Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Bird and UAVs Recognition Detection and Tracking Based on Improved YOLOv9-DeepSORT

JINCAN ZHU[1], CHENHAO MA[2], JIAN RONG, [3], YONG CAO[4]
[1]College of Big Data and Intelligent Engineering, Southwest Forestry University,Kunming650000,China (e-mail: 17872257350@163.com)

Corresponding author: Jian Rong (e-mail: swordrong@163.com).

This work was supported by the Key Laboratory of State Forestry and Grassland Administration on Forestry Ecological Big Data, Southwest Forestry University(No. 2022-BDG-09).

ABSTRACT At present, the protection of birds, especially endangered birds, faces major challenges. In the process of protection, birds are often mixed with various drones, and it is difficult to accurately count the number of endangered birds, which brings great difficulties to bird protection work. So tracking and identifying birds and drones is crucial. To solve these problems, this paper proposes a new multi-target tracking (MOT) model based on the combination of YOLOv9 detection algorithm and DeepSORT tracking algorithm. Firstly, the original RepNSCPELAN4 module is replaced by CAM context feature enhancement module in Backbone to improve the model's ability to extract small target features. Following this, the AFF channel attention mechanism has been integrated with RepNSCPELAN4 in the Head section to create the Repnscpelan4-AFF module, which aims to better address semantic and scale inconsistencies. Finally, a new RepNSCPELAN4-Akconv module has been developed using the AKConv dynamic Convolution module to replace the RepNSCPELAN4 module in the original Head section, enabling the model to more effectively capture detailed and contextual information. On the Bird-Drone visible light comprehensive data set proposed in this study, the improved YOLOv9-DeepSORT model has a mAP0.50 of 81.3% for all categories and 89.1% for individual birds. Compared to the baseline YOLOv9 original model, improvements of 7.9% and 23.9% respectively. On infrared datasets, compared to the original model, the mAP0.50 of the improved model is improved by 3.2% in all categories. The accuracy of identifying individual birds and similarly shaped fixed-wing drones also improved by 2.2% and 7.5% respectively. Moreover, on the mixed visible light and infrared data sets, the model get mAP0.50 of 81.8% higher 0.9% than that of the YOLOv9. These experiments demonstrate the improved YOLOv9-DeepSORT method can expand the multiscene application range of bird recognition and tracking models, effectively promoting the extraction of video frame features in multi-target tracking.

INDEX TERMS Bird protection, AKConv Dynamic convolution, AFF channel attention, CAM feature enhancement, YOLOv9 and DeepSORT.

I. INTRODUCTION

WILD birds are very important for maintaining the ecological balance. In recent years, more and more attention has been paid to the research on bird protection, for example [1], [2], [3], [4]. It faces many challenges, including habitat destruction, climate change, illegal fishing and trade, pollution, invasive species and human activities.

Combined with recent studies, it is clear that the monitoring [5], [9] and population statistics [6], [7]of wild birds are important research direction of bird protection. This is closely linked to technology, particularly the integration of ecological studies with computer vision techniques. Monitoring birds can provide insights into their behavioral patterns [8], reproductive tendencies, and dietary preferences, which are crucial for studying the ecological characteristics of bird populations. Additionally, bird monitoring helps identify rare or endangered species and provides scientific basis for formulating effective conservation strategies [11].

In order to assist in the more rapid and accurate monitoring and counting of wild birds by organizations and institutions, especially in distinguishing drones mixed within bird flocks at long distances, image recognition and tracking technology [17] is particularly important.

At present, detectionV - track model [18] is the most effective monitoring mode, this is largely thanks to the continuous development of target detection algorithms such as YOLO.

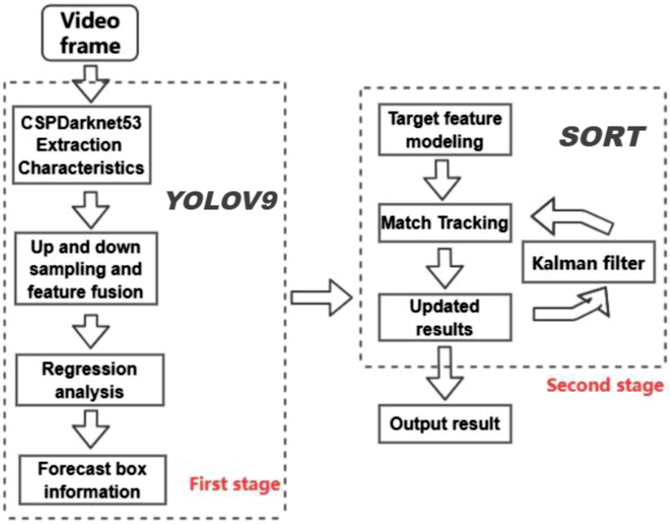


FIGURE 1. The general detection and tracking framework is similar to the combination of the YOLOv9 detection algorithm and the SORT tracking algorithm as illustrated in the figure.

The general pattern of this model is similar to YOLOv9-DeepSORT, and the model described in Figure 1 is a representative example. Generally speaking, the detection and tracking mode consists of two successive stages: the first stage is the target identification and detection stage, and the second stage is the real-time target tracking stage. The first stage usually uses detection methods such as YOLO, ResNet [22], [23] to identify the category of the target and provide specific location information for subsequent stages. In the second stage, after identifying the target position in the previous frame and the next frame of video, the motion trajectory is generated by the correlation algorithm.

However, real-time detection and tracking of airborne birds and drone targets is a major challenge, both now and in the future. There are two significant limitations. First, previous YOLO methods have primarily focused on the detection of a sole bird target, and the data sets mainly included medium and large targets. As an illustrative examples [19]–[21], although the recognition classification accuracy rate can reach more than 90%, but the dataset has only large bird targets and does not take into account interference factors such as drones. Therefore, recognition algorithms based on the YOLO series have not been widely utilized for distinguishing between birds and drone targets in academic research. Secondly, the existing birds and UAVs comprehensive data sets lack remote small target data and multi-attitude data. For example, birds and drones exhibit two different postures [24], [25]: stationary and flying. In particular, the attitude of small fixed-wing drones and birds when flying over considerable distances is strikingly similar, and current methods are not effective enough to clearly distinguish between the two. These factors have a significant impact on the ability to distinguish between birds and drone targets, potentially leading to sub-optimal recognition results.

A review of existing research indicates that there have been few studies that integrate object detection and tracking into a unified model for the identification and tracking of birds and UAVs. Moreover, these studies are conducted independently, without a comprehensive integration of the two methodologies. For instance, Chen et al [26] used YOLOv7 in combination with DeepSORT, adding three GAM modules and Alpha-IoU loss function to achieve better accuracy and bird tracking, with an accuracy of more than 90% on a sole class datasets. Sun et al [27]addressed the issue of inconspicuous sole-frame targets and small target sizes in surveillance videos, proposing a motion information-based algorithm for the detection and localization of birds (FBOD-BMI). Xing et al [28] integrated an efficient target tracker based on the detection module of YOLOv5 to update the target state. They then applied a UAVs classification model to the output of the detection and tracking mechanism, with the objective of further distinguishing UAVs from other background distractions (birds, balloons). Samadzadegan et al [29] employed a deep learning model based on YOLOv4 for recognition and achieved an accuracy of 83% in a dataset comprising three categories: multi-rotor UAVs, helicopters, and birds. Dolph, Chester V, and colleagues [30] also developed an image processing-based aerial object detection and tracking system that combined convolutional neural networks (CNNs) with general aviation aircraft, multi-rotor small unmanned aircraft systems (SUAS), fixed-wing SUAS, and bird classification. This approach led to the creation of improved vision-based perception algorithms, with a cross-validated classification accuracy of 74.4% for both aircraft and birds. Although all of the aforementioned studies employed deep learning algorithms in the field of bird detection, it is important to note that there are certain limitations and shortcomings associated with these methods. For instance, studies related to this topic [26], [29]–[31] have achieved notable recognition results. However, the dataset is limited in terms of the number of categories of birds and drones, the inclusion of interference factors, and the number of categories. And [26] has only a sole category, the models have clear limitations in terms of generalization and portability;Subsequent research has sought to enhance the capabilities of YOLO and other algorithms like [27], [28], [32], with the integration of recognition and tracking algorithms representing a significant advance. Despite these enhancements, these model's performance in terms of recognition accuracy and multi-category tracking still falls short of the current requirements; Recently, studies like [33]–[36]have been conducted on birds in datasets More classification, and the recognition accuracy of more than 70%. It is a pity that the identification process is not integrated with track and trace algorithms, which prevents simultaneous identification and monitoring. In light of the limitations identified in previous studies, this study expanded the data in both visible light and infrared scenarios, collecting four types of dimensional data, including birds and drones. The seamless combination of YOLOv9 and DeepSORT enables the system to achieve real-time detection and tracking with impressive accuracy and minimal latency. This integration effectively meets the need to identify and track bird targets in different scenarios.

To further address the aforementioned issues, this research propose the utilization of the most recent iteration of the YOLOv9 [37] architecture, in conjunction with the DeepSORT [38] trace model for bird target recognition and tracking. The current iteration of YOLOv9 has undergone significant enhancements to its model framework, resulting in a markedly improved recognition efficacy. From the perspective of time and space, the fact that both birds and drones have a similar flying attitude allows for the addition of DeepSORT tracking models, which can be used to track targets in real time for further monitoring. In order to further extract the texture features of long-distance birds and drones, we extend the AKConv variable kernel convolutional network [39]. The experimental results demonstrate that AKConv nucleation convolution is capable of identifying birds with greater clarity, offering significant advantages. Secondly, we extend the CAM context feature enhancement module [40] to enhance the location of important feature information. By further locating important feature information, we can obtain better attitude features, thus improving the recognition accuracy. Finally, we introduce AFF channel attention mechanism [41] to address semantic inconsistencies, scale inconsistencies, and the lack of identifiable features for small targets. The primary contributions of this study are as follows:

(1)A bird recognition and tracking model was constructed based on the proposed backbone network, combined with YOLOv9 and DeepSORT. The model is capable of identifying, locating, and tracking targets such as birds and drones through direct input of video and extraction of video frames, without the necessity of pre-detection or subsequent data association.

(2)The research integrates the new AKConv variable kernel convolution kernel, CAM context feature enhancement module, and AFF channel attention mechanism to maximize the influence of dynamic convolution and attention mechanism on feature extraction. At the same time, the feature enhancement module is used to improve the recognition accuracy of the model, especially when detecting small targets at a long distance. Moreover, a substantial number of comparative ablation experiments have been conducted on methods such as convolution modules and attention mechanisms, with the objective of providing a reference point for subsequent researchers.

(3)Based on the diverse poses and angles of birds and drones, a novel visible light dataset was constructed. The dataset encompasses images of numerous stationary and moving birds, fixed- and rotor-wing drones, small target birds, and long-range drones. Simultaneously, to address the deficiency of infrared datasets, this study compiled an infrared dataset of 10, 614 images comprising birds, rotorcraft, helicopters, and fixed-wing aircraft, covering a variety of small targets.

(4)The study integrated infrared and visible light data into a complete dataset and performed training and validation. Thus, the model does not need to distinguish whether the input data is infrared or visible light and can conduct real-time inference and recognition. This further solves the problem of limited identification of model application scenarios and enhances the generalization ability of the model.

II. MATERIALS AND METHODS

A. DATA SET

Two datasets were used in this study: a homemade visible light dataset and an infrared dataset described as follows:

(1)The homemade visible light dataset comprised four categories: static birds (bird), flying birds (flybird), rotorcraft UAVs, and fixed-wing drones. The dataset comprised a total of 8, 978 images of varying dimensions, accompanied by 8, 978 bounding boxes and labelled files. As illustrated in Figure 2, the distribution of the number of images in our dataset acrossed the four categories—fixed-wing UAVs, rotorcraft UAVs, bird, and Fiybird—is as follows: 41.6%, 20.8%, 12.8%, and 24.8% , respectively. Additionally, there were approximately 1800 small target images, and the sum of all bird images was close to 5000. The entire dataset was divided into three subsets: a training set comprising 8, 414 images, a validation set comprising 413 images, and a test set comprising 151 images. In comparison to the training set, the test set contained a higher percentage of multi-target images and was more realistic. Our homemade dataset as a whole was capable of meeting the demand for bird recognition under different conditions, such as small targets, long distances, multiple categories, and multiple scenarios. Its utility was superior to that of the general sole-bird dataset and large-target dataset.

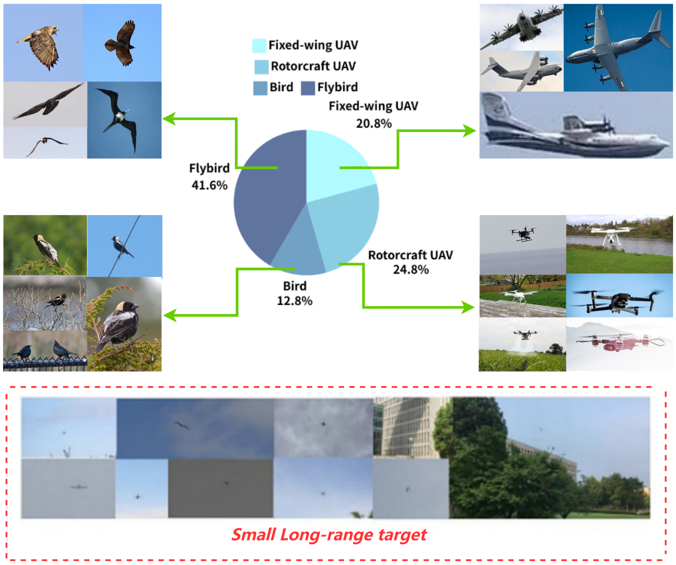


FIGURE 2. The data distribution for homemade visible light dataset is presented above. It includes a total of 8, 978 images of various sizes, approximately 5, 000 images of all birds, and approximately 1, 800 images of small targets.

(2)The infrared dataset produced in this study contains 10, 614 images of different sizes and angles, grouped into four categories: birds, fixed-wing drones, rotorcraft drones, and helicopters. As shown in Figure 3, the subjects of the dataset were birds and common rotorcraft drones. There were 3, 411 images of birds, 4, 466 images of rotorcraft, 1, 655 images of helicopters and 1, 082 images of fixed-wing aircraft. The four categories accounted for 42.1%, 32.1%, 15.6% and 10.2% respectively. Nearly 5500 small target images of birds and various types of aircraft over long distances were collected in the data set, which fully met the demand for small target detection in night scenes, and further verified the improved model's ability to detect and track birds in infrared scenes is further validated.

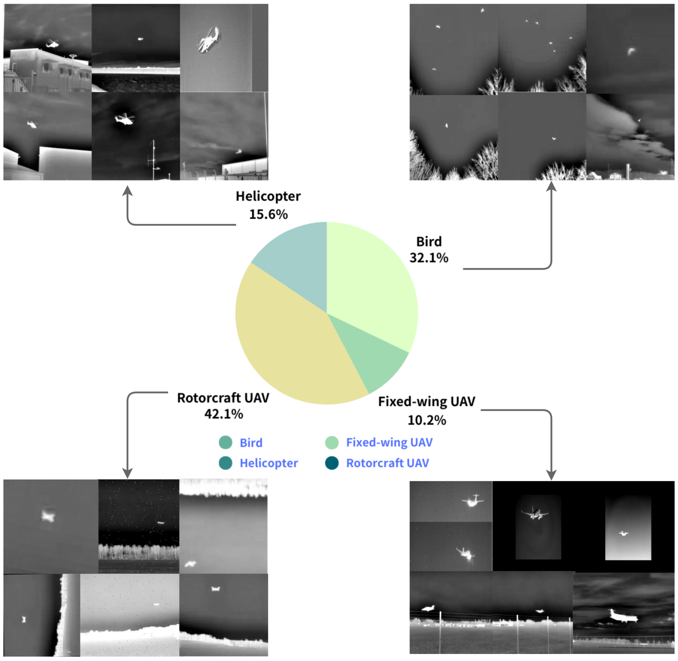


FIGURE 3. The data distribution of the infrared dataset is shown above. In total, it includes 10, 614 images of different sizes, and nearly 5, 500 small target images of distant birds and various types of aircraft. The number of images of birds, rotorcraft, helicopters and fixed-wing aircraft was 3,411, 4,466, 1,655 and 1,082 respectively.

B. IMPROVED YOLOV9 TARGET DETECTION NETWORK

1) The Dialogram of YOLOV9-CAM-AFF-AKConv

YOLOv9 is a new type of high-precision identification model. Its innovative design on the overall architecture makes it perform well in high-precision identification tasks. The main components of the original model include a Backbone Network (RepNSCPELAN4 [42]) and a decoupled Head Structured. Based on this architecture, the complete model of our improved YOLOv9 is shown in Figure 4. The figure demonstrates three significant improvements, including incorporating a CAM context enhancement module in the backbone network, a RepNCSPELAN4-AFF module that fuses AFF attention in the middle layer of the head, and a RepNCSPELAN4-AKConv module that fuses AKConv variable kernel convolution in the last layer of the head. The enhancement of the three modules markedly enhances the model's capacity for generalization, particularly with regard to the extraction of features from small bird targets. This is beneficial for the detection of multiple birds in flight. Furthermore, the improved model demonstrates a notable improvement in accuracy without a commensurate reduction in inference speed, thereby conferring an unparalleled efficiency advantage for real-time object tracking.

(1)The design of its main RepNSCPELAN4 module combines elements of YOLOv5's CSPNet Block [43], YOLOv6's Rep module, and YOLOv7's ELAN module [44]. As shown in Figure 5, It has multiple instances of RepNCSP and Conv module. ReoNCSP refers to the C3 and C2f modules, and integrates the Conv and RepNBottleneck modules. The RepNBottleneck module [45] is a linear combination of RepConvN and Conv module. RepConvN combines two parallel CB and one BN modules to produce the final output via the SILU activation function [46].

(2)Then, the main part of the YOLOv9 head, shown in Figure 6, consists of Conv layers, RepNSCPELAN4 modules, and CB Fuse modules arranged sequentially. This structure enhances feature extraction and representation, which can provide more accurate target detection, bounding box delineation, and classification. The Head-Detect structure remains similar to YOLOv8 [47], with two parallel branches for bounding box error (Reg) and categorization error (Cls). Each branch includes two CBL modules and a Conv layer. As illustrated in Figure 7, for each anchor box, Reg determines the dimensions h and w based on the coordinates (dx, dy, dw, dh). Cls is also calculated based on h and w, this in turn leads to a classification loss. By constantly adjusting the values of width and height (dw, dh), you can more accurately fit the shape of the target object.

2) AKConv dynamic convolutional network

Convolutional networks can be categorized as static or dynamic. Static convolutional neural networks maintain a consistent structure and parameters throughout both the training and inference processes, while dynamic convolutional neural networks allow for the adjustment of the network's structure and parameters based on the characteristics of input data during training and inference. Currently, research on dynamic convolutional networks has made significant progress. In 2019, Liao et al. proposed a variable kernel convolutional network called DKCNN [48], which enables adaptive adjustments to the size and shape of the convolutional kernel to accommodate features of different scales and shapes, overcoming limitations associated with fixed convolutional kernels in traditional CNNs. Subsequently, in 2023, Zhang et al. introduced an innovative variable convolution approach known as AKConv [39], which addresses inherent limitations associated with traditional fixed sample shapes in convolution operations by allowing feature extraction using any number of parameters(e.g., 1, 2, 3, 4, 5, 6, 7, etc.). This advancement has elevated recognition effectiveness to new heights. Additionally, lightweight convolution such as GSConv [49], VoVGSCSP [50] also strikes a good balance between speed

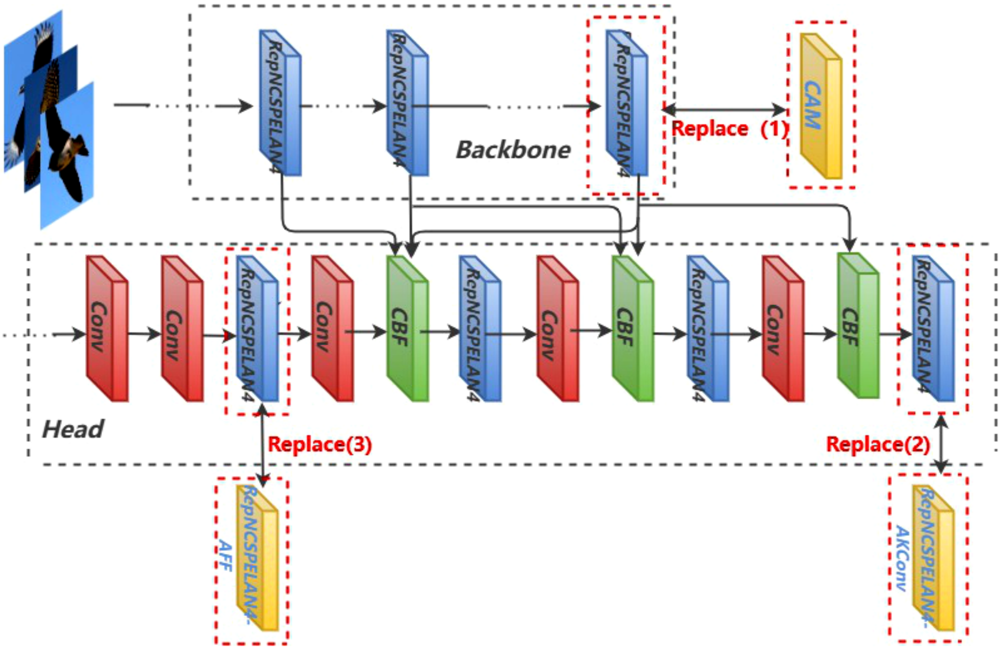


FIGURE 4. The diagram illustrates the integration of the CAM context enhancement module into the final layer of the backbone network, the incorporation of the AFF attention mechanism into the middle layer of the head, and the application of a fusion method with the AKConv variable kernel volume in the final layer of the head.

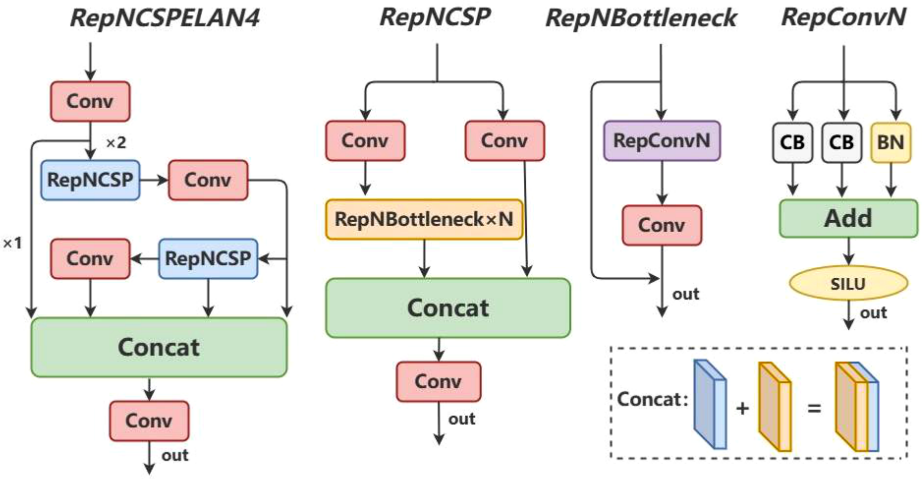


FIGURE 5. Positioned from left to right, RepNSCPELAN4 includes Conv and RepNCSP. The RepNCSP module comprises the Conv and RepNBottleneck modules. The RepNBottleneck module is a direct linear mixture of RepConvN and Conv module. Its submodule, RepConvN, integrates CB and BN modules and provides output via SILU activation functions.

and accuracy. The above methods are improved on the basis of the basic convolutional architecture, and their effectiveness has been continuously verified, providing valuable insights for this study. Considering the specific characteristics of dataset, integrating dynamic variable kernel convolution, especially AKConv convolution network, into detection framework is a very practical approach.

The general form of the dynamic Convolution Integral can be expressed as follows:

$$(K * f)(x) = \int_{\infty}^{\infty} K(x,t)f(t)dt \tag{1}$$

where K(x, t) is a kernel function that depends on x and t, and f(t) is the signal or function to be processed.

As shown in Figure 8, AKConv consists of four main stages: initialization of the sample shape, Conv2d convolution operations, migration, and resampling. The key to AKConv is to adjust the initial sample shape by learning the offset.

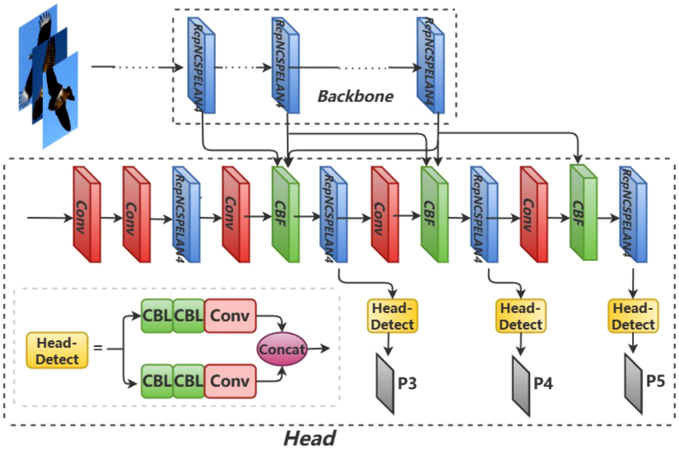


FIGURE 6. The main structure of the Head consists of the connection of multiple Conv convolution layers, the RepNSCPELAN4 module and the CB Fuse module. These modules are used to generate higher-level feature representations, perform cross-stage efficient layer aggregation, and fuse feature maps at different scales.

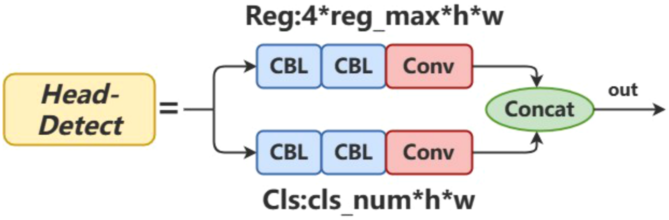


FIGURE 7. YOLOv9 uses two parallel branches on the Head-detect detection structure, one for computing bounding box errors (Reg) and the other for computing classification errors (Cls). Each branch consists of two CBL modules and a convolutional layer.

This allows the convolution kernel to be dynamically adjusted based on the local characteristics of the current input data. Finally, in the training process, the weights and parameters of the AKConv layer are updated through optimization methods such as backpropagation algorithm and gradient descent, so as to minimize the prediction error. According to continuous experiments, we find that replacing the original Conv module with the AKConv module in the last layer of the Head section worked best. Compared with the original RepNCSPELAN4 module, the replaced RepNCSPELAN4-AKConv module has better adaptability to various time series modes and stronger robustness in the presence of noise.

3) CAM feature enhancement module

Feature enhancement modules play a crucial role in deep learning models. Their main function is to refine and enhance features extracted from the backbone. By integrating features of different levels, introducing attention mechanism, and enhancing semantics, the classification and localization capabilities of the model are improved. At present, the mainstream feature enhancement modules include RR Selvaraju

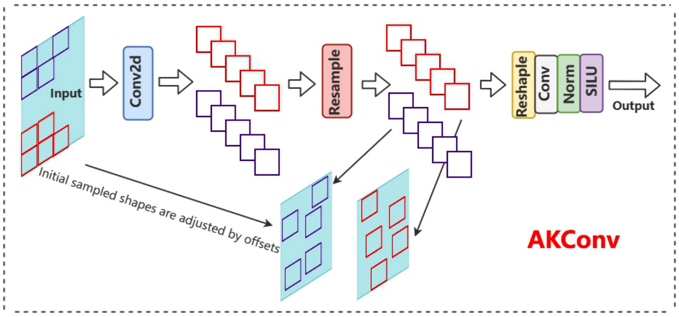


FIGURE 8. AKConv goes through four key steps: initializing the sample shape, Conv2d convolution operation, offsetting, and resampling. Among them, adjusting the initial sample shape by learning the obtained offset is the key of AKConv.

CAM context enhancement module based on CNN network. The module uses convolution at different rates to capture context information from different receiving domains, while enhancing the representation of small targets in combination with the copy-reduce-paste data enhancement technique [51]. The collected data is then integrated from top to bottom into the Feature Pyramid Network (FPN) [52]. This combination of additional contextual information enhances the model's ability to accurately detect and identify targets, especially when dealing with small targets or complex backgrounds. Later, the emergence of feature refinement networks (FRNets) [53] has addressed the limitations of traditional refinement models by providing a fixed representation of features regardless of context. Similarly, Li et al. proposed a novel feature-improved context-aware network (FECANet) [54], which integrates a feature enhancement module and a related recombination module. This method solves the problem of noise and contextual semantic information loss in feature extraction effectively. The results show that different types of functional enhancement modules have different design emphases. Because the focus of this study is to improve the efficiency of small target recognition. Therefore, consider integrating the CAM context enhancement module into the detection framework, which is expected to improve the recognition performance of remote birds.

The framework of CAM modules consists of two main parts: a context-aware module for enhancing CAM and a feature refinement network. As shown in Figure 9, CAM applies convolution to the C5 module at different rates, enabling the capture of context information from different acceptance domains. This method helps to identify key feature areas in the image. The convolution kernel size is 3×3, and the convolution rates are 1, 3, and 5. CAM uses three feature fusion methods :Weight, Adaptive and Concat. Among them, the accuracy of adaptive fusion method is improved the most. According to a large number of experiments, we found that integrating the CAM-based adaptive fusion method into the last layer of the YOLOv9 network backbone can effectively enhance the feature extraction capability of the model, especially for the detection of small targets.

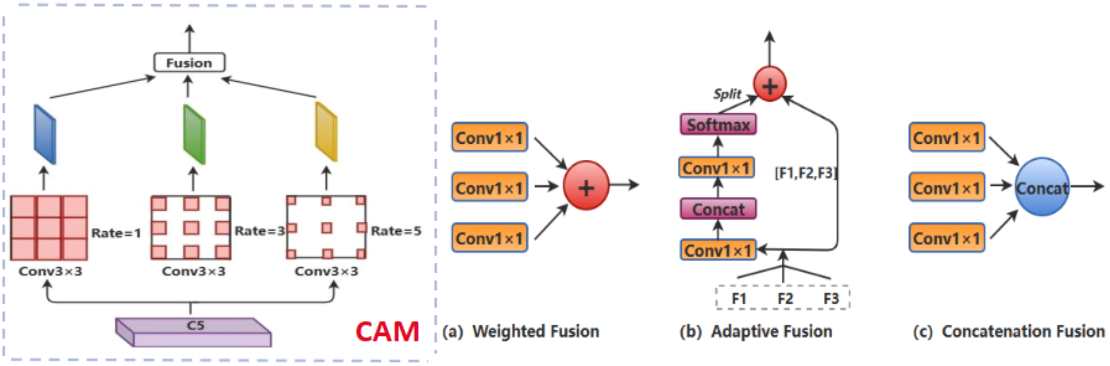


FIGURE 9. The structure:features of the CAM are processed for null convolution at rates of 1, 3, and 5, respectively. Weighted fusion and splicing operations are employed for methods (a) and (c), adding feature mappings directly to the spatial and channel dimensions. Method (b) utilizes an adaptive fusion approach.

4) AFF attention mechanism

The attention mechanism can simulate the workings of the human visual system, allowing the model to prioritize more important and relevant key information when processing input data. At present, most models adopt channel attention mechanism and spatial attention mechanism. In comparison to the spatial attention mechanism, the channel attention mechanism is capable of autonomously discerning the significance of each channel and adjusting the weighting of the feature map, thereby enabling the model to more effectively extract and utilise the feature information. The mainstream channel attention mechanisms include CBAM [55], SE [56], SKNet [57], AFF [41] and so on. The CBAM attention mechanism, proposed by Sanghyun Woo, addresses the limitations of traditional convolutional neural networks in processing information of varying scales, shapes, and orientations by integrating both channel attention mechanism and spatial attention mechanism. On the contrary, Zhou et al. 's proposed implementation of AFF is relatively straightforward and intuitive. It does not necessitate complex calculations or parameter adjustments and can be effectively applied to intricate multi-classification tasks. This makes it particularly accessible to researchers with limited computational resources. Therefore, according to the computational power and data characteristics of this experiment, we propose to integrate the category information capability of AFF attention mechanism in this experiment to filter the respective feature mappings.

As illustrated in Figure 10, the AFF Attention Mechanism presents a fundamental framework for integrating diverse features through the Multiscale Channel Attention Module (MS-CAM). Lower-order feature graph X and higher-order semantic feature graph Y in the higher-order feature pyramid are fused by MS-CAM and other operations to produce output Z. MS-CAM [41]continues SENet's idea of combining local and global features on convolutional neural networks (CNNS). The output of the fused features is denoted by C×H×W, and the symbol "+" represents feature integration. This structure allows the network to perform different selection or weighted averaging between X and Y, effectively addressing semantic and scale inconsistencies. In real-world detection, the fusion of AFF with convolutional modules can better adapt to various input data and task requirements, thereby enhancing their generalization capabilities. Therefore, integrating the AFF attention mechanism with RepNCSPELAN4, a key convolution module in YOLOv9, is considered a key advance. The experimental results show that adding RepNCSPELAN4-AFF module in the middle layer of the head can enhance the detection ability of small target objects.

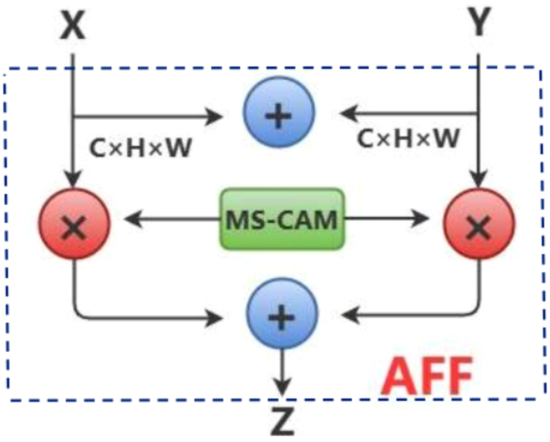


FIGURE 10. The AFF attention mechanism is structured to produce an output Z by fusing the low-feature map X and the higher-order semantic feature map Y in the high-level feature pyramid through MS-CAM and other operations.

C. MULTIPLE TARGET TRACKING BASED ON DEEPSORT

Multi-object tracking (MOT) [58] is an important topic in the field of artificial visual intelligence. As part of this process, the tracking algorithm uses a Kalman filter in each frame of the image to predict the position and speed of the target object. It then matches the target with the observation through the Hungarian algorithm. This iterative process updates the location and speed information of the target, enabling real-time tracking of multiple targets. With continuous advancements in research across various fields, tracking algorithms have made significant breakthroughs in processing complex scenes and high-speed motion. In recent years, researchers have introduced a series of classic end-to-end target tracking algorithms such as Siamese network [59] and Bot-SORT network [60] to achieve end-to-end training and reasoning from raw video data to target tracking results;And the dynamic model of the target trajectory built by Anton Milan and his colleagues using trajectory modeling and probabilistic reasoning. In their study, the researchers employed various techniques, including Kalman filter [61] and particle filter [62], as well as other techniques to enhance target tracking. After the aforementioned major technological advancements, two new and effective tracking algorithms have recently been proposed. YOLO-SORT [38] proposed by Joseph Redmon et al and MotionTrack [63] proposed by Zhengdeng et al are new and effective tracking algorithms. They have carried out new innovations on the previous models and algorithms to achieve a balance between efficiency and accuracy, and have shown great advantages in the application of multiple scenarios. In view of the singularity of current target recognition models in the field of bird detection and the rapid iteration of YOLO series, this study considers combining YOLO and SORT algorithms to better adapt to future complex scenes and accuracy requirements.

The overall working process of the DeepSORT algorithm is shown in Figure 11. Once YOLOv9 has identified the original video frames, located each target, and generated the bounding box, it initiates the tracking process based on the association prediction. For each detected target, the DeepSORT algorithm employs a Kalman filter for association prediction, thereby obtaining the corresponding detection frame. After Kalman filtering, Mahalanobis distance [64], depth representation features, and other correlation indicators are employed to calculate the degree of similarity between the detected target in the video frame and the Kalman filtering predicted track. This is achieved through cascade matching and IOU matching. The results of the above steps are then used as input to the Hungarian algorithm. At the same time, the detection point corresponding to the matching trajectory is updated by Kalman filter, and the corresponding trajectory is obtained. Finally, the process is repeated until the confirmation track or video frame is complete. Among them, the core calculation formula of the Kalman filter [61] is divided into two distinct formulas: the prediction formula and the update formula. The Kalman gain, represented by $\mathrm{G_k}$, is a crucial component of the Kalman filter. When $\mathrm{G_k}$ is zero, the gain is also zero. This means that the value of this loop is the same as the value of the last loop. In this case, the trust values currently measured are not reliable. When $\mathrm{G_k} = 1$, the gain is 1. This means that the estimates for this period are the same as the measurements, and that the estimates for the previous period cannot be trusted. In practice, $\mathrm{G_k}$ is usually between 0 and 1, indicating the degree of trust in the measured value. A is the constant, H is the scale factor, and Q is the covariance. The specific formula is as follows:

Prediction formula :

$$X_k = AX_{k-1} + Bu_k \tag{2}$$

$$P_k = AP_{k-1}A^T + Q \tag{3}$$

Updated formula:

$$G_k = P_kH^T(HP_kH^T + R)^{-1} \tag{4}$$

$$X_k = X_k + G_k(Z_k - HX_k) \tag{5}$$

$$P_k = (1 - G_kH)P_k \tag{6}$$

D. EXPERIMENT DESIGN AND DETAILS

In order to assess the effectiveness of the improved method in various complex scenarios, particularly its suitability for detecting small birds, this study conducted experiments on multiple datasets. Various data enhancement techniques, including HSV saturation [65], value enhancement, translation enhancement, scale enhancement, and Mosaic enhancement, have been adopted.

SGD optimizer [66]was adopted in the training process for random gradient descent. The models run on PyTorch [67]. We trained on NVIDIA's GeForce RTX 3090 graphics card (GPU) with 24G of RAM. The loss function selected in this study includes classification loss and regression loss. Classification Loss was BCE Loss and regression Loss was DFL Loss. The specific loss function formula is as follows: y is the true label (0 or 1) and p is the predicted probability of the output.

Classification loss function formula:

$$BCE\ loss = -y * log(p) - (1 - y) * log(1 - p) \tag{7}$$

Regression loss function formula:

$$S_i = \frac{y_i - y}{y_i + 1 - y_i} \tag{8}$$

$$S_{i+1} = \frac{y - y_i}{y_{i+1} - y_i} \tag{9}$$

$$DFL(S_i, S_{i+1}) = -((y_{i+1})log(S_i) + (y - y_i)log(S_{i+1})) \tag{10}$$

The bachsize is set as 16, the epochs of training are 400, and the YOLOv9-C weights are used as the initialization weights. The size of the input video frames or images are resized to a uniform size of 640×640, and the learning rate (lr) is 0.01.

III. RESULTS

In order to evaluate the effectiveness of the improved method in various complex scenes, different experiments were performed on visible light data, infrared data, mixed data, and video captured in real scenes. All the experiments had achieved good results.

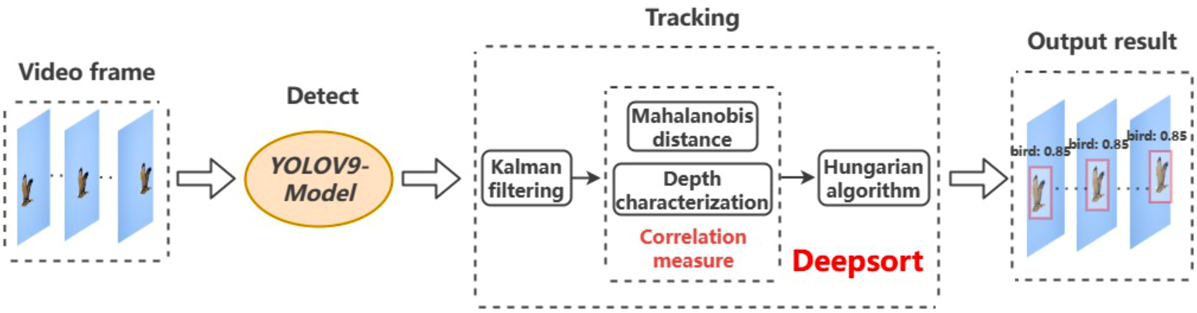


FIGURE 11. The DeepSORT algorithm initially identifies the targets within the video frames through the use of a target detector. Subsequently, the Kalman filter is applied to predict the target location. Finally, the Hungarian algorithm and cascade matching strategy are employed to facilitate the matching and trajectory association between the targets in the front and back frames.

A. RELEVANT REFERENCE INDEX

The performance of the model is evaluated using standard metrics, including three key metrics: recall, mAP0.5, and MAP0.5:0.95. These metrics are employed to assess the model's performance across various dimensions. Recall, also known as the check-all rate, is the proportion of instances correctly identified by the model as a positive class (the true class) out of all instances of the positive class. This is expressed in the mathematical formula as:

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

Where, TP represents the number of samples whose real class is positive and the final predicted result is also positive, FN represents the number of samples whose real class is positive and the final predicted result is negative. In target detection, a higher recall indicates a more comprehensive detection of the target object by the model, with fewer instances of missed cases.

The mAP is employed to evaluate the performance of the model across all categories. First, the Precision is calculated for each category. The formula for Precision is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

FP represents the number of samples whose real class is negative and the final predicted result is positive. Then, the average Precision and Recall for each class should be calculated. Finally, the area under the Precision-Recall curve is denoted as AP. The mAP Indicates the average AP values. AP and mAP is calculated as follows: r stands for Recall.

$$AP = \int_0^1 P(r)dr \tag{13}$$

$$mAP = \frac{\sum_{i=1}^{k} APi}{k} \tag{14}$$

The mAP0.5 denotes the average precision when the IOU threshold is 0.5, while mAP0.5:0.95 denotes the average precision when the IOU threshold is 0.5 to 0.95. The average value of IOU is a measure used to measure the overlap between the predicted frame and the real frame in target detection, and its value ranges from 0 to 1. In general, mAP0.5 considers only the case of high overlap between the detection wild and the real labeling. In contrast, mAP0.5:0.95 considers the case of different overlap degrees. Consequently, mAP0.5:0.95 can be adapted to different scenarios and needs, offering a more comprehensive and accurate assessment while also being less accurate. There is a contradiction between Recall and mAP. In general, mAP decreases when Recall is high, and Recall decreases dramatically when mAP is very high. Consequently, it is necessary to balance the evaluation of model performance by considering both Recall and mAP. In practical target detection scenarios, mAP0.5 is often a more significant metric to prioritize.

B. EXPERIMENTAL RESULTS

(1)On home-made visible light data set

TABLE 1. The effectiveness of each method on all categories was compared on a home-made visible light data set.

Model	Recall	mAP(0.50)	mAP(0.50:0.95)
YOLOv9	0.835	0.734	0.538
YOLOv9+CBAM	0.842	0.774	0.583
YOLOv9+AKConv+CAM	0.77	**0.822**	0.612
YOLOv9+AKConv+CAM+GAM	0.849	0.772	0.592
YOLOv9+AKConv+CAM+VoVGSCSP	0.75	0.795	**0.621**
YOLOv9+AKConv+CAM+CBAM	0.795	0.767	0.597
YOLOv9+AKConv+CAM+AFF (Head with two AFF in the middle)	0.88	0.765	0.582
YOLOv9+AKConv+CAM+AFF (one AFF in Head and one AFF Backbone)	0.838	0.761	0.580
YOLOv9+AKConv+CAM+GAM+AFF (Head with a AFF in the middle)	0.759	0.758	0.576
YOLOv9+AKConv+CAM+AFF (Head with a AFF in the middle)	0.828	**0.813**	**0.621**

As shown in Table 1, the proposed methods achieve better performance than the original model in terms of overall accuracy (mAP0.5, mAP0.5:0.95). With the same dataset and number of training rounds, it can be found that YOLOv9+AKConv+CAM+AFF (where AFF has its only added layer in Head) achieves the best overall results on the full category, with a precision of 81.3% and a recall of 82.8% with the minimum loss of Recall, which is 7.9% better than the original YOLOv9 model. These results show that our YOLOv9+AKConv+CAM+AFF model is effective in aerial bird and UAVs, especially small target detection. Secondly, we also observe that YOLOv9+AKConv+CAM is higher in precision, reaching 82.2%, but the recall is significantly lower. The main reason is that compared with the addition of AFF attention, CAM is structurally designed to focus more on the recognition of small targets, but it is not able to completely notice all the targets with large scales, so there is a decrease in the recall rate.

When the Global Attention Mechanism (GAM) is incorporated into the model, it is observed that the convergence speed is significantly accelerated. Moreover, the model's accuracy does not meet the desired requirements. This is attributed to the integration of the global attention mechanism, which leads to a dispersion of focus on smaller targets and consequently reduces the accuracy of these smaller targets.

TABLE 2. The effectiveness of each method on individual flybird was compared on a home-made visible light data set.

Model	Recall	mAP(0.50)	mAP(0.50:0.95)
YOLOv9	0.98	0.652	0.57
YOLOv9+CBAM	1. 0	0.889	0.831
YOLOv9+AKConv+CAM	0.98	0.890	0.83
YOLOv9+AKConv+CAM+GAM	1. 0	0.85	0.782
YOLOv9+AKConv+CAM+VoVGSCSP	0.961	0.883	0.819
YOLOv9+AKConv+CAM+CBAM	1. 0	0.895	0.836
YOLOv9+AKConv+CAM+AFF (Head with two layer in the middle)	1. 0	0.891	0.828
YOLOv9+AKConv+CAM+AFF (one layer each in Head and Backbone)	0.98	0.82	0.748
YOLOv9+AKConv+CAM+GAM+AFF (Head with a layer in the middle)	0.941	0.793	0.73
YOLOv9+AKConv+CAM+AFF **(Head with a layer in the middle)**	**1. 0**	**0.891**	**0.832**

Table 2 and Table 3 show the recognition effects of individual birds and individual drones respectively. It can be found that the YOLOv9+AKConv +CAM+AFF model has a particularly obvious improvement in the recognition of individual birds, among which mAP0.5 and mAP0.50:0.95 have an increase of 23.9% and 26.2%, respectively. The recognition of fixed-wing UAVs similar to birds also improved, with mAP0.5 and mAP0.50:0.95 improved by 9.4% and 11.1%respectively.

Furthermore, in order to visualize the effect of our improved model, we compare it with the YOLOv9 prototype in the All categories, Flybird, and Fixed-wing UAVs, respectively. Figure 12 clearly demonstrates that the improved model outperforms the original model in all three categories, with particularly notable improvements in the Fiybird category.

(2)On home-made infrared data set

From Table 4, the YOLOv9+AKConv +CAM+AFF improved model demonstrates the most significant improvement. The recall rate only experiences a marginal decrease, with mAP0.50 reaching 83.3% and MAP0.50:0.95 achieving 46.3%. These figures are respectively 3.2% and 2.1% higher than the original YOLOv9 model.

TABLE 3. The effectiveness of each method on individual Fixed-wing UAVs was compared on a home-made visible light data set.

Model	Recall	mAP(0.50)	mAP(0.50:0.95)
YOLOv9	1. 0	0.812	0.617
YOLOv9+CBAM	0.996	0.786	0.603
YOLOv9+AKConv+CAM	0.913	0.90	0.716
YOLOv9+AKConv+CAM+GAM	0.967	0.92	0.756
YOLOv9+AKConv+CAM+VoVGSCSP	0.80	0.906	0.709
YOLOv9+AKConv+CAM+CBAM	0.996	0.831	0.706
YOLOv9+AKConv+CAM+AFF (Head with two layer in the middle)	1. 0	0.839	0.664
YOLOv9+AKConv+CAM+AFF (one layer each in Head and Backbone)	1. 0	0.853	0.704
YOLOv9+AKConv+CAM+GAM+AFF (Head with a layer in the middle)	0.80	0.80	0.642
YOLOv9+AKConv+CAM+AFF **(Head with a layer in the middle)**	0.98	**0.906**	**0.722**

TABLE 4. The recognition effect of each method on the all category was compared on the infrared data set.

Model	Recall	mAP(0.50)	mAP(0.50:0.95)
YOLOv9	0.785	0.801	0.442
YOLOv9+AFF	0.755	0.804	0.444
YOLOv9+CAM	0.808	0.827	0.439
YOLOv9+AKConv	0.739	0.830	0.451
YOLOv9+AKConv+CAM	0.729	0.816	0.453
YOLOv9+AKConv+AFF	0.777	0.818	0.419
YOLOv9+AKConv+CAM+AFF **(Head with a layer in the middle)**	**0.777**	**0.833**	**0.463**

The study also compared the recognition effects of individual birds and individual fixed-wing UAVs, as shown in Table 5 and Table 6 respectively. The data show that the improved model is still superior to the original model in identifying individual birds, with 2.2% higher mAP0.50 and 1.3% higher mAP0.50:0.95. The identification accuracy and recall rate of fixed-wing aircraft similar to birds have also improved compared to the original aircraft. Among them, mAP0.50 increased by 7.5%, mAP0.50:0.95 increased by 3.2%, and the recall rate increased by 1.1%. The experimental results show that the improved method is very effective for the infrared image recognition of birds and fixed-wing UAVs, especially for the detection of small targets.

(3)On the infrared and visible light mixed data set

From figure 13, the experimental results indicate that the improved model has a 2.6% higher recall rate compared to the original model across all categories. Additionally, it shows a 0.9% and 1.0% improvement on mAP0.50 and MAP0.50:0.95, respectively. These data demonstrate that the improved model not only exhibits strong applicability across

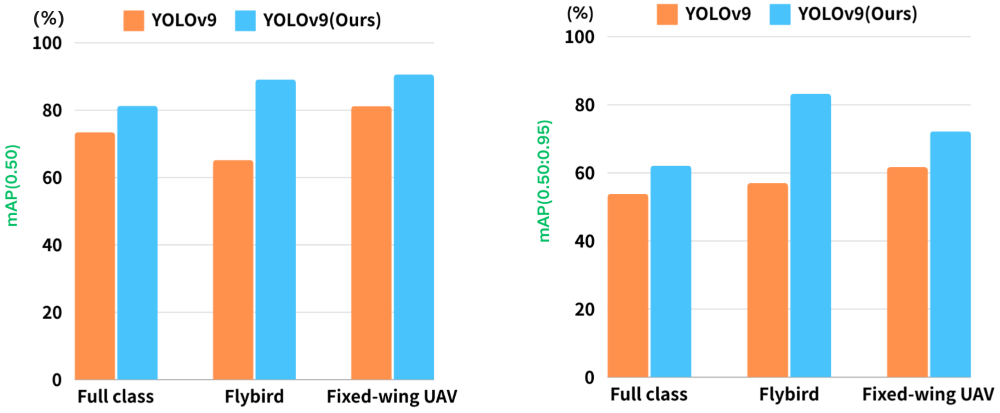


FIGURE 12. Detailed comparison of our improved model with the original YOLOv9 model in terms of All categories, Flybird, and Fixed-wing UAVs.

TABLE 5. The effectiveness of each method on individual bird identification was compared on infrared data set.

Model	Recall	mAP(H0.50)	mAP(0.50:0.95)
YOLOv9	0.814	0.937	0.478
YOLOv9+AFF	0.761	0.940	0.480
YOLOv9+CAM	0.814	0.957	0.473
YOLOv9+AKConv	0.639	0.898	0.406
YOLOv9+AKConv+CAM	0.648	0.911	0.433
YOLOv9+AKConv+AFF	0.887	0.945	0.486
YOLOv9+AKConv+CAM+AFF (Head with a layer in the middle)	**0.766**	**0.959**	**0.491**

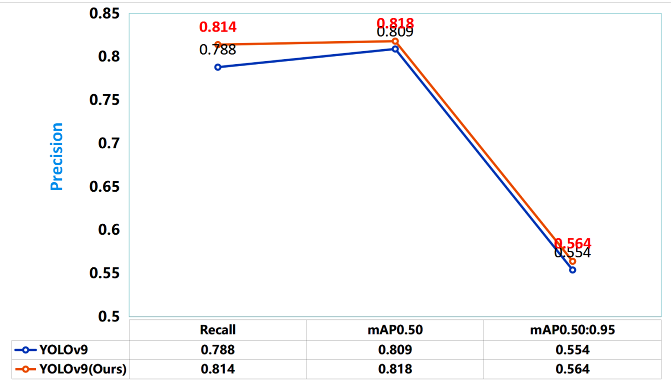
TABLE 6. The effectiveness of each method for individual fixed-wing UAVs identification was compared on infrared data set.

Model	Recall	mAP(0.50)	mAP(0.50:0.95)
YOLOv9	0.841	0.639	0.369
YOLOv9+AFF	0.778	0.579	0.340
YOLOv9+CAM	0.778	0.607	0.291
YOLOv9+AKConv	0.852	0.703	0.391
YOLOv9+AKConv+CAM	0.815	0.639	0.378
YOLOv9+AKConv+AFF	0.852	0.659	0.409
YOLOv9+AKConv+CAM+AFF (Head with a layer in the middle)	**0.852**	**0.714**	**0.401**

different scenarios, but also possesses a high degree of generalization capability.

(4)Effectiveness of our improved YOLOv9-DeepSORT in real-world

To verify the effectiveness of the improved model in real-world scenarios, we made video recordings of birds and fixed-wing drones flying in natural environments. Subsequently, the original YOLOv9-DeepSORT model and the improved YOLOv9-DeepSORT model were used to identify and monitor birds and fixed-wing aircraft respectively on the recorded video clips. Comparative analysis of the effectiveness of the two models in terms of recognition tracking is presented in Figures 14 below. The size of the image is 1280×720, and the size of the bird and aircraft targets is between 40 × 40 and 80 × 80, which appear as small targets at a distance in the original video. It can be observed from the figures that the improved model generally outperforms the original model in real-time recognition and tracking, with an average improvement of 3% across all frames. Therefore, the improved model meets the demands for real-time monitoring during live operations, thus affirming its capability to effectively track and identify aerial bird targets.



	Recall	mAP0.50	mAP0.50:0.95
YOLOv9	0.788	0.809	0.554
YOLOv9(Ours)	0.814	0.818	0.564

FIGURE 13. On the mixed data set, the original model and the improved model's validity are compared across all categories.

IV. DISCUSSION

This paper proposes an efficient automatic identification and tracking framework for birds and UAVs. By introducing a deeper convolutional network structure and an efficient feature extraction mechanism, high-precision bird detection under complex background and multiple environments is realized. The system utilizes the DeepSORT algorithm to continuously track birds in video streams, effectively solving challenges such as small targets and fast movements, thereby

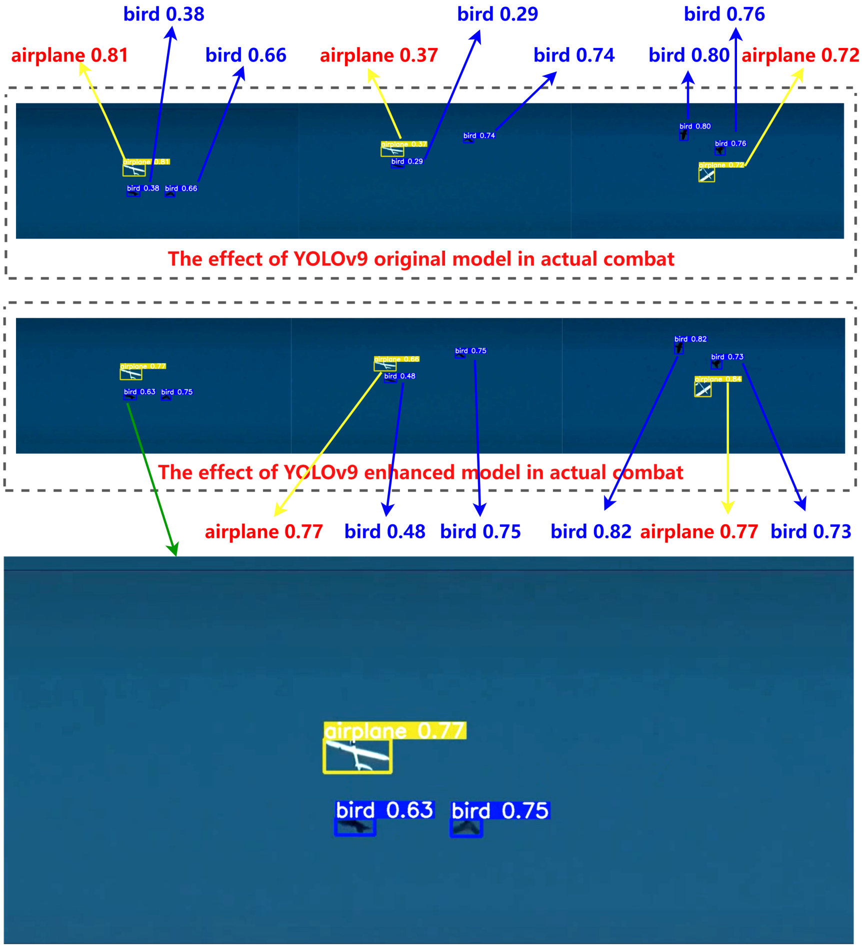


FIGURE 14. The validity of the original YOLOv9-DeepSORT model and the improved YOLOv9-DeepSORT model in real-world.

maintaining high tracking accuracy and stability.

Firstly, a context attention module is introduced to enhance the ability of the model to extract small target features and improve the detection performance significantly. Secondly, combining the attentional feature fusion (AFF) channel attention mechanism, the inconsistencies in semantics and scale are solved, and the robustness of feature extraction is improved. Finally, AKConv dynamic convolution module is used to dynamically adjust the convolution kernel, which further improves the ability of the model to capture details and context information. These improvements have enabled the YOLOv9-DeepSORT model to make significant progress in identifying and tracking birds and other aerial targets. The improvements we proposed effectively solve the challenges brought by the complex environment and improve the overall robustness and accuracy of the tracking system. However, with the iteration of more advanced technologies in the future, we will continue to improve and integrate relevant modules and technologies to further enhance the overall performance of the model.

Meanwhile, from the results of ablation experiments, we can also find some places worth digging and exploring. For example, the study discovered that the integration of AKConv and AFF into different positions of YOLOv9 had varying effects, sometimes even negative. Numerous experiments have indicated that placing AKConv at the end of the Head yielded optimal results, while integrating AFF into the middle of the head is more effective than integrating it into the front and tail of the head. Based on the characteristics of the dataset, it is possible that AKConv's poorer performance in the head may be due to its premature dynamic intervention in features with large differences, which affects the overall sensitive parameters and ultimately leads to undesirable final feature parameters. In contrast, dynamic intervention in the tail reduces the fluctuation of sensitive parameters to some extent, thereby achieving good results. The suboptimal effect of incorporating AFF attention mechanisms at both ends (head and tail) can be attributed to several factors: including the sensitivity of high-resolution detail information in the head feature map to noise interference, and insufficient detail support due to information abstraction in the tail feature map. These combined factors together limit the effectiveness of administering AFF at both ends (head and tail).

Moreover, the combination of infrared and visible light data enhances detection robustness under a variety of environmental conditions. This integration is especially beneficial when the bird is partially obscured by leaves or other obstructions, as the thermal signature can help detect where visual cues are lacking. Systems that utilize both infrared and visible light data have higher accuracy and robustness than systems that rely on a sole mode. This combined approach is excellent at detecting birds in low light conditions, distinguishing birds from other hot objects, and maintaining monitoring capabilities in severe weather conditions. The remarkable thing is that effective data fusion techniques, such as feature level and decision level fusion, are critical to integrating these patterns. Future research will focus on improving data fusion techniques beyond simple data fusion, and developing specialized deep learning models to take full advantage of combining infrared and visible light data.

In summary, the differences in the performance of different methods in different environments highlight the need for further exploration and optimization. Future research will further explore and advance advanced technologies such as AKConv, AFF, CAM, CBAM, etc., to find a more comprehensive and generalized combination. Each of these methods has advantages in feature extraction, attention mechanisms, and dynamic convolution, and is critical to improving bird recognition and tracking systems. In the future, it will provide strong support for ecological research and wildlife protection, and promote the realization of ecological sustainability.

V. CONCLUSION

In order to strengthen the identification and tracking of wild birds and support the conservation work of relevant conservation units, this paper proposes a real-time target detection and tracking model based on YOLOv9-DeepSORT. The model combines AKConv, CAM and AFF to enhance the model's ability to pay attention to target features, especially the characteristics of small target birds. This method enhances the accuracy and generalization of the model, making it suitable for effective monitoring through the parallel integration of YOLOv9 identification network and DeepSORT tracking network. At the same time, this study has generated a comprehensive dataset for bird drones and an infrared dataset. The comprehensive bird drone dataset includes a large number of small targets categorized into 4 categories, comprising 8,978 images of various sizes. The infrared dataset consists of 10,614 images capturing birds and aircraft from different perspectives and sizes, as well as nearly 5,500 images of small targets. On the self-made comprehensive data set, the improved YOLOV9-Deepsort model has a recognition accuracy of 81.3% for all categories, which is 7.9% higher than the original YOLOv9 model. The identification accuracy of bird species and fixed-wing drones with similar appearances reached 89.1% and 90.6% respectively, showing an improvement of 23.9% and 9.4% compared to the original model, demonstrating the effectiveness of the enhanced model in optical recognition. In the infrared data set, the recognition accuracy of the enhanced model for all categories, birds and fixed-wing UAVs reached 83.3%, 95.9% and 71.4%, respectively, which increased by 3.2%, 2.2% and 7.5% compared with the original model under infrared conditions, which confirmed that the enhanced model still has a good recognition performance in the infrared environment. Finally, by integrating infrared and optical data and analyzing bird flight videos under natural conditions, this study has confirmed that the recognition performance of the enhanced model in multiple scenes is still significantly better than that of the original model.

The enhanced YOLOv9-DeepSORT method expands the multi-scene application range of bird recognition and tracking models, effectively promoting the extraction of video frame features in multi-target tracking. It is hoped that this work will serve as an inspiration for future researchers interested in the conservation of wild birds.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Jincan Zhu: Concept, model analysis, improvement, experimentation, writing - manuscript; **Jian Rong**: Method - guidance; **Chenhao Ma**:Software, Validation; **Yong Cao**:Writing – review & editing;

DECLARATION OF COMPETING INTEREST

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

DATA AVAILABILITY

This study generates datasets to open all kinds of datasets. The comprehensive bird drone dataset is available from https://aistudio. baidu. com/datasetdetail/106756 and https://blog. csdn. net/DL_data_set/article/details/139567279. Infrared birds and planes by sifting through data set from the https://universe. roboflow. com/search?q= website.

ACKNOWLEDGMENT

This work was supported by the Key Laboratory of State Forestry and Grassland Administration on Forestry Ecological Big Data, Southwest Forestry University(No. 2022-BDG-09).

REFERENCES

[1] M. He, Z. Dai, X. Mo, Z. Zhang, J. Liu, W. Lei, W. Meng, B. Hu, and W. Xu, "Annual dynamics of bird community at a coastal wetland and their relation to habitat types: The example of beidagang wetland, northern china," *JOURNAL OF MARINE SCIENCE AND ENGINEERING*, vol. 11, no. 2, FEB 2023.

[2] M. A. Marini and F. I. Garcia, "Bird conservation in brazil," *Conservation Biology*, vol. 19, no. 3, pp. 665–671, 2005.

[3] A. Jungandreas, S. Roilo, M. Strauch, T. Václavík, M. Volk, and A. F. Cord, "Response of endangered bird species to land-use changes in an agricultural landscape in germany," *Regional Environmental Change*, vol. 22, no. 1, p. 19, 2022.

[4] D. Lindenmayer, C. MacGregor, and M. J. Evans, "Multi-decadal habitat and fire effects on a threatened bird species," *Biological Conservation*, vol. 283, p. 110124, 2023.

[5] G. Bota, R. Manzano-Rubio, L. Catalán, J. Gómez-Catasús, and C. Pérez-Granados, "Hearing to the unseen: Audiomoth and birdnet as a cheap and easy method for monitoring cryptic bird species," *Sensors*, vol. 23, no. 16, p. 7176, 2023.

[6] S. Rigal, V. Dakos, H. Alonso, A. Auniņš, Z. Benkő, L. Brotons, T. Chodkiewicz, P. Chylarecki, E. De Carli, J. C. Del Moral *et al.*, "Farmland practices are driving bird population decline across europe," *Proceedings of the National Academy of Sciences*, vol. 120, no. 21, p. e2216573120, 2023.

[7] J. E. Cooper, K. E. Plummer, and G. M. Siriwardena, "Using species-habitat models to predict bird counts from urban development plans," *Landscape and Urban Planning*, vol. 230, p. 104629, 2023.

[8] A. Flack, E. O. Aikens, A. Kölzsch, E. Nourani, K. R. Snell, W. Fiedler, N. Linek, H.-G. Bauer, K. Thorup, J. Partecke *et al.*, "New frontiers in bird migration research," *Current Biology*, vol. 32, no. 20, pp. R1187–R1199, 2022.

[9] C. Pérez-Granados and J. Traba, "Estimating bird density using passive acoustic monitoring: a review of methods and suggestions for further research," *Ibis*, vol. 163, no. 3, pp. 765–783, 2021.

[10] Christophe Sausse, Alice Baux, Michel Bertrand, Elsa Bonnaud, Sonia Canavelli, Alexandra Destrez, Page E Klug, Lourdes Olivera, Ethel Rodriguez, Guilllermo Tellechea, et al. Contemporary challenges and opportunities for the management of bird damage at field crop establishment. *Crop Protection*, 148:105736, 2021.

[11] H. S. Pollock, J. D. Toms, C. E. Tarwater, T. J. Benson, J. R. Karr, and J. D. Brawn, "Long-term monitoring reveals widespread and severe declines of understory birds in a protected neotropical forest," *Proceedings of the National Academy of Sciences*, vol. 119, no. 16, p. e2108731119, 2022.

[12] J. Li, S. Tian, and S. Charoenwattana, "Smart iot-based visual target enabled track and field training using image recognition," *SOFT COMPUTING*, vol. 27, no. 17, pp. 12 571–12 585, SEP 2023.

[13] Eduardo B Micaelo, Leonardo GPS Lourenço, Pedro D Gaspar, João MLP Caldeira, and Vasco NGJ Soares. Bird deterrent solutions for crop protection: approaches, challenges, and opportunities. *Agriculture*, 13(4):774, 2023.

[14] Cheng Huang, Kaiwen Zhou, Yuanjun Huang, Pengfei Fan, Yang Liu, and Tien Ming Lee. Insights into the coexistence of birds and humans in cropland through meta-analyses of bird exclosure studies, crop loss mitigation experiments, and social surveys. *PLoS Biology*, 21(7):e3002166, 2023.

[15] Rolf F Storms, Claudio Carere, Robert Musters, Hans Van Gasteren, Simon Verhulst, and Charlotte K Hemelrijk. Deterrence of birds with an artificial predator, the robotfalcon. *Journal of the Royal Society Interface*, 19(195):20220497, 2022.

[16] Moammar Dayoub, Rhoda J Birech, Mohammad-Hashem Haghbayan, Simon Angombe, and Erkki Sutinen. Co-design in bird scaring drone systems: potentials and challenges in agriculture. In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2020*, pages 598–607. Springer, 2021.

[17] Jia Li, Shuya Tian, and Sukanya Charoenwattana. Smart iot-based visual target enabled track and field training using image recognition. *SOFT COMPUTING*, 27(17):12571–12585, SEP 2023.

[18] Xufeng Lin, Chang-Tsun Li, Victor Sanchez, and Carsten Maple. On the detection-to-track association for online multi-object tracking. *PATTERN RECOGNITION LETTERS*, 146:200–207, JUN 2021.

[19] Yutong Chen, Yufen Liu, Zihan Wang, and Jiayi Lu. Research on airport bird recognition based on deep learning. In *2022 IEEE 22nd International Conference on Communication Technology (ICCT)*, pages 1458–1462, 2022.

[20] Samparthi V S Kumar and Hari Kishan Kondaveerti. A comparative study on deep learning techniques for bird species recognition. In *2023 3rd International Conference on Intelligent Communication and Computational Techniques (ICCT)*, pages 1–6, 2023.

[21] Wiam Rabhi, Fatima Eljaimi, Walid Amara, Zakaria Charouh, Amal Ezzouhri, Houssam Benaboud, Moudathirou Ben Saindou, and Fatima Ouardi. An integrated framework for bird recognition using dynamic machine learning-based classification. In *2023 IEEE Symposium on Computers and Communications (ISCC)*, pages 889–892, 2023.

[22] Aekkasit Krueangsai and Siriporn Supratid. Effects of shortcut-level amount in lightweight resnet of resnet on object recognition with distinct number of categories. In *2022 International Electrical Engineering Congress (iEECON)*, pages 1–4, 2022.

[23] U. Archana, Amanulla Khan, Appani Sudarshanam, C. Sathya, Ashok Kumar Koshariya, and R. Krishnamoorthy. Plant disease detection using resnet. In *2023 International Conference on Inventive Computation Technologies (ICICT)*, pages 614–618, 2023.

[24] Chih-Wei Lin, Zhongsheng Chen, and Mengxiang Lin. Video-based bird posture recognition using dual feature-rates deep fusion convolutional neural network. *Ecological Indicators*, 141:109141, 2022.

[25] Chih-Wei Lin, Sidi Hong, Mengxiang Lin, Xiuping Huang, and Jinfu Liu. Bird posture recognition based on target keypoints estimation in dual-task convolutional neural networks. *Ecological Indicators*, 135:108506, 2022.

[26] Xian Chen, Hongli Pu, Yihui He, Mengzhen Lai, Daike Zhang, Junyang Chen, and Haibo Pu. An efficient method for monitoring birds based on object detection and multi-object tracking networks. *Animals*, 13(10):1713, 2023.

[27] Zi-Wei Sun, Ze-Xi Hua, Heng-Chao Li, and Hai-Yan Zhong. Flying bird object detection algorithm in surveillance video based on motion information. *IEEE Transactions on Instrumentation and Measurement*, 2023.

[28] Daitao Xing, Halil Utku Unlu, Nikolaos Evangeliou, and Anthony Tzes. Drone surveillance using detection, tracking and classification techniques. In *International Conference on Image Analysis and Processing*, pages 446–457. Springer, 2022.

[29] Farhad Samadzadegan, Farzaneh Dadrass Javan, Farnaz Ashtari Mahini, and Mehrnaz Gholamshahi. Detection and recognition of drones based on a deep convolutional neural network using visible imagery. *Aerospace*, 9(1):31, 2022.

[30] Chester V Dolph, Corey Ippolito, Louis J Glaab, Michael J Logan, Loc D Tran, B Danette Allen, Mahbubul Alam, Jiang Li, and Khan Iftekharuddin. Adversarial learning improves vision-based perception from drones with imbalanced datasets. *Journal of Aerospace Information Systems*, 20(8):489–507, 2023.

[31] Dimitrios Mpouziotas, Petros Karvelis, Ioannis Tsoulos, and Chrysostomos Stylios. Automated wildlife bird detection from drone footage using computer vision techniques. *Applied Sciences*, 13(13):7787, 2023.

[32] Yunxuan Zhang. Tad: Tracking-aided detection siamese network for visual drone surveillance. In *International Conference on Autonomous Unmanned Systems*, pages 3352–3363. Springer, 2022.

[33] Qi Song, Yu Guan, Xi Guo, Xinhui Guo, Yufeng Chen, Hongfang Wang, Jianping Ge, Tianming Wang, and Lei Bao. Benchmarking wild bird detection in complex forest scenes. *Ecological Informatics*, 80:102466, 2024.

[34] Xi Chen and Zhenyu Zhang. Optimization research of bird detection algorithm based on yolo in deep learning environment. *International Journal of Image and Graphics*, page 2550059, 2024.

[35] Yang Wang, Jiaogen Zhou, Caiyun Zhang, Zhaopeng Luo, Xuexue Han, Yanzhu Ji, and Jihong Guan. Bird object detection: Dataset construction, model performance evaluation, and model lightweighting. *Animals*, 13(18):2924, 2023.

[36] Chao Zhang, Fan Shi, Xinpeng Zhang, and Shengyong Chen. Airport near-altitude flying birds detection based on information compensation multi-scale feature fusion. *IEEE Sensors Journal*, 2023.

[37] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv:2402.13616*, 2024.

[38] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[39] Xin Zhang, Yingze Song, Tingting Song, Degang Yang, Yichen Ye, Jie Zhou, and Liming Zhang. Akconv: Convolutional kernel with arbitrary sampled shapes and arbitrary number of parameters. *arXiv preprint arXiv:2311.11587*, 2023.

[40] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[41] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3560–3569, 2021.

[42] CY Wang, IH Yeh, and HYM Liao. Yolov9: Learning what you want to learn using programmable gradient information. arxiv 2024. *arXiv preprint arXiv:2402.13616*.

[43] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391, 2020.

[44] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *European conference on computer vision*, pages 649–667. Springer, 2022.

[45] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recog-

nition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16519–16529, 2021.

[46] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018.

[47] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-time flying object detection with yolov8. *arXiv preprint arXiv:2305.09972*, 2023.

[48] Jialin Liu, Fei Chao, Chih-Min Lin, Changle Zhou, and Changjing Shang. Dk-cnns: Dynamic kernel convolutional neural networks. *Neurocomputing*, 422:95–108, 2021.

[49] Hulin Li, Jun Li, Hanbing Wei, Zheng Liu, Zhenfei Zhan, and Qiliang Ren. Slim-neck by gsconv: A better design paradigm of detector architectures for autonomous vehicles. *arXiv preprint arXiv:2206.02424*, 2022.

[50] Jubin Huang, Weiwei Zhao, Zhe Lin, and Xuebin Hong. Object detection algorithm for uav images based on lightweight yolo. In *2023 4th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC)*, pages 217–220, 2023.

[51] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. 2020.

[52] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[53] Xiang Xu, Lingdong Kong, Hui Shuai, and Qingshan Liu. Frnet: Frustum-range networks for scalable lidar segmentation. *arXiv preprint arXiv:2312.04484*, 2023.

[54] Huafeng Liu, Pai Peng, Tao Chen, Qiong Wang, Yazhou Yao, and Xian-Sheng Hua. Fecanet: Boosting few-shot semantic segmentation with feature-enhanced context-aware network. *IEEE Transactions on Multimedia*, 2023.

[55] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[56] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[57] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 510–519, 2019.

[58] Chee-Yee Chong. An overview of machine learning methods for multiple target tracking. In *2021 IEEE 24th International Conference on Information Fusion (FUSION)*, pages 1–9, 2021.

[59] Davide Chicco. Siamese neural networks: An overview. *Artificial neural networks*, pages 73–94, 2021.

[60] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022.

[61] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

[62] Petar M Djuric, Jayesh H Kotecha, Jianqui Zhang, Yufei Huang, Tadesse Ghirmai, Mónica F Bugallo, and Joaquin Miguez. Particle filtering. *IEEE signal processing magazine*, 20(5):19–38, 2003.

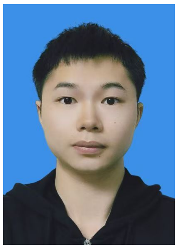
[63] Zheng Qin, Sanping Zhou, Le Wang, Jinghai Duan, Gang Hua, and Wei Tang. Motiontrack: Learning robust short-term and long-term motions for multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17939–17948, 2023.

[64] Goeffrey J McLachlan. Mahalanobis distance. *Resonance*, 4(6):20–26, 1999.

[65] Shamik Sural, Gang Qian, and Sakti Pramanik. Segmentation and histogram generation using the hsv color space for image retrieval. In *Proceedings. International Conference on Image Processing*, volume 2, pages II–II. IEEE, 2002.

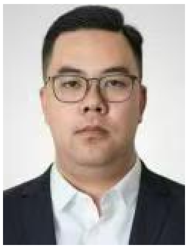
[66] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[67] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

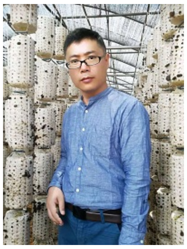


JINCAN ZHU received the bachelor's degree in Data Science and Big Data Technology. In 2023, he graduated from the College of Big Data and Intelligent Engineering of Southwest Forestry University with a bachelor's degree in Data Science and Big Data technology.

He is currently studying for a master's degree in Systems Science at the School of Big Data and Intelligent Engineering, Southwest Forestry University. His research interests include artificial intelligence, computer vision, and deep learning algorithms.



CHENHAO MA graduated from Shanghai Institute of Electrical Engineering with a bachelor's degree in Electronic Information Engineering. He is currently studying for a master's degree in Systems Science at the School of Big Data and Intelligent Engineering, Southwest Forestry University. His research interests include computer vision, image segmentation, artificial intelligence algorithms and their applications.



JIAN RONG male, born in Pengxi, Sichuan Province, graduated from Yunnan University in 2006 with a master's degree in engineering. Associate Professor, College of Big Data and Intelligent Engineering, Southwest Forestry University. His research interests are information systems and integration. Responsible for undergraduate and postgraduate teaching, teaching "Analog electronic circuit basis", "Comprehensive experiment" and other courses. Presided over the completion of one project of the National Natural Science Foundation and one project of the Education Department of Yunnan Province. He has published 2 EI papers and many Chinese journal papers.

He is currently studying for a master's degree in Systems Science at the School of Big Data and Intelligent Engineering, Southwest Forestry University. His research interests include artificial intelligence, computer vision, and deep learning algorithms.

YONG CAO Male, graduated from School of Computer Science and Engineering, University of Electronic Science and Technology of China in December 2010, and obtained a doctor of engineering. He is currently working at the College of Big Data and Intelligent Engineering, Southwest Forestry University. His research interests include fractals in nonlinear science, program correctness proof, software measurement, software engineering, deep learning algorithms, etc. He has published more than 10 papers, including 2 SCI papers, 9 EI papers and 3 ISTP papers.

• • •