ADA-YOLO:Real-Time Object Detection Model for UAVs Infrared Based on YOLOv11

Xiaolin Gu, Ao Shen

Abstract—**In recent years, unmanned aerial vehicles (UAVs) have been widely used in military, civil and commercial fields due to their flexibility, efficiency and versatility. Infrared imaging technology has become an important means of UAV object detection due to its excellent performance in complex environments such as night and foggy days. In this paper, we establish a high-precision detection model for real-time UAV infrared target detection, which effectively realizes object detection of ground objects from high altitude. We call it ADA-YOLO.The main improvements in the technology of this research focus on four main improvement methods based on the ADown module introduced by the YOLOv11n network, the Dynamic Conv module, the AFGCAttention attention mechanism, and the WIoUv3. First, the YOLOv11n algorithm improves the recognition of various types of ground targets such as pedestrians, vehicles, bicycles, etc. by introducing the ADown module, which improves the computational efficiency and the richness of the feature representations; second, we propose the optimization module of Dynamic Conv, which reduces the computation amount of the model to 5.3GFLOPs, and further improves the recognition accuracy of the model. To further optimize the model performance, we introduce the AFGCAttention mechanism, which significantly improves the recognition accuracy of the model to 92.6% with the same number of model parameters and computations. Finally, we introduce the WIoUv3 loss to make the performance of real-time UAV infrared object detection more stable when dealing with low-quality anchor boxes and outliers. Thereby, the model is able to maintain its superior performance at a low computational cost, further improving the recognition accuracy and image processing speed.Based on the experimental data, the ADA-YOLO model we built improved the average accuracy (mAP@50) by 4.8%, and the computational cost and number of parameters were controlled to be within the deployable range of the UAV terminal. In addition, it shows higher detection speed, increasing the frames per second (FPS) from 316 to 347.The ADA-YOLO real-time UAV infrared object detection model improves the average accuracy along with the increased detection speed, ensuring good operational efficiency.**

Index Terms—**UAVs, Object detection, Attention mechanism, YOLOv11n, mAP@50**

I. Introduction

In recent years, Unmanned Aerial Vehicle (UAV) [1] technology has developed rapidly and become an important part of modern technology. With its flexibility, efficiency and versatility, UAV is widely used in military, civil and commercial fields. It is capable of performing tasks in complex terrains and harsh environments, such as disaster rescue, border patrol and infrastructure inspection, while significantly reducing manpower and time costs, and realizing high-precision tasks such as agricultural monitoring, logistics and distribution, and film and television shooting. What's more, the rapid development of UAV technology provides a brand-new data collection platform and application scenario for target detection [2] tasks. With the advantages of flexibility, mobility, wide coverage and low cost, UAVs are able to efficiently acquire high-resolution images and real-time video data, providing a rich source of information for target detection. Compared with traditional imaging technology, infrared imaging technology can effectively distinguish between the target and the background by capturing the thermal radiation characteristics of the target due to its excellent performance in complex environments such as nighttime and foggy days. It has become an important means of UAV target detection. Infrared imaging is robust to light changes, shadows, camouflage and other disturbances, and can more accurately detect hidden or camouflaged targets.

However, the accuracy problem of UAV target detection and recognition accuracy has been one of the key challenges constraining its practical application. In real-world scenarios, images collected by UAVs usually face problems such as target scale diversity, changing lighting conditions, and motion blur, which significantly affect the accuracy and robustness of target detection. Targets in UAV aerial images have limited information about their features and are easily overwhelmed by complex backgrounds, leading to a significant decrease in detection accuracy [3].Targets in UAV aerial images are usually small, most of them are smaller than 32×32 pixels, which can be called target detection of small objects, which is also a major challenge for infrared target detection in UAVs. Traditional target detection algorithms require sliding windows at different locations and scales, are computationally intensive, and are sensitive to noise and appearance changes. Because of its hand-designed limitations, traditional target detection algorithms may not capture all useful information. Moreover, traditional target detection algorithms need to prepare templates for each target type and are poorly adapted to new types. Compared with traditional algorithms, deep learning algorithms can automatically learn hierarchical feature representations from raw data without the need for complex feature engineering by hand. Through the multilayer neural network structure, deep learning [4] can learn complex data representations, which is especially effective for unstructured data such as images, and can improve the accuracy and recognition efficiency of infrared target detection. In addition, deep learning models are able to generalize to new and unseen data by training on large-scale datasets, which is important for practical applications.

In the field of target detection, deep learning algorithms can be categorized into two-stage algorithms and one-stage algorithms, which differ in processing flow and performance. Two-stage algorithms first generate a series of candidate

Manuscript received March 30, 2025; revised September 6, 2025.

Xiaolin Gu is an associate professor at School of Railway Intelligent Engineering, Dalian Jiaotong University, Dalian, China (email: guxiaolin60@126.com).

Ao Shen is a postgraduate student at School of Railway Intelligent Engineering, Dalian Jiaotong University, Dalian, China (corresponding author to provide phone: +086-19811800441; e-mail: shenao19811800441@outlook.com).

regions (Region Proposals), and then perform classification and bounding box regression on these candidate regions. Representative two-stage algorithms are R-CNN [5], Fast R-CNN[6]. one-stage algorithms predict categories and bounding boxes directly on the image without generating candidate regions. Representative one-stage algorithms are mainly YOLO.Compared with two-stage algorithms, one-stage algorithms do not need to generate candidate regions first as two-stage algorithms do, and then classify and regress each candidate region, so they can significantly reduce the amount of computation and improve the detection speed. This makes one-stage algorithms more suitable for application scenarios that require real-time detection. One-stage algorithms typically have simpler network architectures because they make predictions directly on the feature map and do not require additional region proposal networks or classification networks. This simplicity makes the models easier to understand and implement, and more suitable for deployment in UAV target detection embedded systems. However, single-stage target detection algorithms also face limitations for real-time UAV infrared target detection tasks, e.g., in high-resolution images, targets may only account for a small portion of the image, and single-stage target detection algorithms may have difficulty in accurately detecting these small targets because they usually rely on relatively large anchor frames to predict the target position. Therefore, to address the above challenges, researchers are working to develop and optimize lightweight and efficient UAV infrared target detection methods to ensure that the needs for efficient processing and real-time detection are met while maintaining high performance. These studies aim to advance algorithms for fast and accurate UAV infrared target recognition even on resource-constrained platforms.

Jiang et al [7] proposed a UAV thermal infrared image and video target detection framework based on the YOLO model, and found that the YOLOv5-s model performs the best in terms of detection speed and model size, and is able to achieve efficient real-time target detection on resource-constrained UAV platforms.Tanda and Migliazzi [8] compared the effectiveness of two different airborne remote sensing platforms, drones and airplanes, in the solar photovoltaic (PV) system infrared thermography monitoring, and they found that UAVs show higher flexibility and cost-effectiveness in small-scale power plant monitoring. Kong et al [9] proposed a precise landing method for UAVs based on a ground-based infrared stereo vision system. The system expands the field of view through an infrared camera and an adjustable pan-tilt unit (PTU), and utilizes advanced image processing algorithms to achieve tracking and localization of the UAV. Hrúz et al [10] explored the application of UAV-mounted infrared cameras and radio-frequency identification (RFID) technology in the monitoring of the condition of an aircraft airframe. They proposed an intelligent maintenance scheme combining infrared cameras and RFID tags, which can effectively monitor surface and structural damages of aircraft fuselage and improve maintenance efficiency and safety. However, although deep learning algorithms such as YOLO achieve high efficiency in the field of target detection, their improvement in accuracy falls short of the requirements of infrared target recognition for UAVs, and the number of parameters and computation of the algorithms are too large, which limits their usefulness in real-time application scenarios. Therefore, this paper aims to pursue higher UAV target detection accuracy and, at the same time, minimize the amount of computation and parameter count, and ultimately achieve effective detection with minimal computational resources. To improve the recognition accuracy of infrared target detection for UAVs, we introduce Dynamic Convolution, which aims to improve the performance of lightweight convolutional neural networks (CNNs) without increasing the depth or width of the network. The core idea behind our introduction of Dynamic Convolution is to use a set of parallel convolutional kernels instead of using a single convolutional kernel per layer. These convolutional kernels are dynamically aggregated based on the inputs, weighted by an input-dependent attention mechanism. This approach is not only computationally efficient (because of the small convolutional kernels), but also has stronger representation capabilities due to the nonlinear aggregation through the attention mechanism. In order to further improve the detection accuracy and robustness of the UAV infrared target detection algorithm in complex backgrounds, we introduce the ADown module, which effectively reduces the size and computational complexity of the feature map by optimizing the downsampling process, while enhancing the multi-scale feature representation. It also improves the feature extraction efficiency of UAV infrared target detection algorithms in complex environments and adapts to the demand for model lightweighting in embedded devices. In order to improve the detection ability of the UAV infrared target detection model for small targets, we introduced the AFGCAttention mechanism, which enables the model to pay more attention to the key feature regions in the image by dynamically adjusting the weights of the channels in the feature map, while suppressing the interference of irrelevant background information. Finally, in order to significantly improve the overall performance and average accuracy of the UAV infrared target detection model, the performance of the UAV infrared target detection model in the target detection task is optimized by WIoUv3, which is capable of dynamically allocating the gradient gain according to the dynamic characteristics of the IoU and the classification criteria of the quality of the anchoring frames according to the real-time situation. The main contributions of this thesis research are summarized below:

1. In this paper, a novel high-precision UAV target detection model, ADA-YOLO, is proposed for infrared small target detection in complex environments, strongly supported by the new platform YOLOv11 algorithm. The model can effectively improve the average accuracy of UAV recognition, while the complexity and parameters of the model are small and easy to deploy on UAVs.

2.This paper introduces dynamic convolution, which is an effective lightweight CNN design method that can significantly improve the model performance without significantly increasing the computational cost. It enhances the model representation by dynamically aggregating multiple convolutional kernels and can be easily integrated into existing CNN architectures.

3.The introduction of the ADown module aims to optimize the downsampling process in the YOLOv11 network by combining a variety of pooling and convolution operations

to reduce the size of the feature map while enhancing the multi-scale feature representation and improving the model's fine-grained recognition ability in complex backgrounds.

4.The introduction of the AFGCAttention attention mechanism aims to improve the model's ability to recognize small targets by enhancing the network's ability to pay attention to key regions and suppressing the interference of irrelevant background information.

5.The introduction of WIoUv3 is used to optimize the GIoU of the YOLOv11 algorithm with the aim of better adapting to the requirements of the UAV infrared target recognition task, in particular, to achieve a more robust performance when dealing with outliers and low-quality anchored frames.

II. Related Work

The YOLO series of algorithms have been widely used and researched in both industry and academia due to their speed, effectiveness, and ease of deployment.

Zefri et al [11] investigated the use of thermal infrared and visible cameras carried by UAVs to detect images. They detected photovoltaic (PV) target features by generating orthophoto images and developed a semi-automatic hotspot extraction method. This method provides a new idea for efficient detection of high altitude infrared from UAVs and improves the advantages of UAVs in infrared detection.Chrétien et al [12] explored the potential of using UAV-mounted visible and thermal infrared cameras for remote sensing monitoring of multi-class targets. They successfully detected and classified multiple large targets through a multi-criteria goal-oriented image analysis (MOBIA) approach. This study demonstrates the promise of UAVs for large target census, especially for simultaneous multi-category detection.Kelly et al [13] investigated how to obtain accurate temperature data from a non-calibrated UAV thermal infrared camera. Through laboratory and field experiments, they found that simple empirical linear calibration can convert camera digital values to temperature values and proposed a set of best practices to minimize the effects of temperature dependence of UAV thermal infrared cameras. This study shows great promise for the application of UAV thermal infrared cameras in ecophysiology.Gui et al [14] proposed an infrared light-based method for precise landing of UAVs. By placing infrared lights on the runway and utilizing the camera and DSP processor on the UAV to detect and track the infrared lights, the method is able to achieve efficient UAV landing navigation in complex backgrounds.Niu et al [15] proposed a target detection and segmentation model (FFDSM) based on UAV infrared images that combines YOLOv5s-seg, Efficient Channel Attention (ECA) and Spatial Pyramid Pooling Fast Cross-Stage Partial Channel (SPPFCSPC) to improve the detection accuracy for targets of different sizes. Through a series of ablation experiments and comparison experiments, they verified the effectiveness and adaptability of the proposed model in different target detection scenarios.Zhang et al [16] proposed an improved Picodet small target detection method to address the real-time and accuracy problems of small target detection in UAV infrared detection. By introducing a lightweight LCNet network as the backbone network for feature extraction and combining the Squeeze-and-Excitation module and the improved feature pyramid structure, the method significantly improves the real-time (frame rate increased by 31fps) and detection accuracy (average accuracy increased by 7%) of the model. This research provides an efficient solution for UAV infrared small target detection, especially in complex background and multi-scale target scenes.

III. YOLOv11 algorithm

First proposed by Joseph Redmon [17] and other researchers, YOLO (You Only Look Once) has emerged as one of the most important methods in object detection. The algorithm uses a single-stage detection strategy and can predict the bounding box and category probability of an object directly from an image in a single forward propagation. As technology continues to advance, the YOLO family has evolved from the original version to the most recent YOLOv11, which is promoted by Ultralytics, Inc. The family's most recent accomplishment, YOLOv11, has greatly improved in terms of detection speed, accuracy, computational effort, and feature extraction capability.The key elements of the model are highlighted by the YOLOv11 architecture, which is displayed in Figure 1(a). Usually, it is composed of three main parts: the head, neck, and trunk. We provide a brief description of each part and the features that were introduced to improve the architecture as a whole below.

The backbone network, one of the fundamental elements of the YOLOv11 architecture in this study, is primarily responsible for extracting multi-scale important features from the input images. As illustrated in Fig. 1(b), this backbone network has a spatial pyramid fast pooling (SPPF) module, which effectively uses several maximal pooling layers to extract multi-scale features from the input image. As illustrated in Fig. 1(c), feature extraction in the backbone network is based on a sequence of convolutional (Conv) blocks, each of which is composed of a Conv2D layer, a BatchNorm2D layer, and a SiLU activation function.Furthermore, the backbone network integrates a cross-stage component with spatial attention mechanism (C2PSA) [18], as illustrated in Fig. 1(d). By introducing the attention mechanism, the C2PSA module significantly improves the model's detection accuracy. YOLOv11 further optimizes the backbone structure by using multiple C3K2 blocks, which replace the C2f blocks used in YOLOv8 [19]. The C3K2 blocks offer a more computationally efficient implementation of the cross-stage part (CSP) [20]. It is important to note that there are two structural variants of the C3K2 block, corresponding to c3k=false and c3k=true, as illustrated in Fig. 1(e) and Fig. 1(f), both of which aim to increase the efficacy and efficiency of feature extraction.

A key component of the YOLOv11 architecture, the neck component serves as a bridge between the head block and the backbone network [21]. As illustrated in Fig. 1, the neck structure is composed of several convolutional (Conv) layers, C3K2 blocks, feature splicing (Concat) operations, and upsampling (Upsample) blocks, all of which inherit the benefits of the C2PSA mechanism. The neck's design aims for multi-scale feature aggregation, which efficiently integrates feature information from various backbone network scales and guarantees that the features are fully utilized and improved before being passed to the head block.The

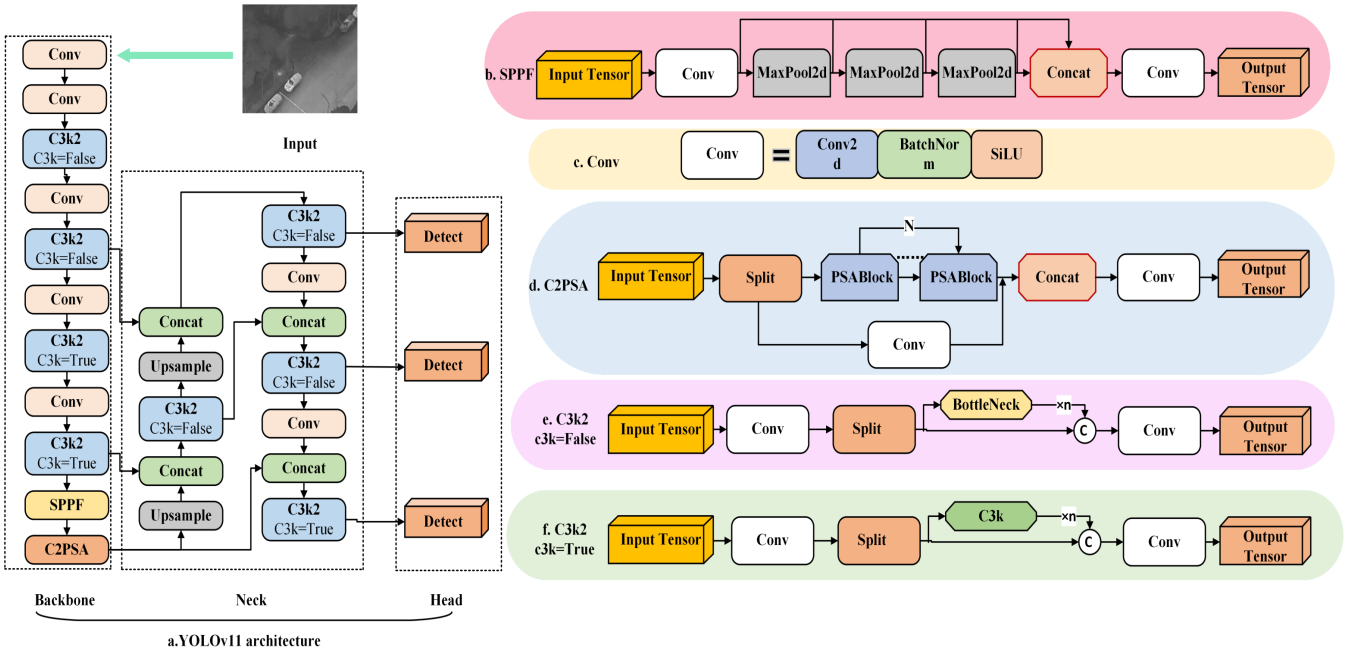

Fig. 1. The YOLOv11 network structure diagram.

neck's design improves the detection framework's overall performance by optimizing the feature delivery pipeline and enhancing the model's capacity to recognize targets at various scales.

The header module [22], the last part of the YOLOv11 architecture, is responsible for producing the final predictions. The head, which is intended to recognize items accurately , is in charge of classifying objects, calculating objectivity scores, and correctly anticipating the bounding box of every object that is detected. Through a sequence of computational stages, the head module generates the final detection results after synthesizing multi-scale features from the neck component, as illustrated in the figure. This procedure, which guarantees the model's excellent performance in the target identification job, consists of, among other things, bounding box regression, objectivity scoring, and categorization prediction.

The final component of the YOLOv11 architecture is the header module, which assumes the central function of generating the final predictions. Designed to accurately recognize objects , the head is responsible for determining object classes, computing objectivity scores, and accurately predicting the bounding box of each identified object. As shown in the Fig. 1, the head module synthesizes multi-scale features from the neck component and outputs the final detection results through a series of computational steps. This process includes, but is not limited to, categorization prediction, bounding box [23] regression, and objectivity scoring, ensuring the model's high performance in the target detection task.

IV. ADA-YOLO ALGORITHM

As we want to increase the UAV infrared target detection accuracy, we try to increase the accuracy of UAV target detection and recognition by combining advanced algorithm design and multiple module improvements. At the same time, we tried to reduce the increase in computation when increasing the recognition accuracy. Therefore, we eliminated models with large parameter sizes and slow detection speeds, and chose YOLOv11, a target detection model with high recognition accuracy and low computational effort, making it fully capable of meeting real-time demands when deployed on edge devices with limited computational resources, such as UAVs. Therefore, this study introduces a real-time UAV infrared target detection model ADA-YOLO based on YOLOv11, as shown in Fig. 2.

In this study, we first propose a dynamic convolutional approach, which is an efficient lightweight convolutional neural network (CNN) design strategy. The strategy significantly improves the performance of the model without significantly increasing the computational burden. Dynamic convolution enhances the model representation by dynamically aggregating multiple convolutional kernels and can be seamlessly integrated into existing CNN architectures, thus improving the generalization ability of the network. Next, we introduce the ADown module, which aims to optimize the downsampling process in the YOLOv11 network.The ADown module not only effectively reduces the size of the feature map but also enhances the multi-scale feature representation by combining multiple pooling and convolution operations, which in turn improves the model's ability to perform fine-grained recognition in complex contexts. In addition, this study proposes the AFGCAttention mechanism, which aims to significantly improve the model's recognition ability for small targets by enhancing the network's attention to critical regions while suppressing the interference of irrelevant background information. Finally, we introduce the WIoUv3 loss function for optimizing the GIoU loss in the YOLOv11 algorithm.WIoUv3 is designed to better adapt to the specific needs of the UAV infrared target recognition task, in particular to achieve a more robust performance when dealing with outliers and low-quality anchored frames. With these improvements, the YOLOv11 algorithm is significantly enhanced in both accuracy and robustness of target detection.

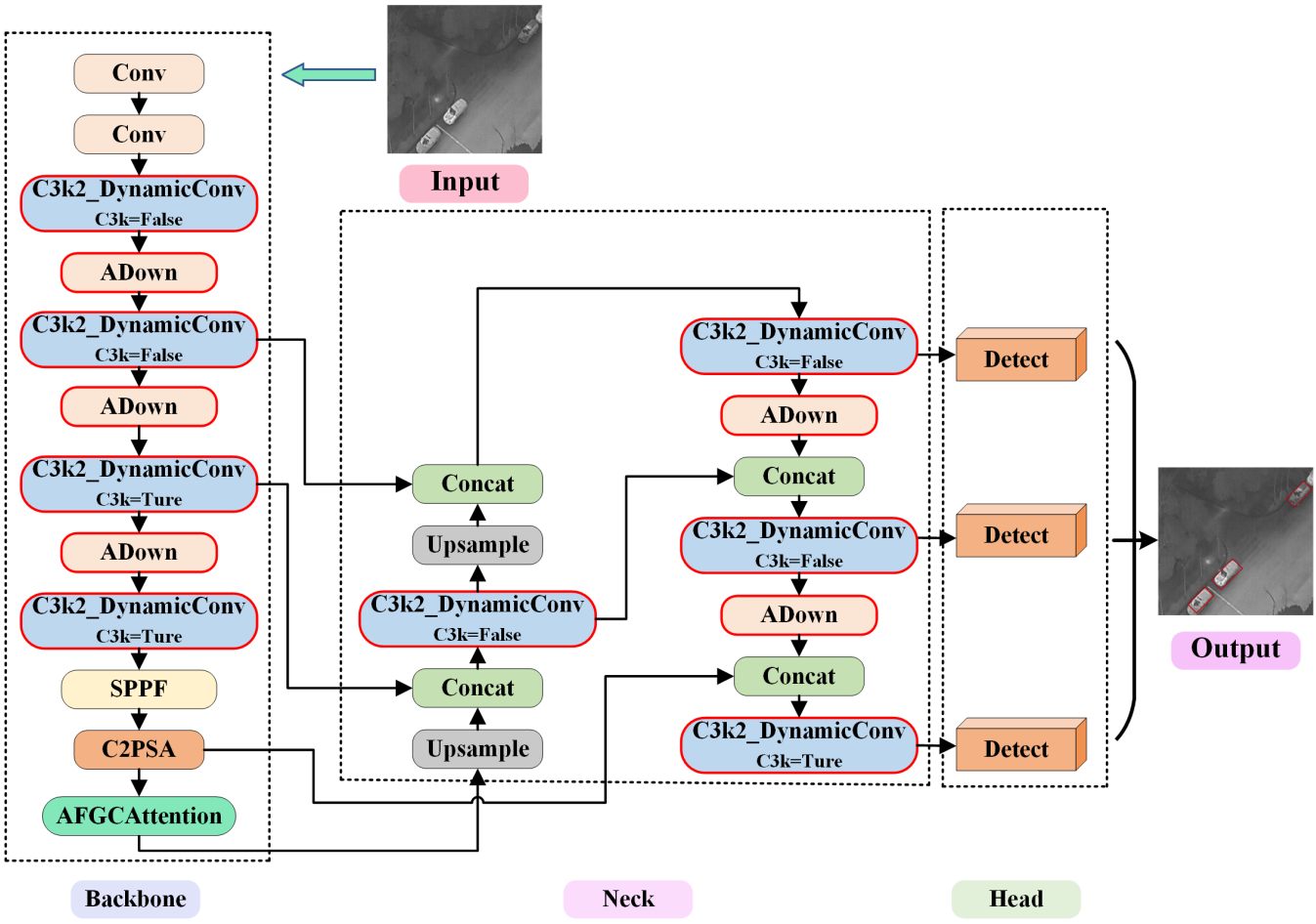

Fig. 2. Diagram of ADA-YOLO network structure. Modules with a red line on the outside indicate that this module has been improved. The details of the three modules proposed in this paper, ADown, DynamicConv and AFGCAttention, will be presented below.

A. ADA-YOLO algorithm detection process

In this study, a model for UAV infrared real-time target detection is proposed. The overall process is shown in Fig. 3. First, the dataset is preprocessed and divided into three subsets as training set, validation set and test set. Subsequently, the optimized YOLOv11n network is used to detect ground targets. During the detection process, ground targets are classified into four categories: human, bicycle, car and other vehicle. ADA-YOLO will generate different recognition accuracies based on the features of the labeled infrared target dataset HIT-UAV for recognition classification. This concludes the testing process.

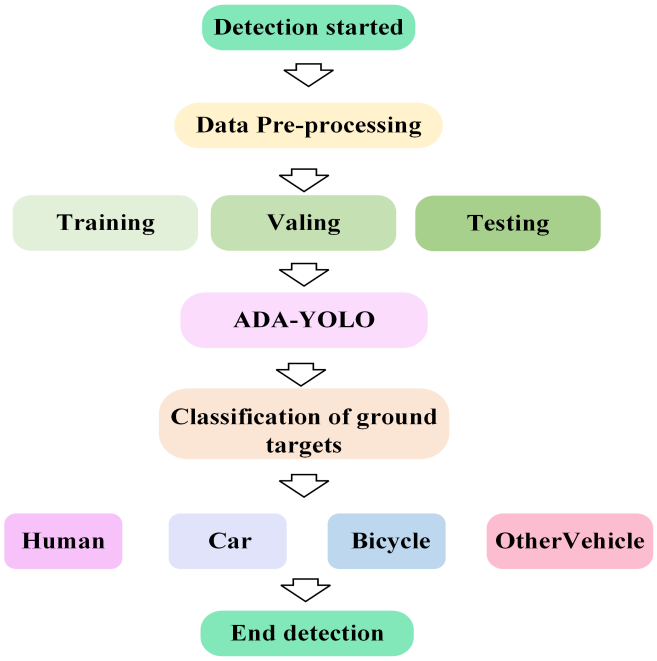

Fig. 3. ADA-YOLO algorithm detection process.

B. ADown Module

In the field of UAV infrared target detection, existing models are able to achieve high recognition accuracy when dealing with medium to large size targets. However, for detection of small-sized targets, the performance of these models degrades significantly. This performance degradation can be traced to the stepped convolution module in the YOLOv11 network architecture. While this module effectively expands the receptive domain, it is inevitably accompanied by a loss of feature information. In contrast, the composition of the ADown module mainly includes maximum pooling, average pooling, 3 × 3 convolution, and 1 × 1 convolution operations. Average pooling is able to preserve the channel information of the feature map during the downsampling process, thus effectively avoiding the information loss. On the other hand, maximum pooling may filter out part of the information

that is considered unimportant during the downsampling process, which may lead to the detection of small targets being interfered by the information of large targets or causing blurring of details.

To overcome this challenge, a combination of 1×1 convolution and 3×3 convolution modules is introduced in this study. In particular, the 3×3 convolution module is able to express more complex functions due to its higher nonlinear properties, thus capturing subtle features in the image more comprehensively and enhancing the completeness of the target information. The small-sized convolutional kernel, on the other hand, helps to extract secondary features, prompting the study to return to traditional convolutional operations for downsampling, as shown in Fig. 4. With this design strategy, we enhance the learning capability of the network, which in turn enhances the performance of the UAV infrared target detection model during training.

During the downsampling process, we successfully generated two sub-feature maps, both of which have a size of (S-1) × (S-1) × (C1/2), which is reduced compared to the size of the original feature map. Subsequently, these subfeature maps underwent different convolution operations. Specifically, each sub-feature map is processed through a specific convolutional layer to extract and refine the feature representation. After the processing is completed, these subfeature maps are connected along the channel dimensions and fused into a new feature map of size (S/2) × (S/2) × (2C), as shown in Fig. 5.

After the ADown feature transformation layer, we finally obtain a feature map with the size of S/2 × S/2 × C2, where the value of C2 is 2C. this feature map is equipped with the number of channels C2 required by the subsequent network module. this process effectively preserves all the discriminative feature information used for target detection and recognition, which ensures that a high feature discriminative power can be maintained in the subsequent processing. With this feature transformation strategy, we not only realize the spatial size reduction of the feature map, but also enhance the richness and robustness of the feature representation, providing high-quality feature inputs for the subsequent network layer processing.

C. Dynamic Convolution Module

In order to enhance the feature selectivity of the UAV infrared target detection model and thus improve the performance of the overall network, we introduce a specific dynamic perceptron implementation, namely dynamic convolution, with which we fuse C3k2 of YOLOv11. The dynamic convolution conforms to the given computational constraints (as shown in Equation 1). Similar to the dynamic perceptron, the dynamic convolution (see Fig. 6) consists of K convolutional kernels that have the same kernel size as well as input and output dimensions. These convolutional kernels are aggregated by attention weights $\{\pi_k\}$ to form the final convolutional output. Following the traditional design of Convolutional Neural Networks (CNNs), we apply activation functions (e.g., ReLU) after Batch Normalization and aggregated convolution operations to build dynamic convolutional layers. Notably, we employ the Squeeze-and-Excitation (SE) mechanism [24] to compute the attention weight $\{\pi_k(x)\}$ of the convolutional kernel, as shown in Fig. 6. In the squeezing and excitation mechanism, the global spatial information is first compressed by Global Average Pooling (GAP) to generate a compressed feature representation containing global context information. Subsequently, we utilize two fully-connected layers (with a ReLU activation function inserted between them) as well as a softmax function to generate normalized attention weights for the K convolutional kernels. The first fully-connected layer reduces the size of the features by 4. Unlike in SENet [25] where the attention is computed on the output channels, our approach computes the attention on the convolutional kernels. This computation is relatively inexpensive as it only involves adjusting the weights of the convolution kernel without additional complex computations. As a result, our dynamic convolutional design maintains computational efficiency while enhancing the model's selectivity of features, thus improving the overall network performance.

$$O\left(W^T\tilde{x}+\tilde{b}\right) \gg O\left(\sum \pi_k \tilde{W_k}\right)+O\left(\sum \pi_k \tilde{b_k}\right)+O\left(\pi\left(x\right)\right) \tag{1}$$

For processing an input feature map of dimension HWC_{in} , the computational complexity analysis of attention shows $O\left(\pi\left(x\right)\right)=HWC_{in}+\frac{C_{in}^2}{4}+\frac{C_{in}^2K}{4}$ Mult-Adds. this computational cost is significantly reduced compared to the convolution operation. The computational complexity of the convolution operation is $O\left(W^T\tilde{x}+\tilde{b}\right)=HWC_{in}C_{out}D_k^2$ Mult-Adds, where D_k represents the size of the convolution kernel and C_{out} denotes the number of output channels. It can be seen that the attention mechanism is more computationally efficient when dealing with input feature maps of the same size, thus reducing the consumption of computational resources while maintaining the performance of the model.

D. AFGCAttention Module

The last module in the backbone of the original YOLOv11 algorithmic architecture is the C2PSA module.The C2PSA module maintains or improves the performance of the network through a parameterized channel attention mechanism [26] where the model is better able to learn important features. However, it also introduces additional computational complexity, especially in lightweight networks, which may offset some of the efficiency gains due to reduced convolutional operations. We therefore introduce AFGCAttention. the AFGCNatten attention mechanism dynamically adjusts the weights of each channel according to its importance in the feature map to further optimize feature selection. This allows the model to focus more on useful features for subsequent tasks and ignore unimportant or irrelevant background noise, thus improving feature representation. This is shown in Fig. 7.

The nth channel of U can be represented by a channel descriptor $U \in R^C$ obtained by global average pooling on the space of feature mappings $F \in R^{C\times H\times W}$.

$$U_n = GAP\left(F_n\right) \tag{2}$$

We define F_n as the set of pixel values of the nth channel feature map. In addition, $GAP(X)$ represents the global average pooling function, which serves to reduce the

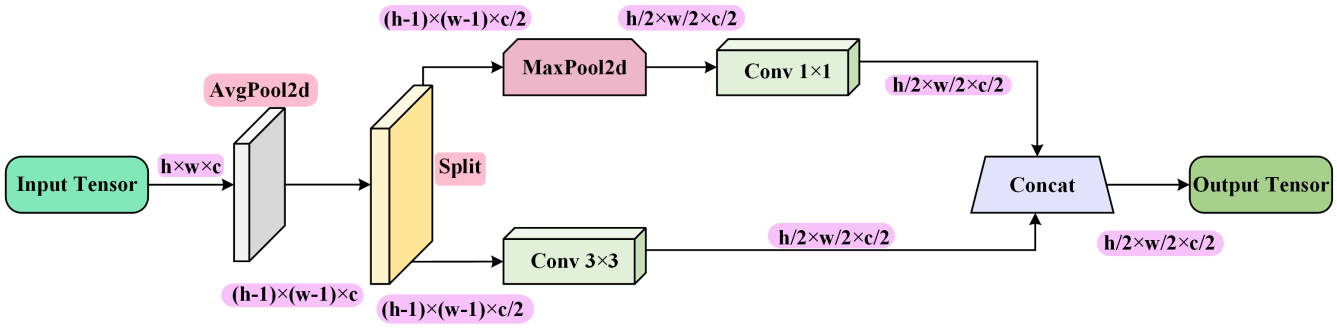

Fig. 4. Structure diagram of the ADown module

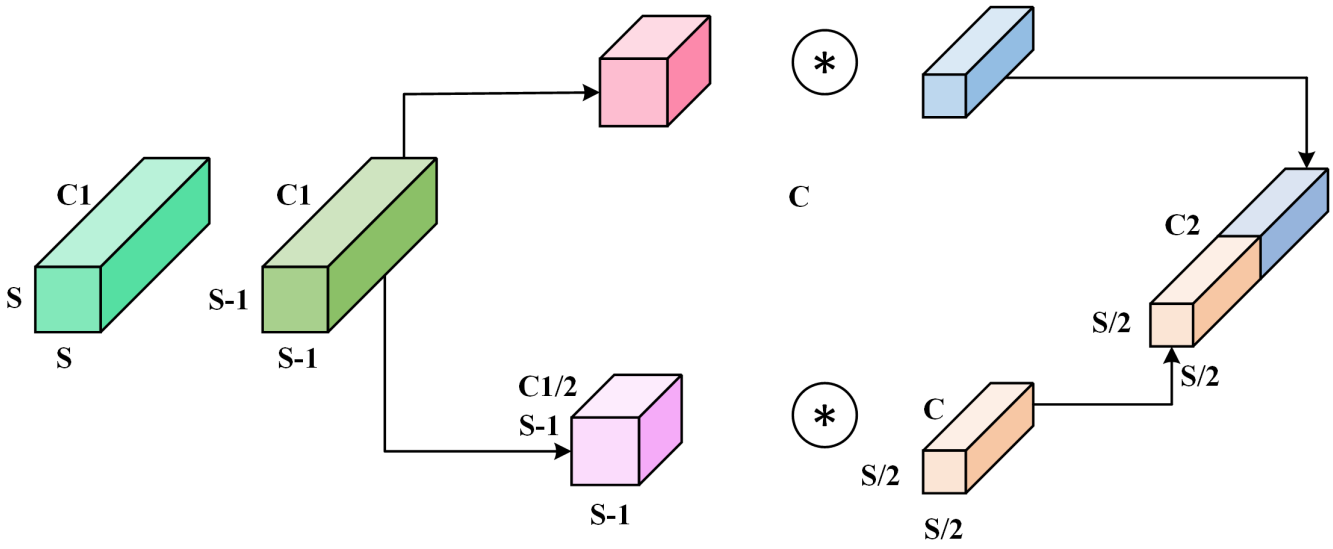

Fig. 5. Structure of ADown feature conversion.

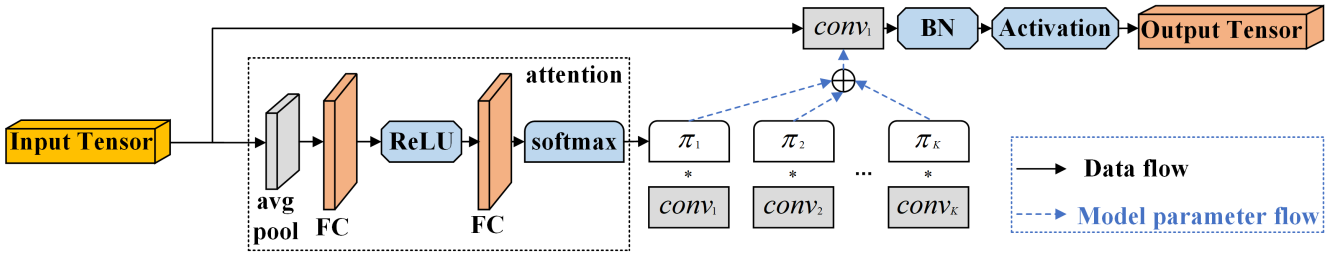

Fig. 6. Structure of Dynamic Convolution module.

dimension of the feature map from $C \times H \times W$ to $C \times 1 \times 1$, thus enabling the aggregation of the global information of the feature map ,as shown in Equation 2. In order to efficiently extract the local inter-channel interaction information while reducing the size of the model parameters, we introduce an interval matrix B. This matrix B is designed to capture the interactions between neighboring channels, and its set of elements is denoted as $B = [b_1, b_2, b_3, ..., b_k]$, which is constructed as described below:

$$U_{lc} = \sum_{i=1}^{k} U \bullet b_i \tag{3}$$

U represents the channel descriptor, while U_{lc} characterizes the local channel information. The parameter k defines the number of neighboring channels considered, as shown in Equation 3. In the experimental implementation, we perform this process through a one-dimensional convolution (Conv1D) operation. In order to further enhance the characterization of the global information and to reveal the dependencies between channels, we employ a diagonal matrix D to extract the global channel information. The diagonal matrix D is constructed as follows: $D = [d_1, d_2, d_3, ..., d_c]$, where each element di corresponds to the global information of the ith channel in the channel descriptor. The specific construction is described below:

$$U_{lc} = \sum_{i=1}^{k} U \bullet b_i \tag{4}$$

U_{gc} represents the global channel information, while the variable C indicates the total number of channels, as shown in Equation 4. In the experimental manipulation, we implement this process through two-dimensional convolution

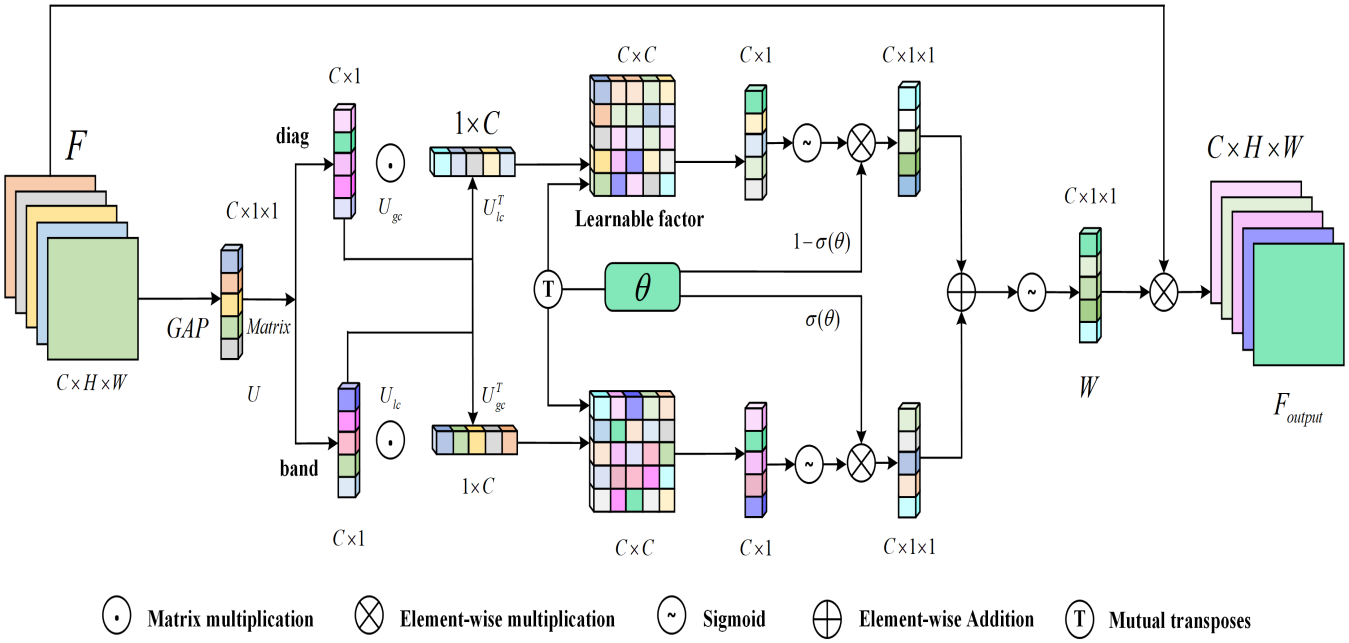

Fig. 7. Structure of AFGCAttention module.

(Conv2D). In order to facilitate the interaction between global and local information, we adopted a strategy of fusing the global information extracted through the diagonal matrix D with the local information obtained through the interval matrix B. The global information is then fused with the local information. In order to quantify the correlation between these two types of information at different scales, we introduce a correlation matrix M. This matrix M is designed to capture the interaction between global channel information and local channel information as follows:

$$M = U_{gc} \bullet U_{lc}^{T} \tag{5}$$

As defined in Equation 5, the correlation matrix M quantifies the interdependencies between global and local channel features. To facilitate precise feature weighting while maintaining computational efficiency, we introduce an adaptive fusion strategy. This strategy derives weight vectors for global and local information directly from the rows of M and the columns of its transpose, M^T, respectively. A learnable parameter then dynamically governs the fusion process, enabling the model to autonomously determine the optimal weighting scheme during training. The mechanism is formally described as follows:

$$U_{gc}^{w} = \sum_{j}^{c} M_{i,j}, i \in 1, 2, 3, ...c \tag{6}$$

$$U_{lc}^{w} = \sum_{j}^{c} \left(U_{lc} \bullet U_{gc}^{T}\right)_{i,j} = \sum_{j}^{c} M_{i,j}^{T}, i \in 1, 2, 3...c \tag{7}$$

$$W = \sigma\left(\sigma\left(\theta\right) \times \sigma\left(U_{gc}^{w}\right) + \left(1 - \sigma\left(\theta\right)\right) \times \sigma\left(U_{lc}^{w}\right)\right) \tag{8}$$

In the proposed framework, U_{gc}^{w} and U_{lc}^{w} represent the fused global and local channel weights, respectively, while c denotes the total number of channels,as shown in Equation 6 and Equation 7. The parameter σ denotes the s-type activation function, which is used to introduce nonlinearity and control the magnitude of the weights, as shown in Equation 8.

This strategy effectively avoids redundant interactions between global and local information and promotes synergy between them. In this way, the model is able to selectively enhance key information while suppressing irrelevant features, thus achieving efficient weight assignment for relevant features. Finally, by performing element-level multiplication operations of the obtained weights with the input feature map, we obtain the final output feature map. In the mathematical expression, F denotes the input feature map, F_{Outout} represents the output feature map, and the $\otimes$ symbol denotes the element-level multiplication operation, as shown in Equation 9.

$$F_{Outout} = F \otimes W \tag{9}$$

E. WIoU Loss

The three loss functions of RTDETR are the border loss function, classification loss, and confidence loss. IoU [27] is the ratio of the intersection of the real target and the candidate detection box to the union of the two components. As shown in Fig. 8. The IoU's value should be set to zero if it surpasses preset threshold values. It still has its original value in any other case, however, is not able to handle some complicated situations due to its insensitivity to small target detection and unequal weight distribution among target categories.

We present Wise-IoU ($WIoU$)[28], an enhanced IoU metric built on a weighting method, to overcome these problems. By adding movable weighting factors for intersection and concatenation regions, $WIoU$ can more adaptably represent the significance of various targets or job requirements. In particular, when determining the intersection and concatenation, $WIoU$ dynamically modifies the weights based on the size, category, or task priority of the target garbage, producing

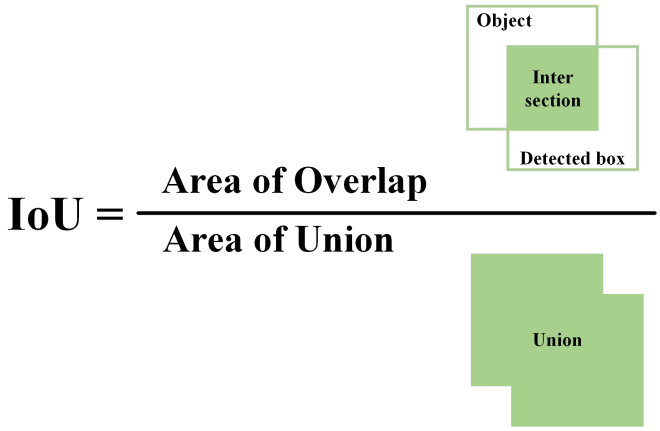

Fig. 8. IoU conceptual diagram

more precise and reliable evaluation findings for the detection results.The benefit of $WIoU$ is that it can greatly enhance the model's detection performance for imbalanced datasets, small targets, and multi-category targets. Three versions of $WIoU$ are available: $v1$ constructs a bounding box based on attentional loss, while $v2$ and $v3$ supplement $v1$ with a focusing mechanism. The performance of $v3$ is superior. Equations (7) and (8) demonstrate how a distance metric is used to construct the $v1$ of $WIoU$.

Wg and Hg stand for the width and height of the real frame x_{gt}, respectively, and y_{gt} for the minimum bounding box, where x and y are the target frame's center coordinates. For high-quality anchor frames, the value of $WIoU$ is greatly reduced by extending the value of the loss function of the traditional IoU (the value of IoU takes the value in the range of [0,1]) to include $WIoU$, which takes the value in the range of [1,e]. The distance between their centroids is the main emphasis when the target frame and the anchor frame overlap significantly. A separation between Wg and Hg from the computational map is shown by an asterisk (*).

$$R_{WIOU} = exp\left(\frac{(x-x_{gt})^2+(y-y_{gt})^2}{\left(W_g^2+H_g^2\right)^*}\right) \tag{10}$$

$$L_{WIoU_{v1}} = R_{WIOU}L_{IoU} \tag{11}$$

The YOLOv11 model's target detection performance is enhanced by the $WIoUv2$ bounding box regression loss function, which lowers the loss values of simple samples while allowing the model to concentrate on the monotonic focusing coefficients of challenging data. Equation (9) displays the $WIoUv2$ loss function formula.

$$L_{WIoU_{v2}} = L_{IoU}^{\gamma *}L_{WIoU_{v1}} \mid \gamma > 0 \tag{12}$$

In the later phases of model training, there is slower convergence since $\gamma * IoU$ reduces as IoU lowers during it. As illustrated in equation (10), a moving average IoU is added to address this issue and maintain the overall $\gamma * IoU/IoU$ at a comparatively high level.

$$L_{WIoU_{v2}} = \left(\frac{L_{IoU}^*}{L_{IoU}}\right)^{\gamma} L_{WIoU_{v1}} \tag{13}$$

An exponential factor is represented by γ. WioUv3 the quality of the anchor box is described by the loss function using anomalies; a lower quality anchor box is indicated by a lower anomaly, and a higher quality anchor box is indicated by a higher anomaly. In equation (11), the degree of abnormality is defined.

$$\beta = \frac{L_{IoU}^*}{L_{IoU}} \tag{14}$$

To improve the attention on common anchor frames, smaller gradient gains are needed for anchor frames with low anomaly values. Conversely, in order to mitigate the substantial adverse effects of low-quality anchor frames, lesser gradient benefits are allocated to anchor frames with high outliers. Equation (12) illustrates how this mechanism operates by building a focusing coefficient that is used to $WIoUv1$ to derive $WIoUv3$.

$$L_{WIoU_{v3}} = rL_{WIoU_{v1}} \mid r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{15}$$

Where the conversion factor is r, the non-monotonic focusing coefficient is β , and the hyperparameters are α and δ. The performance of the RTDETR model in the target detection task is optimized by $WIoUv3$, which is able to dynamically allocate the gradient gain based on the real-time circumstances given the dynamic characteristics of the IoU and the classification criteria of the anchor frame quality.

V. Experiments and Analysis of Results

A. Experimental Platform and Parameter Settings

The operating system for the experiment is Windows 11, Professional, the processor is 12 vCPU Intel(R) Xeon(R) Platinum 8352V CPU @ 2.10GHz, the running memory is 32GB, the GPU model is vGPU-32GB(32GB), and the experiment is performed on PyTorch 1.11.0 Deep Learning framework, Cuda 11.3.0 architecture, and the Python version is 3.8.0 training parameters: batch_size is set to 16, epoch is set to 200, the initial learning rate is 0.01, the final learning rate is 0.01, the momentum is 0.937, the weight decay is set to 0.0005, the warmup epochs is set to 3.0, the warmup momentum is 0.8, the warmup bias learning rate is 0.1, the size of the input image is automatically scaled to 640×640, no pre-training weight values are used, and the other parameters are the default values. As shown in Table I and Table II

TABLE I
Experimental Platform Configuration.

Name	Configuration
Operating System	Windows 11
CPU	12 vCPU Intel(R) Xeon(R) Platinum 8352V CPU @ 2.10GHz
GPU	vGPU-32GB(32GB)
Memory	32GB
Cuda	11.3.0
Pytorch	1.11.0
Python	3.8.0

B. Experimental Dataset

The field of UAV target detection benefits from a collection of high-altitude infrared thermal imaging data provided by the HIT-UAV dataset. As seen in Figure 9, the dataset

TABLE II
MODEL TRAINING PARAMETER SETTINGS.

Parameters	Value
Initial learning rate	0.01
Final learning rate	0.01
Batch size	16
Image size	640×640
Number of epochs	200
Momentum	0.937
Weight decay	0.0005
Warmup epochs	3.0
Warmup momentum	0.8
Warmup Bias Learning Rate	0.1

includes image samples taken by UAVs in various locations, including streets, parking lots, schools, and parks. The dataset encompasses a broad range of shooting conditions, with the shooting angle ranging from 30 to 90 degrees and the UAV flight altitude between 30 and 60 meters.As a result, the targets to be recognized exhibit a range of sizes and shapes, which aids the target detection model in better identifying and comprehending the dataset's richness and complexity. This diversity increases the model's robustness by improving its capacity to generalize to various input data scales.

Fig. 9. The partial image of the HIT-UAV dataset.

We have updated the HIT-UAV dataset to increase the average detection accuracy of UAVs in target detection tasks. The partial dataset image is shown in Fig 9. The four primary categories in the new dataset are human, bicycle, car, and OtherVehicle.With a 7:1:2 division ratio, the dataset is split into three subsets: the training set, validation set, and test set.The training set comprises 2008 photos that are used to train the model; the validation set comprises 287 images that are used to tune the model's parameters; and the test set comprises 571 images that are used to evaluate the model. is employed to modify the model's parameters, and the test set, which includes 571 photos to assess the model's performance, is displayed in Fig 10.The HIT-UAV dataset offers useful picture data support for research in related domains and is a crucial resource for the study of UAV infrared target detection and recognition.

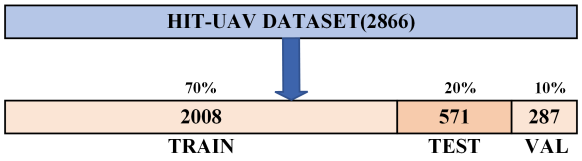

Fig. 10. Scale diagram of dataset categories.

C. Evaluation Indicators

The number of parameters,the amount of calculation, FPS and mean Average Precision (mAP) were selected as the evaluation indexes of the model. Among them, mAP is calculated by precision P and checking rate R. FPS stands for the number of image counts per second that the target detection network is capable of analyzing.

The equation for the precision P is:

$$P = \frac{TP}{TP + FP} \tag{16}$$

The equation for the rate of checking completeness R is:

$$R = \frac{TP}{TP + FN} \tag{17}$$

The equation of AP is:

$$AP = \int_0^1 PdR \tag{18}$$

The equation for mAP is:

$$mAP = \frac{1}{n}\sum_{i}^{n} AP_i \tag{19}$$

The equation for FPS is:

$$FPS = \frac{N}{T} \tag{20}$$

Where TP is the number of positive samples judged correctly, FP is the number of incorrectly detected samples, FN is the number of missed samples, AP is the area of the curve about the axes consisting of the precision P and R; mAP is the average of all APs, and i in mAP indicates the current category. When mAP is higher, it means that the model is trained better. N stands for the count of processed images and T for the total processing time.

D. Ablation experiments

In this experiment, the experimental parameter settings were kept consistent between YOLOv11n and ADA-YOLO during the training process. After improving the ADown module, the number of parameters and computation of the model decreased by 0.5M and 1.0G, respectively, and the FPS increased by 1.5%, mAP@50 increased by 1.3%. After the introduction of the Dynamic Conv module, there is a

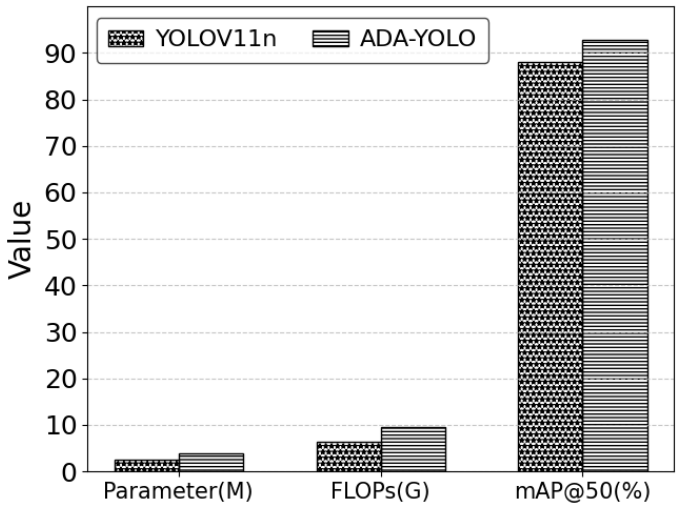

Fig. 11. Comparison of YOLOv11n and ADA-YOLO metrics.

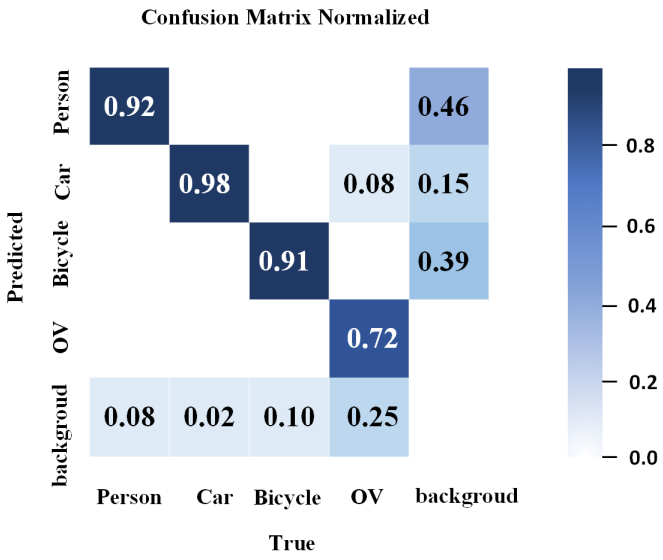

Fig. 12. Confusion matrix of YOLOv11n. OV stands for OtherVehicle.

small increase in the number of covariates and computation of the model, FPS and mAP@50 increased by 2.2%. After the introduction of the improved AFGCAttention attention mechanism, the number of parameters and computation of the model remained the same and the FPS increased by 5.6%, mAP@50 increased by 1.2%. From the Table III, it can be seen that after the introduction of WIoUv3, the number of parameters and the amount of computation of the model remain basically unchanged, and the FPS is increased by 1.2%, mAP@50 increased by 0.2%.

Then, after the introduction of the ADown module to improve the YOLOv11n backbone, we introduced the Dynamic Conv module to refine the model structure. The number of parameters and computation of the model remain basically unchanged. In addition, the FPS is improved by 3.1% over the original model, and the model is significantly lighter than the original model. mAP@50 increased by 3.6%.

Experiments were conducted to improve the model backbone using the AFGCAttention mechanism based on the improvements made to the ADown and Dynamic Conv modules. With the improved YOLOv11n model, the number

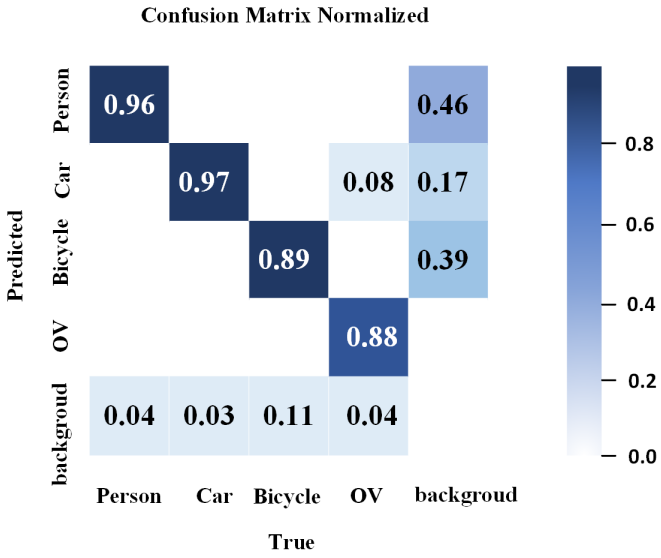

Fig. 13. Confusion matrix of ADA-YOLO. OV stands for OtherVehicle.

of parameters and computation of the model are slightly increased over the original model, but it is still a very lightweight target detection model and suitable for deployment on UAVs. In addition, the FPS increased by 7.2% over the original model and the model is significantly lighter than the original model, mAP@50 increased by 4.6%.

Based on the first three improvements to the YOLOv11n backbone, WIoUv3 is introduced in the final experiment, which constitutes the ADA-YOLO model.The ADA-YOLO model greatly improves the average accuracy of UAV infrared detection, and the advantage of recognizing small targets is more obvious, in particular.The number of parameters and computation of the ADA-YOLO model are slightly increased compared to YOLOv11n, but still controlled within the lightweight range, the FPS increased by 9.8% over the YOLOv11n model, and the average accuracy increased by 4.8% over the YOLOv11n model. As Table III and Fig.11 shown.

Therefore, the real-time UAV infrared target detection model (ADA-YOLO) proposed in this paper is suitable for UAV aerial target detection tasks. In this experiment, three modules are fused and improved with IOUs to achieve the goal of high-precision UAV infrared target detection. In addition, the mAP@50 of the ADA-YOLO model in the recognition task is significantly improved, and the ADA-YOLO model has more obvious advantages for the recognition of small targets. Fig 14 shows a comparison of the detection results of the two models YOLOv11n and ADA-YOLO.

Meanwhile, we give the confusion matrices generated by the YOLOv11n model and the ADA-YOLO model, as shown in Figs. 12 and 13.The ADA-YOLO model reduces the classification confusion and greatly improves the classification accuracy, as shown in our ablation experiments for UAV infrared target detection. The “OtherVehicle” category has the largest improvement in recognition accuracy among the four categories, with a 16% improvement in the mAP@50 metric. In addition, the recognition accuracy of the “Person” category is improved by 4%, and the single-category recognition accuracy reaches 96%. Although, the classification accuracy

TABLE III
ABLATION EXPERIMENTS.

YOLOv11n								
ADown		√				√	√	√
Dynamic Conv			√			√	√	√
AFGCAttention				√			√	√
WIoUv3					√			√
Parameters(M)	2.6	2.1	3.4	2.6	2.6	2.9	3.8	**3.8**
FLOPs(G)	6.3	5.3	6.1	6.3	6.4	5.1	9.6	**9.6**
P(%)	89.0	89.2	93.5	89.6	89.1	93.8	94.1	**94.1**
R(%)	82.7	82.2	83.2	78.9	82.7	83.0	82.9	**82.8**
mAP@50(%)	88.0	89.3	90.2	89.2	88.2	91.6	92.6	**92.8**
FPS(bt=16)	316	321	323	334	320	326	339	**347**

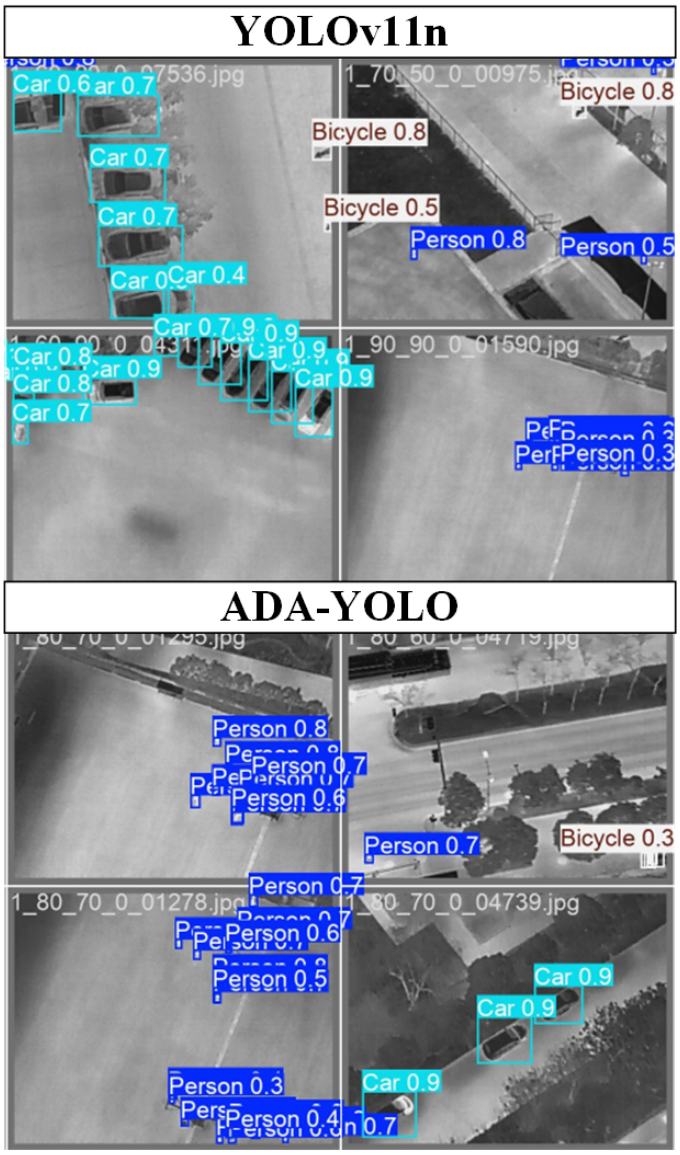

Fig. 14. Comparison of YOLOv11n and ADA-YOLO metrics.

E. Comparative Experiments on Different Attention Mechanisms

The purpose of this experiment is to compare the performance of different attentional mechanisms in the task of UAV infrared target detection and to analyze their respective parametric quantities, computation, FPS, and mAP50.In the field of computer vision, attentional mechanisms have become an important technique to improve the performance of models. We will create five augmented models based on YOLOv11n, namely YOLOv11n+Biformer, YOLOv11n+GAM, YOLOv11n+LSK, YOLOv11n+EMA, YOLOv11n+AFGCAttention.Subsequently, each model is trained until it on the validation set obtains the best performance. Finally, the performance of each modified model is evaluated on the test set to compare its number of parameters, computational effort, and classification accuracy, and then we will select the optimal augmented model as an improvement of our final attention mechanism. The comparison results are shown in Table IV and Fig.15.

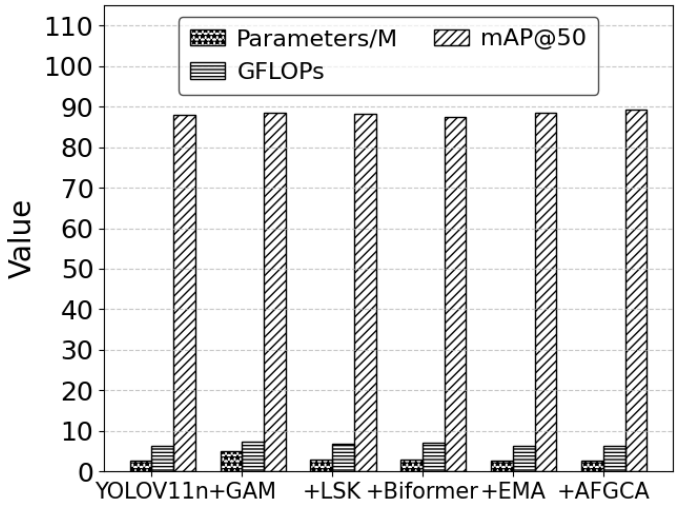

Fig. 15. Experimental comparison of improved attention mechanism models.

of the "Bicycle" category is not much different before and after the improvement, fluctuating around 90%, which also meets the standard of real-time UAV infrared target detection. In addition, the "Car" category has the highest classification accuracy of 97%. It can be seen that our UAV infrared target detection model based on YOLOv11n, after a series of improvements, meets the recognition accuracy requirements of real-time UAV infrared target detection.

As shown in Table IV shows, the YOLOv11n+EMA model is the lightest, with the least amount of computation and number of parameters, and its average accuracy is improved by 0.5% over the original model, which has some improvement effect, but the recognition accuracy is lower. Similarly, the YOLOv11n+LSK model has lower computation and number of parameters, and its average accuracy is improved by 0.2% over the original model, again with lower recognition accuracy.The YOLOv11n+Biformer

TABLE IV
EXPERIMENTAL COMPARISON OF IMPROVED ATTENTION MECHANISM MODELS.

Method	P (%)	R (%)	mAP@50 (%)	Parameters (M)	FLOPs (G)	FPS (bt=16)
YOLOV11n	89.0	82.7	88.0	2.6	6.3	316
YOLOV11n+Biformer	88.7	82.5	87.5	2.9	7.1	327
YOLOV11n+GAM	89.1	82.3	88.6	4.9	7.3	319
YOLOV11n+LSK	89.3	82.2	88.2	3.0	6.9	323
YOLOV11n+EMA	89.4	81.8	88.5	2.6	6.4	331
YOLOV11n+AFGCAttention	**89.6**	**78.9**	**89.2**	**2.6**	**6.3**	**334**

model has an increase in the number of parameters and computation but its average accuracy is lower than that of the original model by 0.5%, respectively.The YOLOv11n+GAM model has the largest computation and number of parameters among the five improvements, but its average accuracy only increases by 0.6%, which shows that the accuracy of the recognition does not meet the performance requirements for UAV infrared target detection. In contrast, the YOLOv11n+AFGCAttention model, which has the same number of parameters and computation as the original model, is the only one among the five improved attention mechanism models with an average accuracy increase of 1.2%. In addition, the YOLOv11n+AFGCAttention model has the highest FPS. Therefore, we believe that the YOLOv11n+AFGCAttention improved model is better in terms of improvement.

From this attention mechanism comparison experiment, we can infer that the YOLOv11n+AFGCAttention enhanced model outperforms the YOLOv11n algorithmic network and has the highest mAP@50 and FPS.What's more, compared with the other four groups of attention mechanism improved models, the YOLOv11n+AFGCAttention improved model has a relatively small number of parameter number is relatively small and less computationally intensive, making it suitable for real-time UAV infrared target detection tasks. Therefore, in order to realize the attention mechanism improvement in this experiment, we chose the YOLOv11n+AFGCAttention improvement model. Meanwhile, this comparison experiment can provide a valuable reference for the research of attention mechanism in the field of real-time UAV infrared target detection.

F. Experiments comparing up-sampling and down-sampling operators

The purpose of this experiment is to compare the performance of different up and down sampling operators in the task of household waste image classification and to analyze their respective parameter counts, computational effort, FPS and mAP50.In the field of computer vision, up and down sampling operators have become an important technique to improve the performance of models. We created five improved models based on YOLOv11n, which are YOLOv11n+ADown, YOLOv11n+LDConv, YOLOv11n+DySample, YOLOv11n+CARAFE, and YOLOv11n+WaveletPool.Subsequently, each model is trained until it achieves optimal performance on the validation set. Finally, the performance of each modified model was evaluated on the test set to compare its number of parameters, computation, average accuracy, and FPS. the comparison results are shown in Table V

TABLE V
EXPERIMENTS COMPARING UP-SAMPLING AND DOWN-SAMPLING OPERATORS.

Algorithmic model	Params (M)	FLOPs (G)	mAP@50 (%)	FPS (bt=16)
WaveletPool	2.2	5.5	88.1	303
LDConv	2.5	5.8	88.4	305
DySample	3.1	6.3	88.5	311
CARAFE	2.3	5.4	89.0	315
ADown	**2.1**	**5.3**	**89.3**	**321**

As shown in Table V shows, the YOLOv11n+DySample model has the largest amount of computation and number of parameters, and the recognition accuracy is 88.5%, which is a poor improvement. Similarly, the YOLOv11n+LDConv model has larger computational and parametric quantities, and its average accuracy is slightly lower than that of the original model, which also suffers from the disadvantage of lower recognition accuracy.The YOLOv11n+WaveletPool model has an increase in the number of parametric quantities and computational quantities, but its average accuracy is only 0.1% higher than that of the original model.The YOLOv11n+ CARAFE model has the most obvious optimization in terms of computation and number of parameters among the first four improvements, but its average accuracy still falls short of the performance requirements for household waste recognition, with an average accuracy of 89.0%. In contrast, the YOLOv11n+ADown model, with a reduced number of parameters and little fluctuation in computation compared to the original model, is the only one among the five improved sets of up- and down-sampling operator models with an average accuracy improvement of 1.3%. In addition, the YOLOv11n+ADown model has the highest FPS, which is 321. Therefore, we believe that the YOLOv11n+ADown improved model has better improvement.

From this up-and-down sampling operator comparison experiment, we can infer that the YOLOv11n+ADown enhanced model outperforms the YOLOv11n algorithmic network and has the highest mAP@50 and FPS. What's more, compared with the other four sets of up-and-down sampling operator improved models, the YOLOv11n+ADown improved model has relatively fewer parameters and the amount of computation is also less, making it suitable for real-time household waste identification and detection tasks. Therefore, in order to realize the up-down sampling operator improvement in this experiment, we choose the YOLOv11n+ADown improvement model. Meanwhile, this comparison experiment can provide a valuable reference for the research of up-down sampling operator in the field of real-time household garbage identification and detection.

TABLE VI
EXPERIMENTAL COMPARISON OF IMPROVED ATTENTION MECHANISM MODELS.

Algorithmic model	Params (M)	FLOPs (G)	mAP@50 (%)	FPS (bt=16)
RTDETR-r18	20.0	56,8	87.8	97
YOLOv3-tiny	12.2	18.8	85.5	52
YOLOv4	52.5	119.8	84.0	63
YOLOv5s	9.1	23.8	91.6	261
YOLOv5m	20.9	48	91.2	273
YOLOv6s	17.2	44.1	89.3	238
YOLOv8s	11.1	28.4	91.7	316
YOLOv9-c	51	238.9	92.1	337
YOLOv10m	15.7	59.8	92.3	343
ADA-YOLO	**3.8**	**9.6**	**92.8**	**347**

G. Comparative experiments with different models

In addition, in order to test the performance of the ADA-YOLO model for target detection, we compare the ADA-YOLO model with other algorithmic models to further investigate the development of the ADA-YOLO model while maintaining the same dataset and hyperparameters.RTDETR-r18, YOLO3-tiny, YOLOv4, YOLOv5s, YOLOv5m , YOLOv6s, YOLOv8s, YOLOv9-c and YOLOv10m are the popular algorithms for comparison. The current evaluation criteria are the number of parameters, arithmetic power, mAP@50 and FPS; Table VI show the comparison results.

According to the Table VI , the maximum computational amount of YOLOv9-c is 238.9 GFLOPS, and the maximum parameter number of YOLOv4 is 52.5M is hardly suitable for real-time UAV infrared target detection.The RTDETR-r18, YOLOv3-tiny, YOLOv5s, YOLOv5m, YOLOv6s, YOLOv8s, YOLOv10m models' The computational and parametric quantities are similarly too large, leading to a lag in the inference process. In contrast, the number of parameters and computation of ADA-YOLO model are 3.8M and 9.6 GFLOPS, respectively. it is easy to see that our ADA-YOLO model has obvious advantages in terms of arithmetic power and parameter lightness, as well as lower device requirements for deployment, compared with RTDETR and previous YOLO series models. More importantly, the ADA-YOLO model achieves the highest FPS and mAP@50, reaching 347 frames per second and 92.8%, respectively. For real-time UAV infrared target detection tasks, the ADA-YOLO model algorithmic approach proposed in this paper is more appropriate.

VI. CONCLUSION

In this study, we improve the YOLOv11n network constituting ADA-YOLO, which is a real-time UAV infrared target detection model. The recognition average accuracy of the ADA-YOLO algorithm is significantly improved compared to the YOLOv11n network, and the efficiency of the target detection process is optimized. We have four improvements for the YOLOv11n algorithm: first, the introduction of the ADown module, which optimizes the downsampling process in the YOLOv11 network and improves the computational efficiency; second, additional optimization through the introduction of the Dynamic Conv module, which is an effective and lightweight CNN design methodology that significantly improves the model performance without significantly increasing the computational cost. At the same time, the recognition accuracy is improved; third, by introducing the AFGCAttention mechanism aims to improve the model's ability to recognize small targets by enhancing the network's ability to pay attention to key regions and suppressing the interference of irrelevant background information, which further improves the recognition accuracy. Moreover, the results show that this attention mechanism enhancement allows the model to increase the mean accuracy value (mAP50) while keeping the number of parameters and computations constant. Fourth, the introduction of WIoUv3 allows the improved model to achieve more robust performance when dealing with outliers and low-quality anchored frames, and it results in a significant increase in the model's FPS. With the introduction of ADA-YOLO, real-time UAV IR target detection will be more effective, practical and affordable in the future. In addition, by reducing hardware requirements and controlling the amount of computation, ADA-YOLO improves the recognition accuracy of IR target detection. This helps to improve the accuracy of IR target detection for better applications in industry and life. The algorithm also contributes to the application of multimodal fusion and end-to-end learning. Our future work will include further testing the improved model on larger and more complex remote sensing datasets to validate its robustness and generalization ability in various application scenarios. We hope more researchers will join us in the future.

REFERENCES

[1] X. Wang, Y. Li, M. Chen, et al., "Prescribed-time attitude tracking control of quadrotor UAVs with unknown disturbances," Engineering Letters, vol. 33, no. 2, pp247-252, 2025.

[2] H. Zhou and H. Dai, "Research on fatigue driving detection based on deep learning," Engineering Letters, vol. 33, no. 2, pp348-356, 2025.

[3] X. Li and Y. Zhang, "A lightweight method for road damage detection based on improved YOLOv8n," Engineering Letters, vol. 33, no. 1, pp114-123, 2025.

[4] A. Vaid, K. W. Johnson, M. A. Badgeley, et al., "Using deep-learning algorithms to simultaneously identify right and left ventricular dysfunction from the electrocardiogram," J. Cardiovasc. Comput. Tomogr., vol. 15, no. 3, pp. 395-410, 2022.

[5] P. Song, P. Li, and L. Dai, "Boosting R-CNN: Reweighting R-CNN samples by RPN's error for underwater object detection," Neurocomputing, vol. 530, pp. 150-164, 2023.

[6] Z. Wang, Y. Ling, X. Wang, et al., "An improved Faster R-CNN model for multi-object tomato maturity detection in complex scenarios," Ecol. Inform., vol. 72, p. 101886, 2022.

[7] C. Jiang, H. Ren, X. Ye, et al., "Object detection from UAV thermal infrared images and videos using YOLO models," Int. J. Appl. Earth Obs. Geoinf., vol. 112, p. 102912, 2022.

[8] G. Tanda and M. Migliazzi, "Infrared thermography monitoring of solar photovoltaic systems: A comparison between UAV and aircraft remote sensing platforms," Therm. Sci. Eng. Prog., vol. 48, p. 102379, 2024.

[9] W. Kong, D. Zhang, X. Wang, et al., "Autonomous landing of an UAV with a ground-based actuated infrared stereo vision system," in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Tokyo, Japan, Nov. 2013, pp. 2963-2970.

[10] M. Hrúz, M. Bugaj, and A. Novák, "The use of UAV with infrared camera and RFID for airframe condition monitoring," Appl. Sci., vol. 11, no. 9, p. 3737, 2021.

[11] Y. Zefri, A. ElKettani, and I. Sebari, "Thermal infrared and visual inspection of photovoltaic installations by UAV photogrammetry—Application case: Morocco," Drones, vol. 2, no. 4, p. 41, 2018.

[12] L. P. Chrétien, J. Théau, and P. Menard, "Wildlife multispecies remote sensing using visible and thermal infrared imagery acquired from an unmanned aerial vehicle (UAV)," Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci., vol. 40, pp. 241-248, 2015.

[13] J. Kelly, N. Kljun, P. O. Olsson, et al., "Challenges and best practices for deriving temperature data from an uncalibrated UAV thermal infrared camera," Remote Sens., vol. 11, no. 5, p. 567, 2019.

[14] Y. Gui, P. Guo, and H. Zhang, "Airborne vision-based navigation method for UAV accuracy landing using infrared lamps," J. Intell. Robot. Syst., vol. 72, no. 2, pp. 197-218, 2013.
[15] K. Niu, C. Wang, J. Xu, et al., "An improved YOLOv5s-Seg detection and segmentation model for the accurate identification of forest fires based on UAV infrared image," Remote Sens., vol. 15, no. 19, p. 4694, 2023.
[16] Q. Zhang, L. Zhou, and J. An, "Real-time recognition algorithm of small target for UAV infrared detection," Sensors, vol. 24, no. 10, p. 3075, 2024.
[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 779-788.
[18] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," arXiv preprint arXiv:2410.17725, Oct. 2024.
[19] G. Wang, Y. Chen, P. An, et al., "UAV-YOLOv8: A small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios," Sensors, vol. 23, no. 16, p. 7190, 2023.
[20] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "Scaled-yolov4: Scaling cross stage partial network," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Nashville, TN, USA, Jun. 2021, pp. 13029-13038.
[21] T. Liang, X. Chu, Y. Liu, et al., "Cbnet: A composite backbone network architecture for object detection," IEEE Trans. Image Process., vol. 31, pp. 6893-6906, 2022.
[22] M. Tömösközi, M. Reisslein, and F. H. P. Fitzek, "Packet header compression: A principle-based survey of standards and recent research studies," IEEE Commun. Surv. Tutor., vol. 24, no. 1, pp. 698-740, 1st Quart., 2022.
[23] S. Ma and Y. Xu, "Mpdiou: a loss for efficient and accurate bounding box regression," arXiv preprint arXiv:2307.07662, Jul. 2023.
[24] C. Zeng, Y. Zhao, Z. Wang, et al., "Squeeze-and-excitation self-attention mechanism enhanced digital audio source recognition based on transfer learning," Circuits Syst. Signal Process., pp. 1-33, 2024, doi: 10.1007/s00034-024-02601-1.
[25] M. Li, Y. Zheng, D. Li, et al., "Ms-senet: Enhancing speech emotion recognition through multi-scale feature fusion with squeeze-and-excitation blocks," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Seoul, South Korea, Apr. 2024, pp. 12271-12275.
[26] X. Li, M. Li, P. Yan, et al., "Deep learning attention mechanism in medical image analysis: Basics and beyonds," Int. J. Netw. Dyn. Intell., vol. 2, no. 2, pp. 93-116, 2023.
[27] H. Zhang and S. Zhang, "Focaler-iou: More focused intersection over union loss," arXiv preprint arXiv:2401.10525, Jan. 2024.
[28] C. Wang, Q. Zhang, and J. Huang, "An improved multi-target detection algorithm in UAV aerial images based on YOLOv8s framework," Engineering Letters, vol. 33, no. 4, pp998-1007, 2025.