mAP@0.5:0.95, with results ranging from 87.9 % to 89.2 %, compared to 85.3 % to 87.0 % for YOLOv5s. In terms of total loss, CD-ViT-YOLO also exhibits lower and more stable values (0.0387–0.0419), indicating better accuracy and convergence.

Table 6 provides a detailed analysis of detection performance on six daily behaviours of caged ducks. It reveals that CD-ViT-YOLO outperforms YOLOv5s in recognizing these behaviours, as indicated by its higher mAP@0.5:0.95. This improvement is attributed to CD-ViT-YOLO's incorporation of a novel backbone network and EfficientViT-m0, which captures richer features and enables the model to more accurately distinguish between different behaviours.

Further analysis of Table 6 indicates that among the six daily behaviours, lying, standing, spreading, and eating have distinct characteristics, thereby resulting in higher detection performance for these behaviours. Notably, both the CD-ViT-YOLO and YOLOv5s models show relatively weak performance in detecting the preening behaviour. This is because the preening behaviour involves ducks using their beaks to groom their feathers. From the model's perspective, the extracted features during preening behaviour are very similar, regardless of the duck's posture, making it difficult to distinguish between these situations and thus affect detection performance. Additionally, detecting

Table 5
Ablation experiments results.

Basic Model (YOLOv5s)	+EfficientViT-m0	+DRE	+DAD Head	mAP@ 0.5(%)	mAP@ 0.5:0.95(%)	#Params(M)	GFLOPs
√				96.3	86.3	7.02	15.8
√	√			95.8	87.4	3.98	7.4
√		√		96.6	87.8	7.04	16.0
√			√	96.4	87.2	7.02	15.8
√	√	√		96.8	88.1	4.7	8.8
√	√		√	95.9	87.6	3.99	7.4
√		√	√	97.0	88.3	7.05	16.0
√	√	√	√	97.4	88.6	4.75	8.7

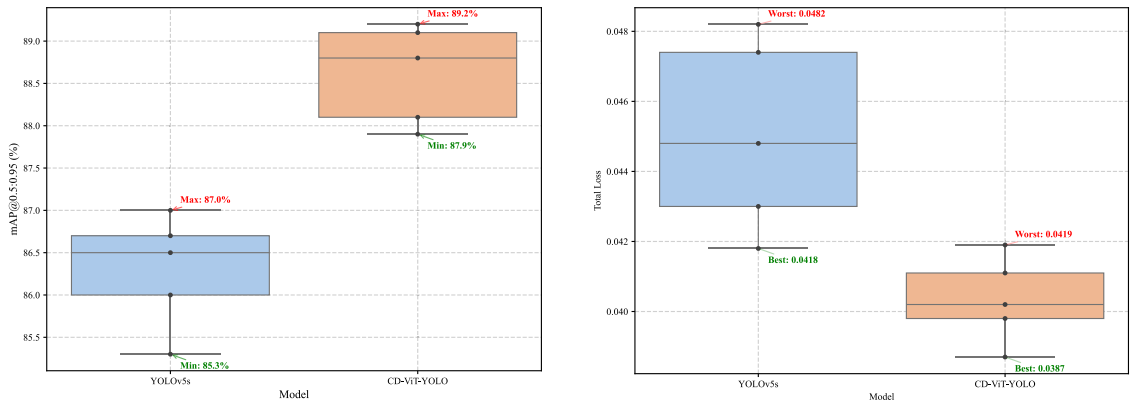

Fig. 8. Performance comparison of CD-ViT-YOLO and YOLOv5s under Five-Fold Cross-Validation. The left provides mAP@0.5:0.95 and the right shows the total loss values.

Table 6
Comparison of YOLOv5s and CD-ViT-YOLO in recognizing different duck behaviours.

Model	behaviour	P (%)	R (%)	F1 (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv5s	drinking	89.9	94.2	92.0	95.1	86.3
	lying	98.0	97.1	97.5	98.2	88.3
	standing	96.1	91.7	93.7	95.9	85.9
	eating	94.6	97.9	96.3	98.0	93.9
	preening	90.2	91.0	90.6	93.9	83.6
	spreading	87.5	97.6	92.2	96.8	79.8
CD-ViT-YOLO	drinking	96.9	93.3	95.1	96.5	88.4
	lying	98.2	96.5	97.4	99.2	88.9
	standing	97.3	93.8	95.5	96.8	88.8
	eating	95.6	98.1	96.8	98.7	94.7
	preening	90.1	91.3	90.7	95.6	87.7
	spreading	90.0	97.8	93.8	97.5	83.3

drinking behaviour poses another challenge, as the water pipes that supply the drinking water are located at the rear of the cage and are rather narrow, making them less prominent in the images. As a result, the models may struggle to determine whether the duck is drinking.

Detection outcomes for different behaviours in caged ducks are illustrated in Fig. 9, facilitating an intuitive comparison between YOLOv5s and CD-ViT-YOLO. Comparing Fig. 9(a), (b), and (c), it is evident that YOLOv5s produces redundant detection boxes, showing an extra box compared with the ground truth, such as an additional box labelled as lying down in the cage on the right. However, the proposed CD-ViT-YOLO accurately detects the duck's behaviour without such errors. By comparing Fig. 9(d), (e), and (f), it can be seen that YOLOv5s mistakenly detects a duck's behaviour as preening, resulting in an erroneous detection box. In contrast, the CD-ViT-YOLO correctly identifies the behaviour. Furthermore, Fig. 9(g), (h), and (i) show that occlusion between ducks poses a challenge for the model, causing YOLOv5s to miss a severely occluded duck. However, CD-ViT-YOLO successfully detects the previously missed duck and accurately recognises its behaviour. Finally, by comparing Fig. 9(j), (k), and (l), it is clear that YOLOv5s misidentify a duck's behaviour as spreading and fail to detect another duck due to the occlusion between them. In contrast, CD-ViT-YOLO, benefiting from the enhanced feature extraction capabilities provided by the DRE block, can mitigate the occlusion problem to some extent. As a result, CD-ViT-YOLO accurately detects the occluded duck, demonstrating superior performance compared with YOLOv5s.

3.2.2. *Performance under different lighting conditions*

Moreover, Table 7 presents a comprehensive comparison between YOLOv5s and the proposed CD-ViT-YOLO model under varying lighting conditions. In this study, a corresponding relationship between light changes and time periods has been established. That is, an adequate light environment corresponds to daytime (12:00 pm to 5:00 pm, approximately 300–500 lx), a medium light intensity environment corresponds to evening (5:00 pm to 7:00 pm, approximately 50–200 lx), and an insufficient light environment corresponds to nighttime (7:00 pm to 9:00 pm, approximately 20–50 lx). Under varying lighting conditions, CD-ViT-YOLO consistently outperformed YOLOv5s, with particularly evident improvements observed under adequate and medium lighting. For most behaviours, higher scores were achieved in both mAP@0.5 and mAP@0.5:0.95. For example, under medium light, CD-ViT-YOLO reached 98.5 % mAP@0.5 and 94.5 % mAP@0.5:0.95 for eating, representing an improvement of 1.1 percentage points over YOLOv5s. Under insufficient lighting conditions, the detection performance of both models declined noticeably due to reduced image contrast and increased noise, which obscure key features of duck behaviours. Actions such as preening and spreading become particularly difficult to distinguish because of partial occlusions and diminished visual clarity. Despite these challenges, CD-ViT-YOLO maintained a clear performance advantage. For example, in recognising lying behaviour, CD-ViT-YOLO achieved a mAP@0.5 of 98.8 %, surpassing YOLOv5s by 2.4 percentage points. Future work will focus on further enhancing behaviour detection performance under low-light conditions. It is worth noting that a marked shift in the behavioural patterns of the caged ducks was observed during this period: no eating behaviour occurred, and drinking samples were extremely limited. Therefore, only four frequently occurring behaviours were retained for comparison. Overall, these results demonstrate that the proposed CD-ViT-YOLO model achieves robust and reliable detection performance across varying lighting conditions, effectively handling the challenges posed by changes in illumination.

Across all lighting conditions, the detection performance for preening and spreading remained consistently lower than that of other behaviours, as shown in Table 7. Combined with the analysis in Fig. 8(e), it can be observed that preening is prone to misclassification due to several factors. It involves only small, localised movements of the head and neck, while the rest of the body remains still, making it visually similar to static behaviours such as lying. Additionally, the duck often lowers its head or tucks it close to the body during preening, and the head may even be obscured by wings, hiding key visual features. Furthermore, in the densely populated cage environment, ducks frequently overlap or cluster together, increasing occlusion and making it more difficult for the model to distinguish individual behaviours. These challenges collectively contribute to the reduced recognition accuracy for preening. The difficulties in detecting spreading can be attributed to several factors. First, this behaviour is typically characterised by a large extension of the wings, causing significant shape variation and increasing the likelihood of body parts being cropped at the image boundaries. Second,

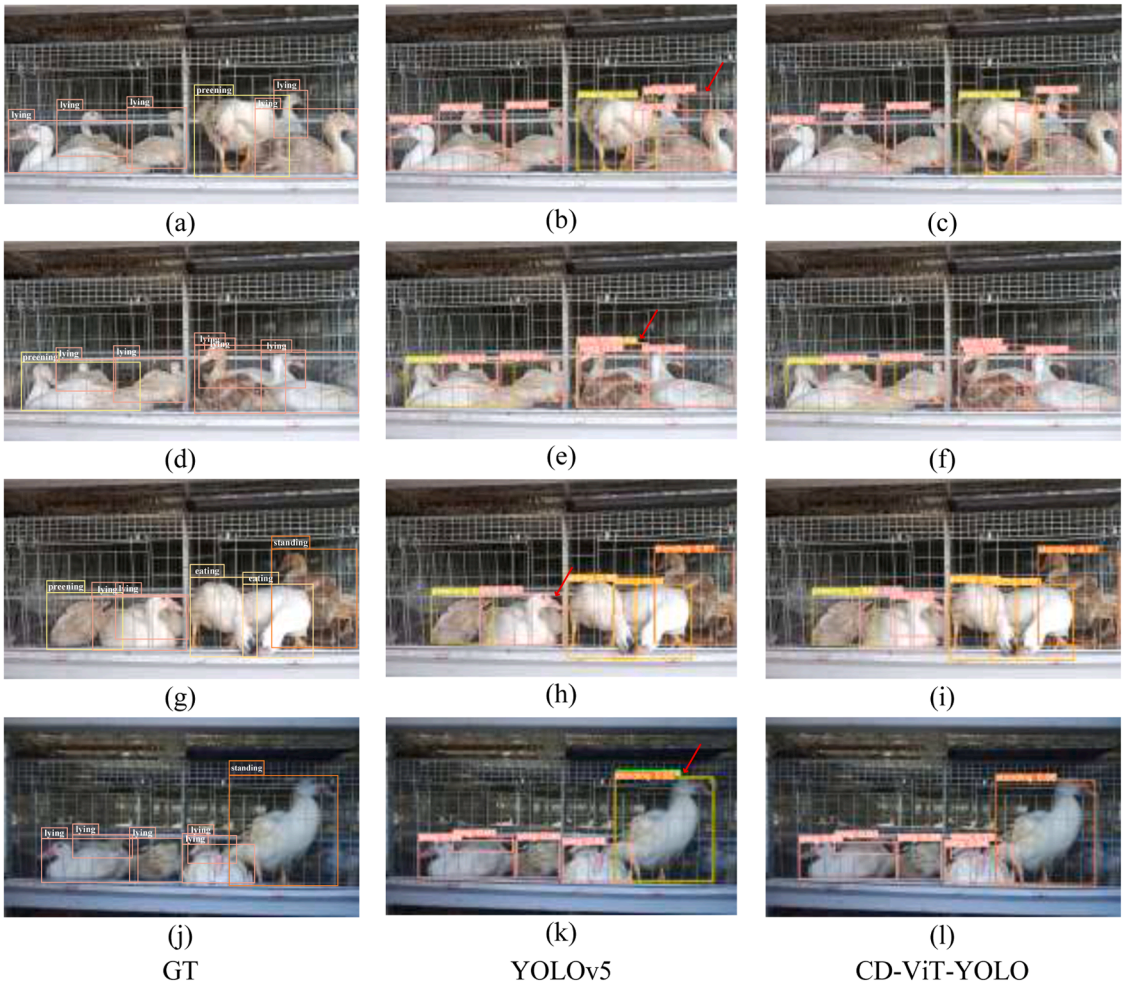

Fig. 9. Comparison of detection results between YOLOv5s and CD-ViT-YOLO, where Ground Truth (GT) denotes the real behavioural annotation. The first column is the original image of the input, the middle column is the detection result of the YOLOv5s model, and the rightmost column is the detection result of the CD-ViT-YOLO model. Red arrows indicate redundant, missed, and wrong detection boxes.

Table 7
Performance comparison CD-ViT-YOLO and YOLOv5 under varying lighting conditions.

Lighting conditions	Behaviour	YOLOv5s		CD-ViT-YOLO	
		mAP@0.5(%)	mAP@0.5:0.95(%)	mAP@0.5(%)	mAP@0.5:0.95(%)
Adequate light (In daylight)	drinking	95.3	86.8	96.8	88.9
	lying	99.5	89.6	99.5	90.1
	standing	96.9	86.6	97.1	89.6
	eating	98.7	94.3	98.9	94.9
	preening	94.5	84.7	96.6	88.4
	spreading	97.8	81.0	97.8	83.9
Medium light (In the evening)	drinking	94.7	86.0	96.2	87.9
	lying	98.6	88.6	99.3	89.0
	standing	95.8	86.0	96.8	89.3
	eating	97.4	93.4	98.5	94.5
	preening	94.0	84.4	95.8	88.0
	spreading	96.9	80.4	97.4	83.4
Insufficient light (At night)	lying	96.4	86.7	98.8	87.6
	standing	95.0	85.1	96.5	87.5
	preening	93.1	81.7	94.4	86.7
	spreading	95.7	78.0	97.3	82.6

it frequently occurs in densely populated areas, where the wings or body are prone to occlusion by nearby individuals, compromising the completeness of the target region.

In order to offer a clearer insight into detection results across different lighting conditions, the schematic diagram of the comparison of detection results between YOLOv5s and CD-ViT-YOLO under different light intensities is presented in Fig. 10. It can be observed that in different lighting conditions, both YOLOv5s and CD-ViT-YOLO can detect behaviours of caged ducks. Notably, when the ambient light intensity decreased, as shown in Fig. 10(b) in the first row, YOLOv5s failed

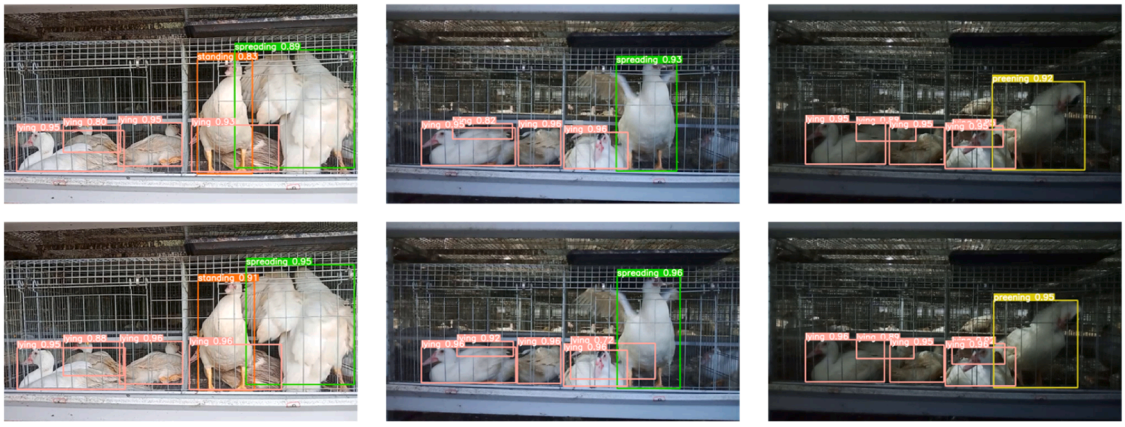

Fig. 10. Comparison of detection results between YOLOv5s and CD-ViT-YOLO under different lighting conditions.(The first row shows the results of YOLOv5s, and the second row shows the detection results of CD-ViT-YOLO).

to detect the "lying" behaviour, resulting in a missed detection. In contrast, CD-ViT-YOLO accurately identified this behaviour. The combination of deformable convolution and the global modelling capability of the Transformer enables CD-ViT-YOLO to adaptively capture spatial variations and contextual information, thereby enhancing its robustness in insufficient-light conditions. This example clearly demonstrates that CD-ViT-YOLO is more robust under varying lighting conditions, while YOLOv5s may struggle to maintain reliable detection performance when lighting is insufficient. Such robustness makes CD-ViT-YOLO particularly well suited for practical deployment in poultry farming environments, where lighting conditions differ greatly between day and night. The improved stability and accuracy of CD-ViT-YOLO across these varying light levels ensure consistent behaviour detection, which is crucial for maintaining animal welfare and enabling timely interventions. This enhanced robustness significantly improves the practicality and effectiveness of deploying CD-ViT-YOLO in real-world agricultural settings, supporting continuous monitoring and better farm management.

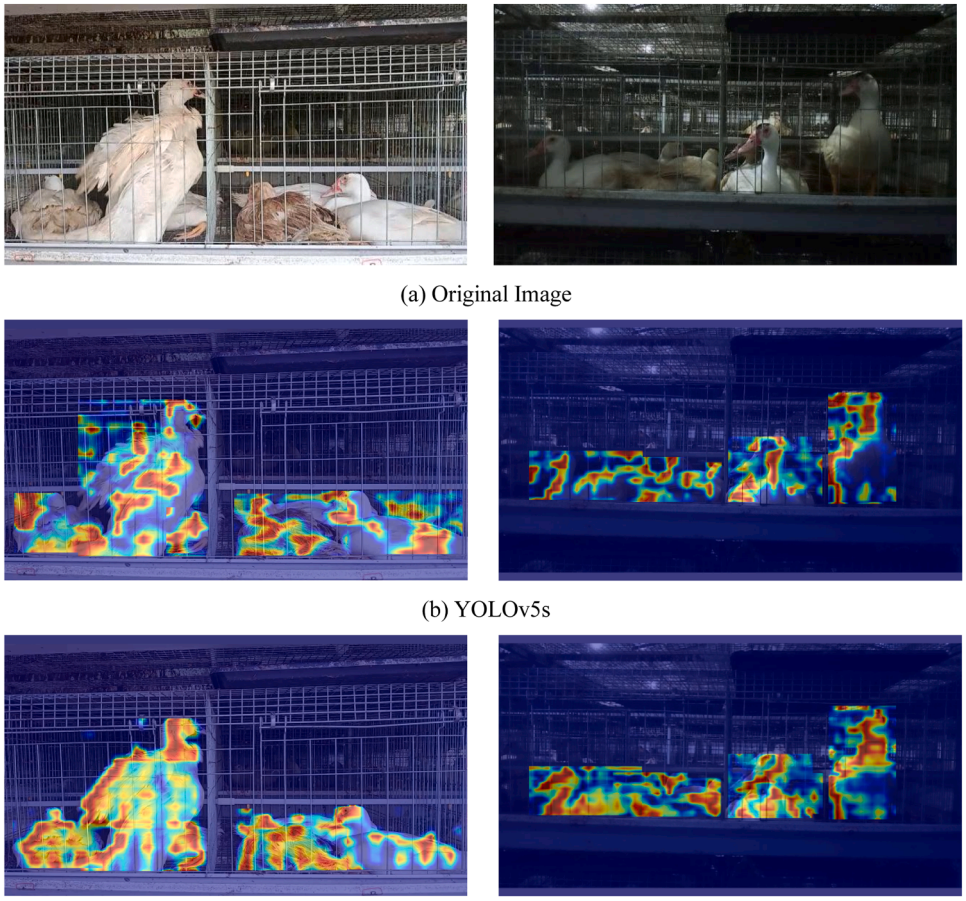

Fig. 11. Comparison of Heatmaps Between YOLOv5s and CD-ViT-YOLO.

The differences in feature representation between the baseline model and the improved model were further explored by visualising the heatmaps of YOLOv5s and CD-ViT-YOLO under identical input conditions, as shown in Fig. 11. The heatmaps of YOLOv5s (Fig. 11(b)) shows relatively dispersed and shallow activation, indicating a limited ability to focus on the key behavioural regions of the ducks. Additionally, irrelevant background regions (such as cage structures) exhibit unnecessary activation, which may mislead the model and interfere with accurate behaviour recognition.

In contrast, the heatmaps generated by CD-ViT-YOLO (Fig. 11(c)) display more concentrated and intense responses in behaviour-related areas. The model effectively captures the body contours and posture features of the ducks while suppressing activation in non-target regions. This phenomenon demonstrates that certain proposed strategies play a role. Specifically, the EfficientViT-m0, as a Transformer-based backbone, brings global information, which is beneficial. Moreover, the DRE block, by using deformable convolution (DCNv4) to endow the model with dynamic receptive fields, is crucial for making the heatmap activation more in line with the duck contours. These components work together to some extent to strengthen the model. Moreover, the heatmaps of CD-ViT-YOLO present clearer and more well-defined target boundaries, indicating more precise spatial localisation and stronger attention to behaviour-related regions. These visualisations confirm that CD-ViT-YOLO exhibits superior localisation performance, particularly under complex cage conditions involving densely distributed ducks and diverse behaviour types.

3.3. Comparison of different models on caged duck behaviour detection performance

The YOLO model has been extensively applied across various detection tasks due to its robust capabilities, leading to the development of numerous versions, each offering unique advantages. Given the limited research on daily behaviour detection for caged ducks, this study primarily compares CD-ViT-YOLO with other models from the YOLO series. To thoroughly evaluate the effectiveness of the CD-ViT-YOLO model for detecting caged duck behaviours, its performance was compared against YOLOv5s, YOLOv6s [21], YOLOv7-tiny [48], YOLOv8s, YOLOv9s [49], YOLOv10s [46], YOLOv11s, YOLOv12s [41] and RT-DETR [63]. All models were trained on the same dataset, following the parameter configurations outlined in Section 2.4. Table 8 presents the detection results of all models on the test set.

As shown in Table 8, CD-ViT-YOLO achieved mAP@0.5 and mAP@0.5:0.95 scores of 97.4 % and 88.6 %, respectively. Compared with YOLOv8s, CD-ViT-YOLO improved mAP@0.5:0.95 by 1.8 percentage points, reduced the number of parameters by approximately 52 %, and cut GFLOPs by 60 %, while maintaining nearly the same inference speed (77.6 FPS vs. 77.5 FPS), demonstrating a better balance of performance and efficiency. Compared with the latest YOLO model, YOLOv12s, CD-ViT-YOLO achieved improvements of 1.5 % and 2.5 % in mAP@0.5 and mAP@0.5:0.95, respectively, while reducing parameters by 48 % and computation by nearly 60 %, indicating marked model lightweight while maintaining high performance. Additionally, CD-ViT-YOLO obtained the highest F1-score (95.1 %), further validating its overall advantages in accuracy and real-time performance.

Besides the YOLO series, this study also compared a Transformer-based detector, RT-DETR. Although RT-DETRv2-S exhibited excellent detection performance, it requires 21.7 M parameters and 57.5 GFLOPs of computation. In contrast, CD-ViT-YOLO consumes only about 15 % of the computational cost, while still achieving 0.3 % and 0.4 % higher mAP@0.5 and mAP@0.5:0.95, respectively, and maintaining a faster inference speed, showing greater advantages in lightweight deployment and real-time detection.

To provide a more intuitive comparison of the models' performance across various aspects, Fig. 12 presents line charts of five evaluation metrics: Precision, Recall, F1-score, mAP@0.5, and mAP@0.5:0.95. As shown in the figure, CD-ViT-YOLO consistently ranks among the top performers across all metrics, particularly excelling in F1-score and mAP@0.5:0.95, demonstrating strong overall detection capability.

In summary, CD-ViT-YOLO achieves excellent detection performance while maintaining a relatively small model size. This advantage primarily results from the integration of a lightweight ViT backbone into the YOLO framework, which enhances global feature extraction. In addition, the replacement of the original C3 block with the proposed DRE block improves detection capability, while the introduction of a DAD head enables the model to better handle behaviour samples of varying complexity. Benefiting from these architectural improvements, CD-ViT-YOLO effectively and accurately recognises caged duck behaviours.

With the rapid development of large language models, vision–language models (VLMs) have been increasingly applied across visual tasks, where the integration of textual information can strengthen semantic alignment and improve generalization. Such characteristics may also benefit fine-grained behaviour recognition, where subtle distinctions between categories are often challenging to capture. In this study, two YOLO-based vision–language detectors, YOLOE-v8s [47] and YOLOv8s-worldv2, are included for comparison. In the experiments, both models use the class names as textual embeddings, thereby promoting a closer alignment between linguistic labels and visual representations. To ensure a fair evaluation, small-scale experiments were conducted with 50 epochs of full fine-tuning (Table 9).

The experimental results show that both YOLOE-v8s and YOLO-Worldv2-v8s achieve competitive performance, with precision, recall, and mAP values better than those of the proposed CD-ViT-YOLO. Specifically, YOLOE-v8s attains the highest precision and mAP@0.5, while YOLOv8s-worldv2 achieves the best recall, reflecting the potential benefits of incorporating textual embeddings into detection. However, CD-ViT-YOLO offers a more favorable trade-off, achieving comparable accuracy with significantly fewer parameters and lower computational cost. These findings suggest that while vision–language models show promise for fine-grained behaviour recognition, lightweight designs such as CD-ViT-YOLO remain advantageous for efficient deployment in

Table 8
The Experimental results of performance comparison of different models.

Model	P (%)	R (%)	F1 (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	#Params(M)	GFLOPs	FPS
YOLOv5s	92.7	94.9	93.8	96.3	86.3	7.02	15.8	76.468
YOLOv6s	92.7	93.4	93.0	94.7	86.5	16.3	44.0	72.463
YOLOv7-tiny	95.0	93.2	94.0	95.7	86.9	6.02	13.2	74.388
YOLOv8s	93.9	93.1	93.5	96.0	86.8	9.87	21.8	77.519
YOLOv9s	92.6	93.2	92.9	96.5	87.6	6.31	22.7	61.349
YOLOv10s	95.1	92.6	93.8	96.2	87.2	8.07	24.8	85.470
YOLOv11s	91.4	95.2	93.2	96.2	87.0	9.41	21.3	69.930
YOLOv12s	92.8	94.3	93.5	95.9	86.1	9.20	21.5	64.935
RT-DETR-r50	95.2	94.6	94.9	97.4	89.3	42.7	130.5	51.032
RT-DETRv2-S	94.5	95.0	94.7	97.1	88.2	21.7	57.5	90.335
CD-ViT-YOLO	94.9	95.1	95.0	97.4	88.6	4.75	8.7	77.568

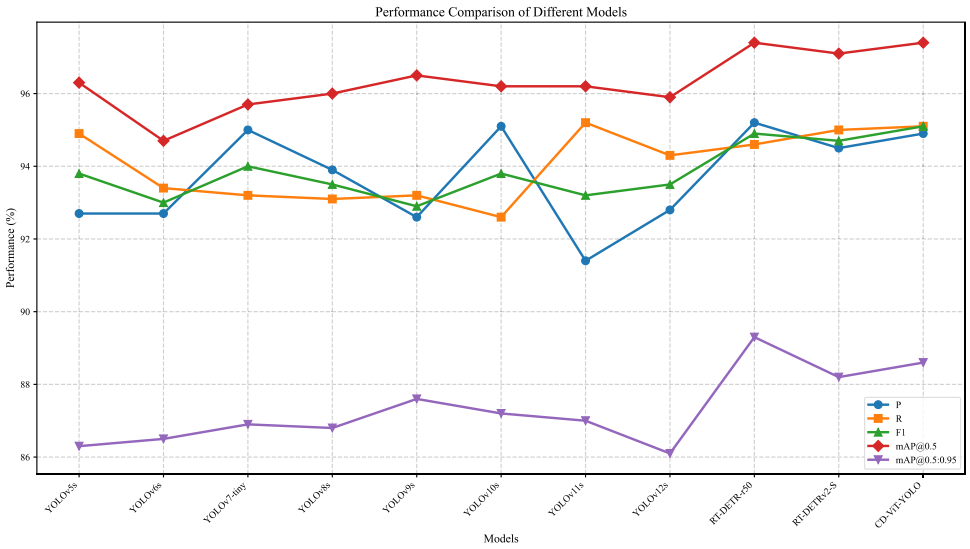

Fig. 12. Comparison of YOLO variants and transformer-based models on key detection metrics.

Table 9
Experimental results of vision–language models.

Model	P (%)	R (%)	F1 (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	#Params(M)	GFLOPs
YOLOE-v8s	96.1	94.7	95.3	97.6	90.9	11.1	28.4
YOLOv8s-worldv2	94.4	95.7	95.0	96.9	90.6	12.7	33.3
CD-ViT-YOLO	94.9	95.1	95.0	97.4	88.6	4.75	8.7

practical applications.

3.4. Comparison of different models on an external animal behaviour dataset

While CD-ViT-YOLO demonstrates strong performance on the self-built DuckBehaviour dataset, its generalisability to other animal behaviour detection scenarios remains untested. To assess the model's robustness beyond the original dataset, CD-ViT-YOLO was evaluated on a publicly available cattle behaviour dataset [23], which contains five behaviour categories: standing, lying, foraging, rumination, and drinking. The experiments were conducted under consistent conditions to ensure a fair comparison. These settings provide supporting evidence for the generalisation ability of CD-ViT-YOLO across different species and environmental conditions.

As shown in Table 10, CD-ViT-YOLO achieves a favourable balance between accuracy and efficiency compared with other YOLO series models. In particular, relative to performance-leading YOLOv10s and YOLOv12s, CD-ViT-YOLO maintains nearly the same level of detection performance while requiring only about half of the parameters and considerably fewer computations. This is because the proposed optimization strategies are specifically designed for behaviour recognition of caged ducks rather than cattles, which is reasonable as the conducted optimization directs towards a specific rather than general research field. Such a balance highlights its potential for deployment in resource-constrained environments.

4. Discussion and limitations

4.1. Discussion

The analysis of existing literature reveals that the research on animal behaviour recognition has predominantly focused on free-range poultry and large livestock [22], with relatively limited studies on the behaviour of caged ducks. However, as caged ducks become increasingly prevalent in modern farming systems, precise monitoring of their behaviour has become crucial, not only to enhance farming efficiency but also to ensure better animal welfare.

With the rapid development of computer vision and deep learning technologies, non-contact behaviour detection methods have increasingly emerged as effective alternatives to traditional contact-based sensor techniques. These non-contact approaches, which utilise video

Table 10
Cross-Dataset performance comparison of CD-ViT-YOLO on cattle behaviour Dataset.

Model	P (%)	R (%)	F1 (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	#Params(M)	GFLOPs
YOLOv5s	91.1	92.0	91.5	92.3	82.6	7.02	15.8
YOLOv6s	92.7	92.4	92.5	93.7	83.7	16.3	44.0
YOLOv7-tiny	92.0	91.2	91.6	92.7	83.2	6.02	13.2
YOLOv8s	92.9	93.1	93.0	93.0	83.7	9.87	21.8
YOLOv9s	93.6	92.2	92.9	93.0	84.6	6.31	22.7
YOLOv10s	93.3	94.0	93.6	94.2	85.2	8.07	24.8
YOLOv11s	93.6	94.2	93.9	93.1	84.3	9.41	21.3
YOLOv12s	93.8	93.9	93.8	93.9	85.1	9.20	21.5
CD-ViT-YOLO	93.4	94.0	93.7	93.5	84.7	4.75	8.7

capture and image processing algorithms, minimise disturbances to animals while enabling efficient and real-time monitoring of behaviours in large-scale farming environments. By observing animal activities without physical interference, these methods offer wider application prospects and greater potential for advancing intelligent livestock management.

The YOLO model has recently gained considerable attention for its strong performance in recognising livestock and poultry behaviours, especially in object detection tasks [24,34]. Nonetheless, much of the prior research has concentrated on extracting local features, often overlooking the critical role of global feature representation in complex behaviour recognition scenarios. Conventional YOLO architectures primarily focus on local information, which can restrict their effectiveness in interpreting scenes with multiple targets, occlusions, or complicated backgrounds. To overcome these challenges, this study introduces the CD-ViT-YOLO model, integrating the EfficientViT-m0 network to improve global feature extraction. This enhancement enables CD-ViT-YOLO to better capture long-distance and cross-target behaviour patterns, thereby boosting detection accuracy and robustness, particularly in the context of caged duck behaviour recognition.

Currently, detecting the behaviour of caged ducks presents several challenges, including mutual occlusion among ducks, low-light conditions, and complex environmental backgrounds. To address these issues, this study proposes the CD-ViT-YOLO model, which is designed to detect six key daily behaviours of caged ducks: standing, drinking, eating, lying, preening, and spreading. These behaviours were selected because they comprehensively reflect the health and physiological status of the ducks, providing critical data for farm management. Ablation experiments demonstrated that the introduction of the DRE block and the EfficientViT-m0 network improved CD-ViT-YOLO's recognition accuracy, resulting in more stable and efficient performance, particularly in complex environments.

Although research on caged duck behaviour remains limited, this study validates the effectiveness of the proposed CD-ViT-YOLO model through comparisons with various representative lightweight and high-performance detectors. CD-ViT-YOLO achieves a good balance between detection performance and model efficiency. This advantage results from the synergy of several architectural innovations: the ViT backbone enhances the ability to capture global contextual information; the DRE block introduces deformable convolution and dynamic receptive fields, improving adaptability to spatial variations in dense scenes; and the DAD Head improves the detection of behaviour categories with fewer samples. These innovations collectively reinforce CD-ViT-YOLO's performance and practicality in caged duck behaviour detection.

4.2. Limitations

Although CD-ViT-YOLO has demonstrated strong performance in detecting caged duck behaviours, there are still several limitations to address. For instance, in low-light conditions such as nighttime, reduced image quality may negatively impact detection accuracy. Furthermore, the boundaries between different behaviours are often ambiguous. Behaviours such as "drinking" and "pecking" share similar visual characteristics, which may lead to confusion during classification. This semantic ambiguity, combined with the unavoidable subjectivity in manual annotations, introduces label noise that could affect model performance. Additionally, in densely populated farming environments, it is challenging to distinguish between individual ducks due to their similar appearances. As a result, the current detection approach primarily operates at the category level rather than the individual level, limiting its applicability in precise health monitoring. Future work could explore the use of image enhancement techniques to improve low-light images, leverage temporal information for behaviour disambiguation, and incorporate object tracking to enable individual-level behaviour recognition, thereby improving the model's real-world utility. In particular, achieving stable real-time inference on resource-constrained edge devices remains a critical challenge for future research.

5. Conclusion

Detecting the behaviour of caged ducks is essential for advancing intelligent farming practices in duck husbandry. This study addresses challenges such as occlusion by introducing CD-ViT-YOLO, an enhanced model based on YOLOv5, specifically designed to detect the daily behaviours of caged ducks. In CD-ViT-YOLO, the DRE block was developed to replace the C3 block in the Neck module of YOLOv5, thereby improving feature extraction from irregularly shaped targets and enhancing detection performance. Additionally, the lightweight and high-performance EfficientViT-m0 network was employed as a substitute for the YOLOv5 backbone, reducing model complexity while maintaining baseline detection capability. Furthermore, a DAD head was designed to adaptively allocate learning focus based on sample difficulty, thereby enhancing the detection of difficult samples. Through the combination of these strategies, the proposed CD-ViT-YOLO model was developed. Extensive simulation experiments were conducted on the self-built dataset DuckBehaviour, which contains various scenarios. The results demonstrate that the CD-ViT-YOLO model accurately detects six daily behaviours of caged ducks in diverse conditions such as occluded instances, and insufficient light environments. With a parameter count of 4.75M and 8.7 GFLOPs, CD-ViT-YOLO achieved an impressive mAP@0.5:0.95 of 88.6 %. Compared with the baseline model, YOLOv5s, CD-ViT-YOLO shows an improvement of 2.3 percentage points in mAP@0.5:0.95, at the cost of 32.3 % parameter count and 44.9 % GFLOPs. When compared with the most recent model, YOLOv12s, CD-ViT-YOLO improves the mAP@0.5 by 1.5 percentage points, with reductions in parameter count and GFLOPs of 48.4 % and 59.5 %, respectively. These results indicate that the proposed CD-ViT-YOLO model provides accurate and efficient detection of the daily behaviours of caged ducks, supporting the development of intelligent farming systems in duck husbandry.

Future research will focus on enhancing the robustness and adaptability of CD-ViT-YOLO in practical scenarios. Key directions include improving image quality under poor lighting conditions, leveraging multimodal data such as infrared or depth information, and incorporating temporal context and object tracking to refine behaviour recognition in dense groups. In addition, optimising the network structure and inference efficiency will be essential for achieving stable, real-time deployment on resource-constrained edge devices. These directions will help advance CD-ViT-YOLO into a more intelligent and practical behaviour monitoring system for poultry farming.

Ethics statement

If this manuscript involves research on animals or humans, it is imperative to disclose all approval details.

If Yes, please provide your text here:All animal protocols used in the present study were approved by the South China Agricultural University Institutional Animal Care and Use Committee (SCAU-10,564).

CRediT authorship contribution statement

Yujin Gong: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Conceptualization. **Gen Zhang:** Writing – original draft, Visualization, Validation, Formal analysis. **Chuntao Wang:** Writing – review & editing, Methodology, Funding acquisition. **Deqin Xiao:** Visualization, Resources, Funding acquisition.

Declaration of competing interest

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

Acknowledgement

This work was supported by the earmarked fund for China Agriculture Research System [CARS-42-13] , the National Natural Science Foundation of China [62172165] and Innovation Team Project in Modern Agricultural Industrial Technology System of Guangdong [2024CXTD28].

Data availability

Data will be made available on request.

References

[1] A. Alameer, S. Buijs, N. O'Connell, L. Dalton, M. Larsen, L. Pedersen, I. Kyriazakis, Automated detection and quantification of contact behaviour in pigs using deep learning, Biosyst. Eng. 224 (2022) 118–130, https://doi.org/10.1016/j.biosystemseng.2022.10.002.

[2] Z.E. Barker, J.A. Vázquez Diosdado, E.A. Codling, N.J. Bell, H.R. Hodges, D. P. Croft, J.R. Amory, Use of novel sensors combining local positioning and acceleration to measure feeding behaviour differences associated with lameness in dairy cattle, J. Dairy. Sci. 101 (7) (2018) 6310–6321, https://doi.org/10.3168/jds.2016-12172.

[3] T.M. Brown-Brandl, G.A. Rohrer, R.A. Eigenberg, Analysis of feeding behaviour of group housed growing–finishing pigs, Comput. Electron. Agric. 96 (2013) 246–252, https://doi.org/10.1016/j.compag.2013.06.002.

[4] H. Cai, J. Li, M. Hu, C. Gan, S. Han, EfficientViT: lightweight multi-scale attention for high-resolution dense prediction, in: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17256–17267, https://doi.org/10.1109/ICCV51070.2023.01587.

[5] C.-L. Chang, B.-X. Xie, R.-Y. Xu, Spatiotemporal analysis using deep learning and fuzzy inference for evaluating broiler activities, Smart Agricul. Technol. 9 (2024) 100534, https://doi.org/10.1016/j.atech.2024.100534.

[6] G.B. Chang, X.P. Liu, H. Chang, G.H. Chen, W.M. Zhao, D.J. Ji, G.S. Hu, Behaviour differentiation between wild Japanese quail, domestic quail, and their first filial generation, Poult. Sci. 88 (6) (2009) 1137–1142, https://doi.org/10.3382/ps.2008-00320.

[7] C. Chen, W. Zhu, J. Steibel, J. Siegford, J. Han, T. Norton, Recognition of feeding behaviour of pigs and determination of feeding time of each pig by a video-based deep learning method, Comput. Electron. Agric. 176 (2020) 105642, https://doi.org/10.1016/j.compag.2020.105642.

[8] J. Chen, S.-H. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, S.-H.G. Chan, Run, don't walk: chasing higher FLOPS for faster neural networks, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12021–12031, https://doi.org/10.1109/CVPR52729.2023.01157.

[9] Cho Aye, C., Zin, T., & Kobayashi, I. (2022). Black cow tracking by using deep learning-based algorithms, 13(12), 1313–1319. https://doi.org/10.24507/icicelb.13.12.1313.

[10] H. Chung, H. Vu, Y. Kim, C.Y. Choi, Subcutaneous temperature monitoring through ear tag for heat stress detection in dairy cows, Biosyst. Eng. 235 (2023) 202–214, https://doi.org/10.1016/j.biosystemseng.2023.10.001.

[11] C. Fang, T. Zhang, H. Zheng, J. Huang, K. Cuan, Pose estimation and behaviour classification of broiler chickens based on deep neural networks, Comput. Electron. Agric. 180 (2021) 105863, https://doi.org/10.1016/j.compag.2020.105863.

[12] Y. Gu, S. Wang, Y. Yan, S. Tang, S. Zhao, Identification and analysis of emergency behaviour of cage-reared laying ducks based on YOLOv5, Agriculture 12 (4) (2022) 485, https://doi.org/10.3390/agriculture12040485.

[13] J. Guo, G. He, H. Deng, W. Fan, L. Xu, L. Cao, S. Gul Hassan, Pigeon cleaning behaviour detection algorithm based on light-weight network, Comput. Electron. Agric. 199 (2022) 107032, https://doi.org/10.1016/j.compag.2022.107032.

[14] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, GhostNet: more features from cheap operations, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1577–1586, https://doi.org/10.1109/CVPR42600.2020.00165.

[15] H. Hao, P. Fang, W. Jiang, X. Sun, L. Wang, H. Wang, Research on laying hens feeding behaviour detection and model visualization based on convolutional neural network, Agriculture 12 (12) (2022), https://doi.org/10.3390/agriculture12122141, 2141–2141.

[16] Q.T. Hoang, C.P.K. Phung, T.N. Bui, T.P.D. Chu, D.T. Tran, Cow behavior monitoring using a multidimensional acceleration sensor and multiclass SVM, Int. J. Mach. Learn. Netwo. Collabor. Eng. 2 (3) (2018) 110–118, https://doi.org/10.30991/ijmlnce.2018v02i03.003.

[17] Q. Hou, C.-Z. Lu, M.-M. Cheng, J. Feng, Conv2Former: a simple transformer-style ConvNet for visual recognition, IEEe Trans. Pattern. Anal. Mach. Intell. (2024) 1–10, https://doi.org/10.1109/tpami.2024.3401450.

[18] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, Q. Le, Searching for MobileNetV3, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1314–1324, https://doi.org/10.1109/ICCV.2019.00140.

[19] D.B. Jensen, M. Toscano, E. van der Heide, M. Grønvig, F. Hakansson, Comparison of strategies for automatic video-based detection of piling behaviour in laying hens, Smart Agric. Technol. 10 (2025) 100745, https://doi.org/10.1016/j.atech.2024.100745.

[20] A. Katharopoulos, A. Vyas, N. Pappas, F. Fleuret, Transformers are RNNs: fast autoregressive transformers with linear attention, in: International Conference on Machine Learning, 2020, https://doi.org/10.48550/arXiv.2006.16236 arXiv: 2006.16236.

[21] Li, C., Li, L., Geng, Y., Jiang, H., Cheng, M., Zhang, B., & Chu, X. (2023). YOLOv6 v3.0: a full-scale reloading. arXiv:2301.05586. https://doi.org/10.48550/arXiv.2301.05586.

[22] D. Li, B. Dai, Y. Li, P. Song, X. Dai, Y. He, W. Shen, IATEFF-YOLO: focus on cow mounting detection during nighttime, Biosyst. Eng. 246 (2024) 54–66, https://doi.org/10.1016/j.biosystemseng.2024.07.017.

[23] K. Li, D. Fan, H. Wu, A. Zhao, A new dataset for video-based cow behavior recognition, Sci. Rep. 14 (2024) 18702, https://doi.org/10.1038/s41598-024-65953-x.

[24] R. Li, B. Dai, Y. Hu, X. Dai, J. Fang, Y. Yin, W. Shen, Multi-behaviour detection of group-housed pigs based on YOLOX and SCTS-SlowFast, Comput. Electron. Agric. 225 (2024) 109286, https://doi.org/10.1016/j.compag.2024.109286.

[25] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, B. Guo, Swin Transformer V2: scaling up capacity and resolution, in: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11999–12009, https://doi.org/10.1109/CVPR52688.2022.01170.

[26] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: hierarchical vision transformer using shifted windows, in: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9992–10002, https://doi.org/10.1109/ICCV48922.2021.00986.

[27] Ma, N., Zhang, X., Zheng, H.-T., & Sun, J. (2018). ShuffleNet V2: practical guidelines for efficient CNN architecture design. *arXiv:1807.11164*. https://arxiv.org/abs/1807.11164.

[28] M.A. Islam, S. Lomax, A.K. Doughty, M.K. Islam, P. Thomson, C. Clark, Revealing the diversity in cattle behavioural response to high environmental heat using accelerometer-based ear tag sensors, Comput. Electron. Agric. 191 (2021) 106511, https://doi.org/10.1016/j.compag.2021.106511.

[29] Mehta, S., & Rastegari, M. (2022). *Separable self-attention for mobile vision transformers. arXiv:2206.02680*. https://arxiv.org/abs/2206.02680.

[30] V.R. Merenda, V.U.C. Bodempudi, M.D. Pairis-Garcia, G. Li, Development and validation of machine-learning models for monitoring individual behaviors in group-housed broiler chickens, Poult. Sci. 103 (12) (2024) 104374, https://doi.org/10.1016/j.psj.2024.104374.

[31] A. Mujahid, M. Furuse, behavioural responses of neonatal chicks exposed to low environmental temperature, Poult. Sci. 88 (5) (2009) 917–922, https://doi.org/10.3382/ps.2008-00472.

[32] S. Neethirajan, Automated tracking systems for the assessment of farmed poultry, Animals 12 (3) (2022) 232, https://doi.org/10.3390/ani12030232.

[33] B. Paneru, R. Bist, X. Yang, L. Chai, Tracking dustbathing behavior of cage-free laying hens with machine vision technologies, Poult. Sci. 103 (12) (2024) 104289, https://doi.org/10.1016/j.psj.2024.104289.

[34] B. Paneru, R. Bist, X. Yang, L. Chai, Tracking perching behaviour of cage-free laying hens with deep learning technologies, Poult. Sci. 103 (12) (2024) 104281, https://doi.org/10.1016/j.psj.2024.104281.

[35] Pratama, Y.P., Kurnia Basuki, D., Sukaridhoto, S., Yusuf, A.A., Yulianus, H., Faruq, F., & Putra, F.B. (2019). Designing of a smart collar for dairy cow behaviour monitoring with application monitoring in microservices and Internet of things-based systems. https://doi.org/10.1109/ELECSYM.2019.8901676.

[36] I. Radosavovic, J. Johnson, S. Xie, W.-Y. Lo, P. Dollar, On network design spaces for visual recognition, in: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1882–1890, https://doi.org/10.1109/ICCV.2019.00197.

[37] I. Radosavovic, R.P. Kosaraju, R. Girshick, K. He, P. Dollár, Designing network design spaces, in: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10425–10433, https://doi.org/10.1109/CVPR42600.2020.01044.

[38] Sachin Subedi, Ramesh Bahadur Bist, X. Yang, L. Chai, Tracking pecking behaviours and damages of cage-free laying hens with machine vision technologies, Comput. Electron. Agric. 204 (2023), https://doi.org/10.1016/j.compag.2022.107545, 107545–107545.

[39] A. Sammad, H. Luo, W. Qiu, J.M. Galindez, Y. Wang, G. Guo, Huang Xixia, Automated monitoring of seasonal and diurnal variation of rumination behaviour: insights into thermotolerance management of Holstein cows, Biosyst. Eng. 223 (2021) 115–128, https://doi.org/10.1016/j.biosystemseng.2021.12.002.

[40] P.R. Shorten, L.B. Hunter, Acoustic sensors for automated detection of cow vocalization duration and type, Comput. Electron. Agric. 208 (2023) 107760, https://doi.org/10.1016/j.compag.2023.107760.

[41] Tian, Y., Ye, Q., & Doermann, D. (2025). *YOLOv12: attention-centric real-time object detectors. arXiv:2502.12524*. https://arxiv.org/abs/2502.12524.

[42] D.D. Tran, N.D. Thanh, Pig health abnormality detection based on behaviour patterns in activity periods using Deep Learning, Int. J. Adv. Comput. Sci. Appl. 14 (5) (2023), https://doi.org/10.14569/ijacsa.2023.0140564.

[43] S. Tu, Y. Cai, Y. Liang, H. Lei, Y. Huang, H. Liu, D. Xiao, Tracking and monitoring of individual pig behaviour based on YOLOv5-byte, Comput. Electron. Agric. 221 (2024) 108997, https://doi.org/10.1016/j.compag.2024.108997.

[44] P.K.A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, A. Ranjan, FastViT: a fast hybrid vision transformer using structural reparameterization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5785–5795.

[45] A. Wang, H. Chen, Z. Lin, J. Han, G. Ding, RepViT: revisiting mobile CNN from ViT perspective, in: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15909–15920, https://doi.org/10.1109/CVPR52733.2024.01506.
[46] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, G. Ding, YOLOv10: real-time end-to-end object detection, in: Proceedings of the Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS), 2024.
[47] Wang, A., Liu, L., Chen, H., Lin, Z., Han, J., & Ding, G. (2025). YOLOE: real-time seeing anything. *arXiv:2503.07465*. https://arxiv.org/abs/2503.07465.
[48] C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao, YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7464–7475, https://doi.org/10.1109/CVPR52729.2023.00721.
[49] Wang, C.-Y., Yeh, I.-H., & Liao, H.-Y.M. (2024). YOLOv9: learning what you want to learn using programmable gradient information. *arXiv:2402.13616*. https://arxiv.org/abs/2402.13616.
[50] J. Wang, Y. Zhang, J. Wang, K. Zhao, X. Li, B. Liu, Using machine-learning technique for estrus onset detection in dairy cows from acceleration and location data acquired by a neck-tag, Biosyst. Eng. 214 (2022) 193–206, https://doi.org/10.1016/j.biosystemseng.2021.12.025.
[51] R. Wang, Q. Bai, R. Gao, Q. Li, C. Zhao, S. Li, H. Zhang, Oestrus detection in dairy cows by using atrous spatial pyramid and attention mechanism, Biosyst. Eng. 223 (2022) 259–276, https://doi.org/10.1016/j.biosystemseng.2022.08.018.
[52] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid vision transformer: a versatile backbone for dense prediction without convolutions, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 548–558, https://doi.org/10.1109/ICCV48922.2021.00061.
[53] D. Wu, Y. Wang, M. Han, L. Song, Y. Shang, X. Zhang, H. Song, Using a CNN-LSTM for basic behaviours detection of a single dairy cow in a complex environment, Comput. Electron. Agric. 182 (2021) 106016, https://doi.org/10.1016/j.compag.2021.106016.
[54] D. Xiao, H. Wang, Y. Liu, W. Li, H. Li, DHSW-YOLO: a duck flock daily behaviour recognition model adaptable to bright and dark conditions, Comput. Electron. Agric. 225 (2024) 109281, https://doi.org/10.1016/j.compag.2024.109281.
[55] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: simple and efficient design for semantic segmentation with transformers, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, J.Wortman Vaughan (Eds.), Advances in Neural Information Processing Systems (Vol. 34, Curran Associates, Inc, 2021, pp. 12077–12090, in: https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf.
[56] Y. Xiong, Z. Li, Y. Chen, F. Wang, X. Zhu, J. Luo, W. Wang, T. Lu, H. Li, Y. Qiao, L. Lu, J. Zhou, J. Dai, Efficient deformable ConvNets: rethinking dynamic and sparse operator for vision applications, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 5652–5661.
[57] J. Xu, J. Ye, S. Zhou, A. Xu, Automatic quantification and assessment of grouped pig movement using the XGBoost and YOLOv5s models, Biosyst. Eng. 230 (2023) 145–158, https://doi.org/10.1016/j.biosystemseng.2023.04.010.
[58] W. Xu, T. Xu, J.Alex Thomasson, W. Chen, R. Karthikeyan, G. Tian …, Q. Su, A lightweight SSV2-YOLO based model for detection of sugarcane aphids in unstructured natural environments, Comput. Electron. Agric. 211 (2023) 107961, https://doi.org/10.1016/j.compag.2023.107961.
[59] X. Yin, D. Wu, Y. Shang, B. Jiang, H. Song, Using an EfficientNet-LSTM for the recognition of single Cow's motion behaviours in a complicated environment, Comput. Electron. Agric. 177 (2020) 105707, https://doi.org/10.1016/j.compag.2020.105707.
[60] N. Zehner, J.J. Niederhauser, M. Schick, C. Umstatter, Development and validation of a predictive model for calving time based on sensor measurements of ingestive behaviour in dairy cows, Comput. Electron. Agric. 161 (2019) 62–71, https://doi.org/10.1016/j.compag.2018.08.037.
[61] K. Zhao, Y. Duan, J. Chen, Q. Li, X. Hong, R. Zhang, M. Wang, Detection of Respiratory rate of dairy cows based on infrared thermography and deep learning, Agriculture 13 (10) (2023) 1939, https://doi.org/10.3390/agriculture13101939.
[62] S. Zhao, Z. Bai, L. Meng, G. Han, E. Duan, Pose estimation and behaviour classification of Jinling White Duck based on improved HRNet, Animals 13 (18) (2023) 2878, https://doi.org/10.3390/ani13182878.
[63] 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. arXiv:2304.08069*. https://arxiv.org/abs/2304.08069.
[64] Z. Zheng, L. Qin, PrunedYOLO-Tracker: an efficient multi-cows basic behaviour recognition and tracking technique, Comput. Electron. Agric. 213 (2023) 108172, https://doi.org/10.1016/j.compag.2023.108172.
[65] X. Zhu, C. Chen, B. Zheng, X. Yang, H. Gan, C. Zheng …, Y. Xue, Automatic recognition of lactating sow postures by refined two-stream RGB-D faster R-CNN, Biosyst. Eng. 189 (2020) 116–132, https://doi.org/10.1016/j.biosystemseng.2019.11.013.