Research on Lightweight Infrared Target Detection Algorithm Based on Deep Learning

Jingshi Zhang, Yujun Zhang*, Jianhan Zhou

Abstract—**Infrared target detection has become widely applied in drone-based detection systems. However, factors such as small target size and complex contextual information have significantly increased the difficulty of detection tasks. To address these challenges, we propose a lightweight infrared target detection model based on YOLOv9. First, we introduce a combination of a spatial-to-depth layer (SPD) and a convolution-free stride layer (Conv) to form SPD_Conv, replacing the original adaptive downsampling layer. This modification reduces the model's parameters and computational load. Next, we replace the original convolutional layers with a dual convolution (DualConv) that integrates group convolutions and heterogeneous convolutions, effectively reducing both computational cost and parameter size while improving detection accuracy. Additionally, Partial Convolution (PConv) is applied to the efficient channel attention mechanism (ECA). Adaptive global max pooling and adaptive global average pooling are incorporated at the input layer, summed together, with a fixed convolution kernel size. Furthermore, we design a novel cross-channel interaction module, PC_EPCA, to enhance the efficiency of spatial feature extraction. These innovations result in improved target detection performance while maintaining a lightweight design, making the model particularly well-suited for multi-scale target detection tasks. The improved model reduces parameter size by 43.9%, computational load by 48.2%, and increases mAP50 accuracy by 1.1% on the HIT-UAV dataset. The model demonstrates exceptional performance in infrared small target detection for drones, effectively reducing both parameters and computation while maintaining high detection accuracy.**

Index Terms—**Infrared Small Target Detection; YOLOv9; Lightweight; Attention Mechanism; Dual Convolution**

I. Introduction

In recent years, with the rapid development of technology, the application of infrared image target detection has become increasingly widespread.Infrared small target detection represents a vital research focus within the domains of computer vision and object detection. It aims to identify small, irregularly shaped, and low-contrast targets from infrared images or videos.Due to the inherently small size of targets, substantial background interference, and low target-to-background contrast, the detection of infrared small targets poses considerable challenges in real-world applications[1]. Infrared imaging generates images by the thermal radiation emitted by objects, which is particularly effective in nighttime or low-light environments. Compared to visible light imaging, infrared images have stronger anti-interference capabilities and are suitable for harsh weather conditions, dark environments, and other scenarios. Aerial images, due to their wider field of view, typically contain more object instances. Moreover, infrared small targets usually have a small temperature difference and low contrast with the background, making it difficult to distinguish them effectively. The challenges are further compounded by issues such as small scale and frequent occlusion of infrared small targets[2]. Compared with detection tasks in natural environments, aerial detection faces even more severe challenges. Therefore, the research on this model holds significant value in the field of computer vision.

Infrared small target detection can generally be divided into two primary approaches: conventional detection techniques and deep learning-based methods.[3]. Both traditional and deep learning methods have achieved remarkable success in the visible light image domain. However, traditional image processing techniques, such as background modeling and suppression, edge detection and region extraction, and small target enhancement, struggle when dealing with the complex background clutter of infrared small targets[4]. These methods often fail to effectively suppress noise, leading to suboptimal detection performance in complex infrared scenes.The emergence of deep learning, especially the use of Convolutional Neural Networks (CNNs), has ushered in a new era for infrared small target detection.Deep learning is a machine learning method based on artificial neural networks, which solves complex data problems by simulating the pattern recognition mechanisms of the human brain. Through forward and backward propagation, deep learning models continuously optimize weights to minimize prediction errors. The impressive capabilities of convolutional neural networks (CNNs) in image processing have established them as the mainstream approach for infrared small target detection. By leveraging end-to-end training, CNNs [5] are able to extract intricate features from infrared imagery, thereby enhancing the accuracy of small target detection. Moreover, CNNs demonstrate strong robustness and are well-suited for handling large-scale tasks. Deep learning-based target detection models can generally be categorized into two types: two-stage models and single-stage models. Among the two-stage approaches, R-CNN, introduced by Ross Girshick and his team in 2014 [6], stands out as a pivotal method in deep learning for object detection. The R-CNN series is particularly effective in achieving high detection accuracy, especially for small objects, and is renowned for its precise localization capabilities. These algorithms also offer significant advantages in scenarios

Manuscript received February 26, 2025; revised May 12, 2025.

The work was supported by the Natural Science Foundation of China (No.62272093) and by Liaoning Key Laboratory of the Internet of Things Applicaiton Technology on Intelligent Construction.

Jingshi Zhang is a graduate student of School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (e-mail: 1006229587@qq.com).

Yujun Zhang is a Professor of School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (Corresponding author to provide e-mail: 1997zyj@163.com).

Jianhan Zhou is a teacher of School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (e-mail: zjhzjhan@163.com).

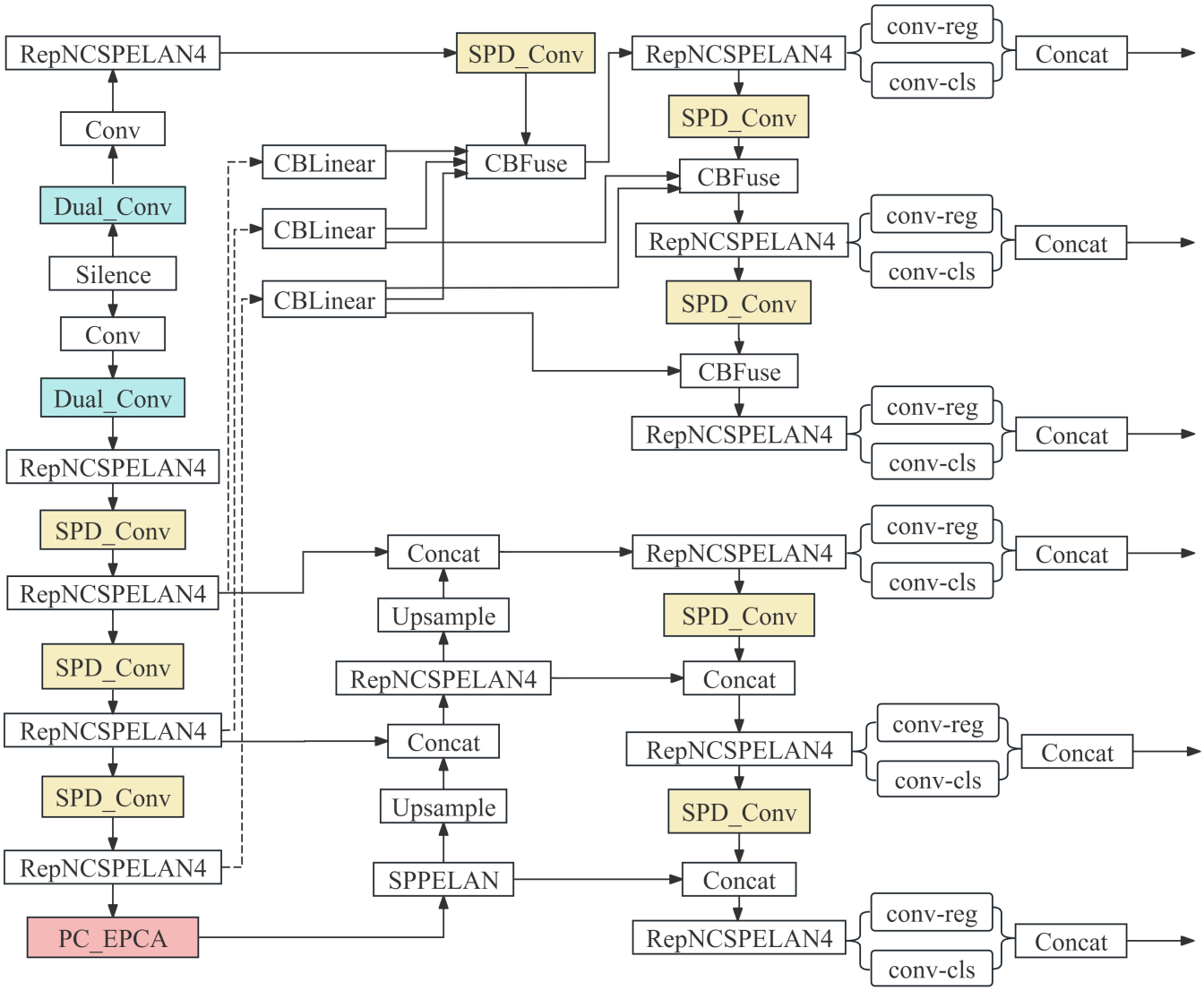

Fig. 1: Improved Network Structure

that require complex feature extraction and multi-scale detection. Fast R-CNN[7], another important work by Ross Girshick in 2015, builds on R-CNN by improving training speeds and accuracy. Faster R-CNN[8], which also uses VGG16 as its backbone, achieves an inference speed of 5 fps on a GPU (including the generation of candidate regions) while improving accuracy. However, two-stage models tend to be computationally expensive, involve complex algorithm designs and implementations, and require substantial memory to store large amounts of candidate regions and intermediate feature data, leading to high memory consumption. In contrast, the single-stage YOLO model is faster, enabling real-time detection with high accuracy. In scenarios with dense object distributions, YOLO can efficiently handle the detection and classification of multiple objects. As the YOLO model continues to evolve[9], it strikes a good balance between accuracy and speed. Therefore, this paper is based on YOLOv9.

Infrared small target detection has garnered considerable interest from researchers in recent years. Yu et al. integrated neural networks with the concept of local contrast and proposed a novel architecture called the Multi-Scale Local Contrast Learning Network (MLCL-Net) [10]. Similarly, Zhao et al. presented the Gradient-Guided Learning Network (GGL-Net) [11], which also incorporates local contrast learning (LCL). Within their feature extraction module, they introduced a Gradient Supplement Module (GSM) to encode original gradient information into deeper layers of the network and effectively utilized an attention mechanism to enhance feature representation. Wu et al. developed the Deep Interactive U-shaped Network (DI-U-Net) [12], which employs a multi-level residual U-shaped structure with both long and short skip connections to preserve feature resolution. Hou et al. proposed the Robust Infrared Small Target Detection Network (RISTD-Net) [13], featuring a hybrid feature extraction framework that integrates hand-crafted features. In another study, Wang et al. introduced RLPGB-Net [14], a pyramid feature fusion detection network that leverages reinforcement learning to emphasize the salient characteristics of targets.

Accordingly, to address the challenge of detecting small targets in complex background environments—particularly in infrared imagery captured by aerial drones [15]—this paper presents a lightweight detection algorithm built upon

the YOLOv9 architecture.The lightweight model, which is an improved version based on YOLOv9, is illustrated in Figure 1. The original adaptive downsampling layer of the model is replaced with a combination of a spatial-to-depth layer (SPD) [16] and a convolution-free stride layer (Conv), forming SPD_Conv. This modification decreases the number of model parameters and lowers computational complexity. Additionally, the original convolution layers are replaced with a dual convolution (DualConv) [17], which combines group convolutions and heterogeneous convolutions. This reduces computational cost and parameter size while improving detection accuracy. Finally, a novel and improved cross-channel interaction mechanism, the high-efficiency channel attention (PC_EPCA), is introduced in the backbone part to enhance spatial feature extraction efficiency.This leads to enhanced target detection performance while preserving a lightweight architecture. Additionally, the algorithm is further examined for its potential applications in the domain of infrared small target detection.

II. RELATED ALGORITHMS

YOLOv9 [18] is a deep learning model specifically designed for real-time object detection tasks. It employs a lightweight architecture built upon a backbone network, offering distinct advantages. This design guarantees high detection accuracy while greatly improving processing speed, thereby making it well-suited for real-time applications. In the context of deep network processing for multi-object detection tasks, the model often needs to accommodate a wide range of variations. To address this challenge, researchers introduced the concept of Programmable Gradient Information (PGI). PGI delivers complete input information to the objective function of the target tasks, thereby ensuring precise gradient calculations and facilitating effective updates to the network weights. Furthermore, leveraging gradient path planning, the researchers developed the General Efficient Layer Aggregation Network (GELAN). This novel lightweight network architecture further highlights the superior performance of PGI in optimizing lightweight models.

The backbone network of YOLOv9 is used to extract low-level and high-level features from input images, playing a central role in feature extraction and image representation. YOLOv9 introduces a completely new backbone design. Compared to the CSPDarkNet [19] and EfficientNet in YOLOv4 and YOLOv5, YOLOv9 employs a hybrid network architecture that combines the strengths of Convolutional Neural Networks (CNNs) and Transformer architectures. This design enhances the efficiency of feature extraction and improves the handling of complex scenes and small object detection. The feature extraction module is a key innovation of YOLOv9, designed to effectively extract multi-scale feature information. Common modules such as Spatial Pyramid Pooling (SPP) and Deformable Convolutions (DCN) are used to capture details and variations at different scales. The neck of the network further processes and integrates features from different scales to provide rich feature representations for subsequent object detection tasks. In the head of the network, the processed features are applied to object localization, class prediction, and bounding box regression, ultimately generating detection results. YOLOv9 also introduces innovations in its loss function, including a region balancing mechanism. This mechanism helps reduce false detections caused by background noise, especially in dense scenes, and enhances the detection capability for small objects. When handling high-resolution inputs, YOLOv9 automatically adjusts the network layer parameters, ensuring that the model maintains efficient inference speed while ensuring detection accuracy.

YOLOv9 excels in multi-scale object detection, outperforming traditional Feature Pyramid Networks (FPN) by effectively recognizing both small and large targets through the integration of feature pyramids with a self-attention mechanism. During the feature map fusion process, YOLOv9 incorporates a spatial attention mechanism, enabling the model to prioritize key regions of interest in the image, suppress background noise, and enhance detection accuracy. Additionally, YOLOv9 introduces extra convolutional layers in the top-down path, facilitating more effective integration of information across multiple scales. This optimization enables the model to capture object features at various scales with greater precision, significantly improving small object detection. Its multi-scale feature fusion strategy further allows the model to perform object detection across different scales simultaneously, greatly enhancing its ability to detect targets in complex environments. YOLOv9 has demonstrated exceptional performance in object detection tasks, particularly excelling in challenging scenarios.

III. IMPROVEMENTS

A. SPD_Conv Module

The working principle of Convolutional Neural Networks (CNNs) involves progressively extracting increasingly complex features from input data through multiple layers of convolution operations. Models such as AlexNet [20], VGGNet [21], and ResNet [22] have shown outstanding performance in image classification tasks. However, in tasks such as low-resolution image detection or small object detection, the performance of existing CNN models often deteriorates significantly. This decline can be attributed to several factors, including the inherently low resolution of small objects, limited contextual information, and the tendency for large objects to dominate the feature learning process when they coexist within the same image.Moreover, the commonly used stride convolutions and pooling operations in contemporary CNN architectures often lead to the loss of fine details and suboptimal feature representation, especially when dealing with low-resolution images or small objects. SPD_Conv consists of a spatial-to-depth layer and a stride-less convolution layer. It transforms the spatial dimensions of the input feature map into depth dimensions, effectively increasing the feature map's depth while preserving the original information. The operation of SPD on the feature map is illustrated in Figure 2.The initial feature map is split into multiple sub-feature maps and subsequently concatenated along the channel axis, producing an intermediate feature map with decreased spatial resolution and expanded channel depth. This approach mitigates information loss, particularly when processing low-resolution images and small objects, thereby preserving more spatial details.

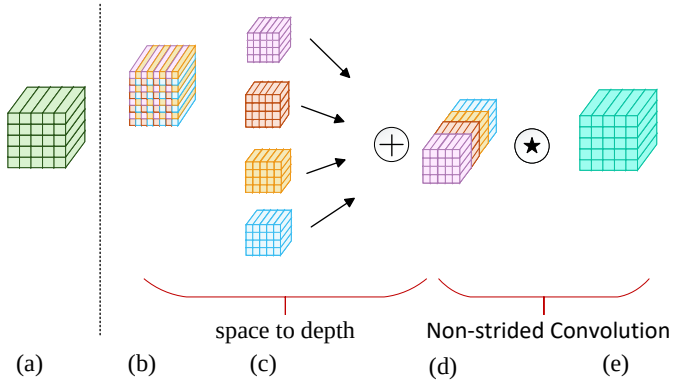

Fig. 2: Working principle of SPD_Conv

The specific process works as follows: for an intermediate feature map X of size $S\times S\times C_1$, a series of sub-feature maps are sliced according to a specific rule. For example,$f_{0,0} = X[0 : S : scale, 0 : S : scale]$, which means that every scale pixel is taken from X n both the horizontal and vertical directions to form a sub-feature map $f_{0,0}$. Similarly $f_{1,0} = X[1 : S : scale, 0 : S : scale]$is obtained by taking pixels with the same interval starting from the next row (index 1). This process is repeated to obtain multiple sub-feature maps. The slicing approach for the sub-feature maps at different positions under a given scale value is illustrated here, covering different starting positions in both horizontal and vertical directions.

In general, for a given original feature map X , the sub-map $f_{x,y}$ consists of all $X(i, j)$ elements such that $i+x$ and $j+y$ are divisible by scale. This means that by varying the values of x and y , one can determine the position and element selection rules for the sub-feature maps in the original feature map, thereby extracting multiple sub-feature maps from the original feature map.

B. Dual_Conv Module

Due to the large parameter and computational resource requirements of modern CNN models, it is difficult to deploy them on embedded systems and mobile platforms.For instance, MobileNetV1 [23] tackles this issue by breaking down standard convolutions into depthwise and pointwise convolutions, which greatly reduces the parameter count and computational cost, albeit at the expense of increased network complexity.Group convolution divides the input channels into several groups, with each group performing convolution independently, and then the results of the groups are concatenated. By grouping the convolution filters and input feature map channels, this method reduces computational costs. This technique was first used in AlexNet for training with dual GPUs. Heterogeneous convolution refers to using convolution kernels of different shapes and sizes within the same layer, and then combining the convolution results. This approach allows for feature extraction at different receptive fields, enhancing the model's expressive power. Simultaneously, heterogeneous convolution integrates both 3×3 and 1×1 convolutions within a single convolution filter, reducing computational complexity. However, this approach may interrupt the continuous integration of cross-channel information, which could potentially impact the network's accuracy.

DualConv leverages the strengths of both group convolution and heterogeneous convolution by partitioning the N convolutional filters into G distinct groups.Some convolution kernels perform both 3×3 and 1×1 convolutions simultaneously, while others perform only 1×1 convolutions. This enhances information sharing and feature extraction capabilities without the need for channel shuffling operations. Compared to standard convolutions, the computational cost (FLOPs) is significantly reduced. The structural layout is shown in Figure 3, where M is the number of input channels, N is the number of convolution filters (also the number of output channels), and G is the number of groups in both group convolution and dual convolution.

In DualConv, for a given G, the ratio of combined convolution kernels with a size of (K×K + 1×1) to all channels is 1/G, while the remaining 1×1 convolution kernels have a ratio of (1 - 1/G). Therefore, in the dual convolution layer consisting of G convolution filter groups, the floating-point operations (FLOPs) for the combined convolutions kernels are:

$$FL_{CC} = (D_0^2 \times K^2 \times M \times N + D_0^2 \times M \times N) \quad (1)$$

The floating-point operations (FLOPs) of the remaining 1×1 pointwise convolution kernels are:

$$FL_{PC} = (D_0^2 \times M \times N) \times (1 - 1/G) \quad (2)$$

The total number of floating-point operations (FLOPs) is:"

$$\begin{aligned} FL_{DC} &= FL_{CC} + FL_{PC} \\ &= D_0^2 \times K^2 \times M \times N/G + D_0^2 \times M \times N \end{aligned} \quad (3)$$

Here, D_0 represents the width and height dimensions of the output feature map, M denotes the number of input feature channels,N indicates the number of convolution filters, which is also the number of output feature channels, and G represents the number of convolution filter groups. Through this design, DualConv addresses the issue of poor channel information exchange in group convolution and improves the negative impact of heterogeneous convolution on information preservation. Since convolution is applied to all input channels, it better retains the original information, helping deeper convolution layers extract features more effectively, without the need for channel shuffling operations. In terms of computational cost, the use of the group convolution strategy reduces the parameters of the original backbone network model, effectively lowering both the number of parameters and the computational cost.

C. PC_EPCA Module

Traditional channel attention mechanisms, such as the SE module in SENet, perform dimensionality reduction when learning channel attention. While this can control model complexity, it disrupts the direct correspondence between channels and weights, negatively affecting performance. Current attention mechanisms, like SE-Net [24] and CBAM [25], often focus on developing complex attention modules to improve performance, but this increases model complexity. The Efficient Channel Attention (ECA) module is a lightweight and effective mechanism designed to enhance

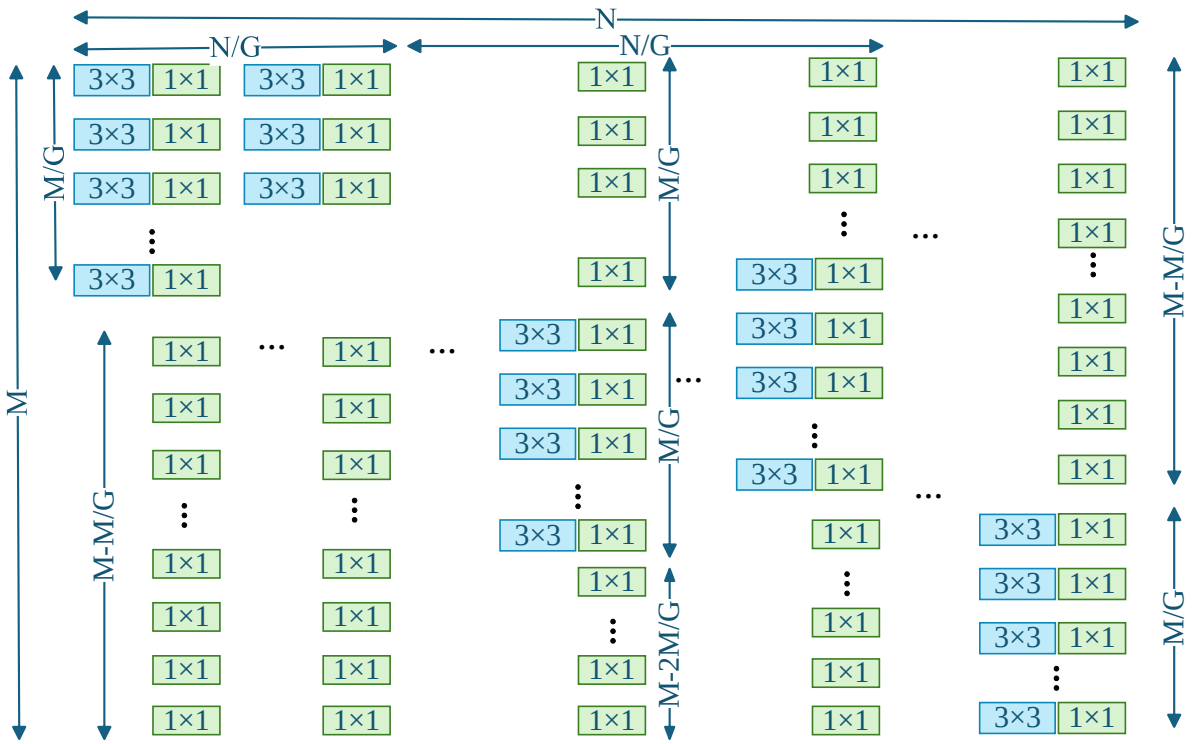

Fig. 3: Structural layout of DualConv

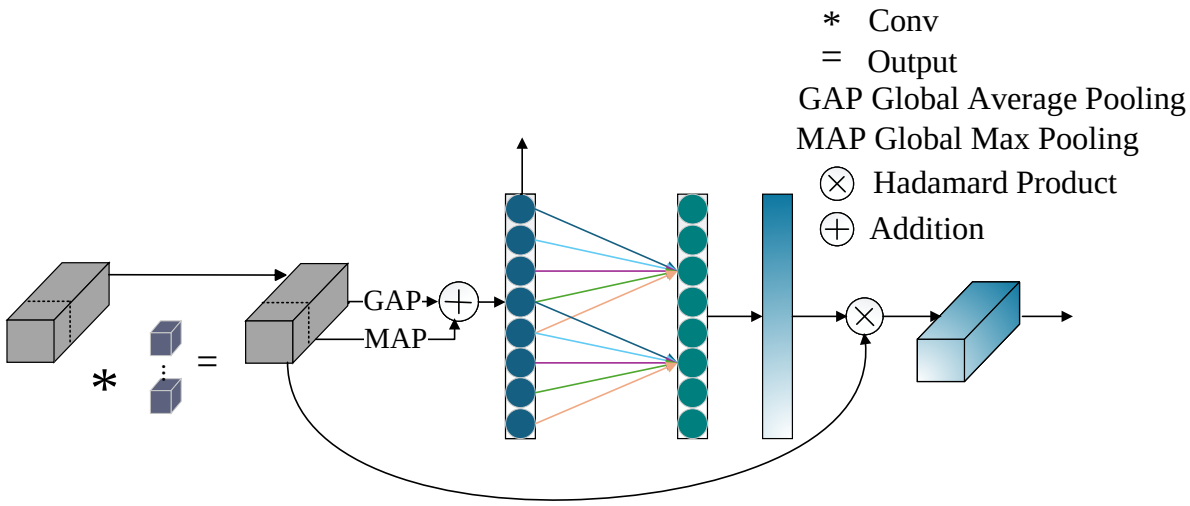

Fig. 4: Structural layout of DualConv

channel-wise feature representation in deep convolutional neural networks. The ECA module focuses on local cross-channel interactions to achieve efficient channel attention learning. With an extremely streamlined parameter configuration, it significantly strengthens the network's capacity to capture essential features and facilitates more efficient learning, significantly boosting the model's overall effectiveness and practical application value.

Firstly, the Efficient Channel Attention (ECA) module incorporates both adaptive global max pooling and adaptive global average pooling to enhance the extraction of salient channel-wise features.

Would you like the next sentence to explain how these operations contribute to attention.These operations compute the maximum and mean values across spatial dimensions of feature maps respectively, with their outputs being fused to generate a comprehensive representation of channel-wise feature statistics. Furthermore, a partial convolution (PConv) strategy is proposed to mitigate feature map redundancy. This approach performs standard convolution on only a fraction of input channels while maintaining the remaining channels intact, thereby significantly reducing computational overhead and memory access costs without compromising complete channel information preservation.The harmonious integration of these components boosts the model's performance without compromising computational efficiency.

PConv exploits the redundancy across feature map channels by performing standard convolution on only a portion of the input channels to extract spatial features, while keeping the rest of the channels intact. The floating-point

TABLE I: Ablation experiments were conducted on the HIT-UAV dataset.

Models	SPD_Conv	Dual_Conv	PC_EPCA	Params/M(%)	GFLOPS(%)	mAP50(%)	FPS
YOLOv9				48.61	237.7	87.8	22.4
YOLOv9	√			27.94	32.3	76.1	52.2
YOLOv9	√	√		27.86	123.7	86.2	33.5
YOLOv9	√	√	√	**27.23**	**123.1**	**88.9**	**22.94**

operations per second (FLOPs) for PConv are lower than those of conventional convolution, but higher than depthwise convolution and group convolution. To enable continuous or structured memory access, either the first or the last c_p contiguous channels are selected to represent the entire feature map during computation. Without loss of generality, it is assumed that the input and output feature maps have the same number of channels. Consequently, the floating-point operations (FLOPs) for PConv can be expressed as:

$$h \times w \times k^2 \times c_p^2 \tag{4}$$

Moreover, the memory access for PConv is smaller, that is

$$h \times w \times 2c_p + k^2 \times c_p^2 \approx h \times w \times 2c_p \tag{5}$$

In this context, h and w refer to the height and width of the feature map, k represents the size of the convolution kernel,c indicates the total number of channels in the input feature map, and c_p denotes the number of channels utilized for spatial feature extraction in PConv. PConv significantly reduces the computational complexity and memory access demands by decreasing the number of channels involved in computation, thus improving computational efficiency.ECA dynamically determines the kernel size k, which can be adjusted according to the number of channels. For example, $k = 5$ means that each channel interacts with its five neighboring channels. This adaptive approach eliminates the need for manually adjusting the kernel size through cross-validation, saving computational resources and enhancing the model's adaptability and generalization ability. Additionally, a fast 1D convolution of size k is used to implement local cross-channel interaction. Each channel interacts with its k neighboring channels, learning the relationships between the channels and generating channel attention weights.Compared to traditional methods, this local interaction approach requires fewer parameters and is more efficient. The output of the 1D convolution is passed through a Sigmoid activation function, which normalizes the values between 0 and 1, generating the attention weights for each channel. These attention weights are then applied to the original input feature map through element-wise multiplication, recalibrating the map to amplify the response of significant channels while attenuating the response of less relevant channels, thus enhancing the network's feature representation capabilities.

IV. DATASETS AND EVALUATION METRICS

A. Experimental Dataset

This study utilizes the HIT-UAV public dataset [26], which consists of 2,898 infrared thermal images curated from 43,470 drone video frames captured in a variety of scenarios, including campuses, parking lots, roads, and playgrounds. Designed to meet the core needs of search and rescue operations, the dataset offers detailed annotations for four key object categories: pedestrians, cars, bicycles, and special-purpose vehicles, along with a "Don'tCare" category to handle ambiguous or low-confidence detections. To ensure robust model validation, the entire dataset is randomly divided into training, validation, and test sets at an 8:1:1 ratio, thereby establishing a standardized benchmark framework for machine learning evaluation.

B. Evaluation Metrics

This study establishes a dual-dimensional evaluation framework that encompasses both detection accuracy and computational efficiency. For detection accuracy, the framework incorporates Precision, Recall, class-specific Average Precision (AP), and the mean Average Precision (mAP) across categories, with the global mAP at a 50% Intersection-over-Union (IoU) threshold (mAP50) serving as the primary metric. In terms of computational efficiency, parameters (Params), floating-point operations (FLOPs), real-time frame rate (FPS), and model storage requirements are used to systematically assess the algorithm's deployability in hardware environments, thus creating a comprehensive performance evaluation system.

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

$$mAP = \frac{\sum_{n=1}^{Num(class)} AP(n)}{TP + TN + FP + FN} \tag{8}$$

TP (True Positive) represents the number of correctly identified positive samples, meaning instances where the model accurately classifies actual positive cases. FP (False Positive) indicates the number of negative samples that the model mistakenly identifies as positive. FN (False Negative) refers to the number of positive samples that the model incorrectly classifies as negative. The number of parameters (Params) indicates the total count of learnable weights within the model, serving as a primary measure of model complexity—larger parameter counts generally imply a more complex model. GFLOPs (Giga Floating Point Operations) quantify the computational load required for inference or training, reflecting the model's demand on processing resources; higher GFLOP values correspond to increased computational requirements. FPS (Frames Per

TABLE II: Comparative experiments on attention mechanisms

Model	Params/M(%)	GFLOPS(%)	FPS(%)	mAP50(AP%)	mAP50-95(AP%)
YOLOv9	48.6	237.7	22.4	87.8	59.1
YOLOv9+SPD+Dual+MCA	27.24	123.3	35.1	86.6	55.9
YOLOv9+SPD+Dual+DASI	29.13	123.7	35.5	87.1	56.3
YOLOv9+SPD+Dual+NAM	27.23	123,2	27.5	86.1	55.7
YOLOv9+SPD+Dual+CBAM	27.45	123.3	29.1	86.4	57.1
YOLOv9+SPD+Dual+Skattention	45.77	138.7	26.3	85.0	55.2
YOLOv9+SPD+Dual+DoubleAttention	27.79	123.6	28.7	87.2	56.7
Ours	**27.2**	**123.1**	**26.9**	**88.9**	**61.3**

Second) evaluates the model's processing speed by indicating how many image frames can be processed each second, with higher FPS values denoting stronger real-time performance.

V. EXPERIMENTS AND ANALYSIS

A. Ablation Study

In order to verify the effectiveness of each improved module in this paper on YOLOv9, we gradually incorporated the improved modules into YOLOv9 for ablation experiments. As shown in Table 1, first, the original downsampling layer of the model was replaced with the SPD_Conv module. After the replacement, although the mAP of the model decreased by 11.5%, the number of parameters dropped by 27.94M, with a reduction rate of 42.5%, and the amount of computation decreased significantly to 32.3 GFLOPS, while the FPS increased by 29.8%. Next, the original convolutional layer was changed to the DualConv convolutional module. After the improvement, the mAP increased by 10.1%, the number of parameters dropped to 27.86M, the amount of computation became 123.7 GFLOPS, and the FPS was 33.5. Finally, the improved attention mechanism PC_EPCA was introduced into the backbone. As a result, the mAP increased by 2.7%, the number of parameters dropped to 27.23M, the amount of computation decreased to 123.1 GFLOPS, and the FPS was 26.94. The results of the ablation experiments demonstrate that the enhanced model presented in this paper not only improves accuracy but also significantly reduces computational complexity and the number of parameters. This indicates that the improved model holds notable research value for infrared small target detection.

B. Comparison Experiment On Attention Mechanisms.

To evaluate the effectiveness of the enhanced attention mechanism, PC_EPCA, this paper conducted comparison experiments across various attention mechanisms based on prior improvements, as presented in Table 2. The table clearly shows that, when other attention mechanisms are applied, both the number of parameters and computational cost exceed those of PC_EPCA, and accuracy tends to decrease. In contrast, the PC_EPCA mechanism not only reduces the number of parameters and computation but also improves mAP50 by 1.1%. This demonstrates that by optimizing convolution and integrating global max pooling with adaptive global average pooling, the PC_EPCA mechanism allows the model to capture both local and global features more efficiently, significantly minimizing information loss. Moreover, it can capture local and global features more effectively, significantly enhancing the performance in complex backgrounds. Meanwhile, by reducing computational redundancy, the model has achieved remarkable results in reducing the amount of computation and the number of parameters, effectively promoting the lightweight process of the model.

TABLE III: Comparison of Methods on the HIT-UAV Dataset

Model	Params/M(%)	GFLOPS(%)	FPS(%)	mAP50(AP%)
Faster-RCNN	48.1	176.8	37.6	76.8
SSD	46.9	138.5	81.2	85.6
YOLOv5s	8.7	23.8	87.3	79.4
YOLOv6	4.04	26.3	82.5	72.3
YOLOv8	10.6	30.4	78.7	79.8
YOLOv9	48.6	237.7	22.4	87.8
YOLOv10s	7.7	29.7	40.3	83.7
RTDETR	30.5	27.79	23.6	85.4
Ours	**27.2**	**123.1**	**26.9**	**88.9**

C. Comparative Experiment

To further validate the performance of the improved model algorithm proposed in this paper, several classic models were selected for comparative experiments, with the results presented in Table 3. As shown in the table, the improved model demonstrates a substantial reduction in both the number of parameters and computational cost compared to classic models such as Faster R-CNN, SSD, and RT-DETR, while also achieving an improvement in accuracy.By comparing with the models such as YOLOv5s, YOLOv6, YOLOv8, and YOLOv10s, it can be observed that the mAP50 has been significantly improved, with increases

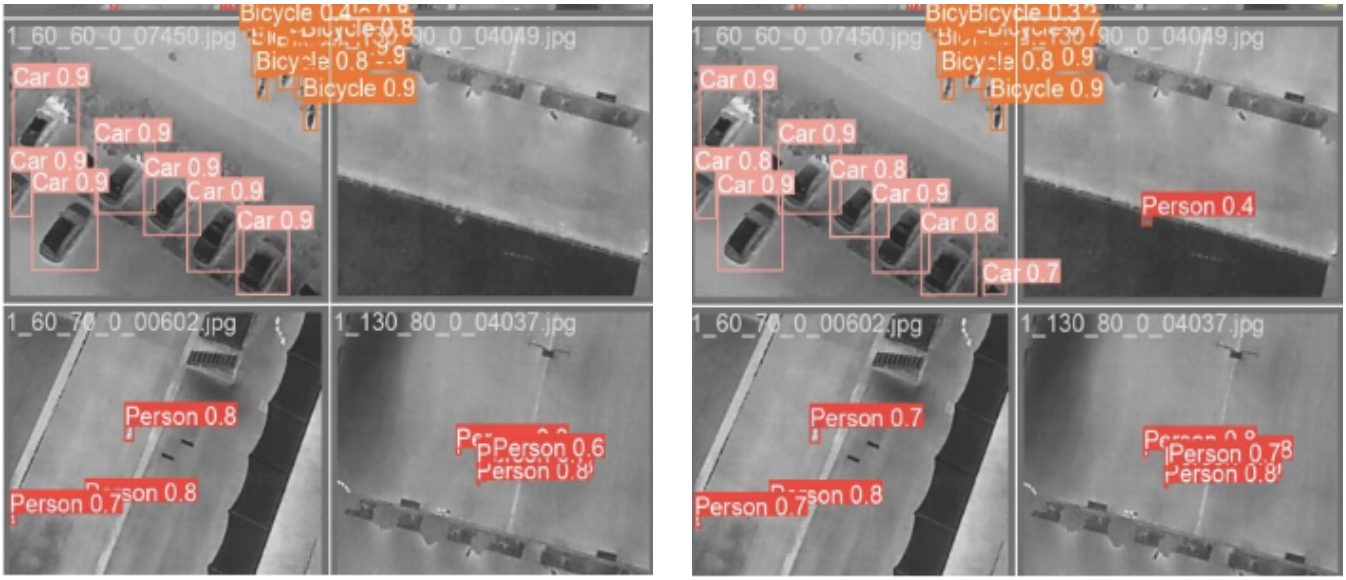

Fig. 5: Comparison of detection effects

of 9.5%, 16.6%, 9.1%, and 5.2% respectively.In comparison with YOLOv9, the improved model presented in this paper achieves a 1.1% increase in mAP50, reduces the number of parameters by 21.4 million, and decreases computational cost by 114.6 GFLOPS. These results clearly demonstrate that the enhanced YOLO model offers significant advantages in terms of model lightweighting.It not only ensures lightweighting but also improves the accuracy, reduces the usage cost, and enhances the practicability of the model.

VI. Visual Analysis Of Detection Effects

To verify the effectiveness of the improved model algorithm in this paper compared to YOLOv9, two sets of images from the HIT-UAV dataset were selected for comparison of detection performance under different models. Figure 5(a) presents the detection results using the YOLOv9 model, while Figure 5(b) illustrates the detection performance of the improved model algorithm. As shown in the figures, the YOLOv9 model tends to miss the detection of small targets in complex backgrounds. In contrast, the improved model enhances the ability to extract features from small targets, thereby improving detection accuracy. The improved algorithm effectively addresses the challenge of detecting tiny targets, boosts detection precision, reduces computational redundancy, and significantly lowers both the number of parameters and computational load. This model meets the real-time performance requirements and is better suited for infrared small target detection applications.

VII. Conclusion

In the field of infrared small target detection, issues such as real-time detection effects and the tendency of missing detections for small targets in complex backgrounds have always attracted much attention. In response to this, this paper proposes an innovative lightweight algorithm based on enhancements to YOLOv9.Specifically, a series of measures were taken during the improvement process. Firstly, the adaptive downsampling layers of the original model were uniformly replaced with SPD_Conv. Meanwhile, some of the original convolutional layers were replaced with dual convolutions. Moreover, an original attention mechanism PC_EPCA was introduced in the backbone part. Through these improvement measures, the algorithm has achieved optimizations in multiple aspects.The proposed algorithm not only substantially reduces computational cost and parameter count but also achieves a notable improvement in detection accuracy. According to the experimental results on the public HIT-UAV dataset, the improved algorithm demonstrates outstanding performance—achieving a 1.1% increase in mAP50, a 43.9% reduction in the number of parameters, and a 48.2% decrease in computation. These results indicate that the algorithm's real-time detection capability has been significantly enhanced, along with its effectiveness in detecting small targets in complex backgrounds.

While the proposed algorithm achieves promising performance, certain aspects still require further improvement. Future research will aim to further enhance the detection of small targets and improve overall detection accuracy, thereby promoting the continued evolution of infrared small target detection algorithms.

References

[1] Rawat S S, Verma S K, Kumar Y. Review on recent development in infrared small target detection algorithms[J]. Procedia Computer Science, 2020, 167: 2496-2505.

[2] Ma J, Tang L, Xu M, et al. STDFusionNet: An infrared and visible image fusion network based on salient target detection[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-13.

[3] Zhao M, Li W, Li L, et al. Single-frame infrared small-target detection: A survey[J]. IEEE Geoscience and Remote Sensing Magazine, 2022, 10(2): 87-119.

[4] Zhang M, Wang Y, Guo J, et al. IRSAM: Advancing segment anything model for infrared small target detection[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 233-249.

[5] Mathew A, Amudha P, Sivakumari S. Deep learning techniques: an overview[J]. Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2020, 2021: 599-608.

[6] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.
[7] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448
[8] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.
[9] Terven J, Córdova-Esparza D M, Romero-González J A. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas[J]. Machine learning and knowledge extraction, 2023, 5(4): 1680-1716.
[10] Yu C, Liu Y, Wu S, et al. Infrared small target detection based on multiscale local contrast learning networks[J]. Infrared Physics
[11] Zhao J, Yu C, Shi Z, et al. Gradient-guided learning network for infrared small target detection[J]. IEEE Geoscience and Remote Sensing Letters, 2023, 20: 1-5.
[12] Wu X, Hong D, Huang Z, et al. Infrared small object detection using deep interactive U-Net[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5.
[13] Hou Q, Wang Z, Tan F, et al. RISTDnet: Robust infrared small target detection network[J]. IEEE Geoscience and Remote Sensing Letters, 2021, 19: 1-5.
[14] Wang Z, Zang T, Fu Z, et al. RLPGB-Net: Reinforcement learning of feature fusion and global context boundary attention for infrared dim small target detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-15.
[15] Chen C, Zheng Z, Xu T, et al. Yolo-based uav technology: A review of the research and its applications[J]. Drones, 2023, 7(3): 190.
[16] Sunkara R, Luo T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects[C]//Joint European conference on machine learning and knowledge discovery in databases. Cham: Springer Nature Switzerland, 2022: 443-459.
[17] Zhong J, Chen J, Mian A. DualConv: Dual convolutional kernels for lightweight deep neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 34(11): 9528-9535.
[18] Wang C Y, Yeh I H, Mark Liao H Y. Yolov9: Learning what you want to learn using programmable gradient information[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2024: 1-21..
[19] Yaseen, M. What is YOLOv9: An In-Depth Exploration of the Internal Features of the Next-Generation Object Detector[J]. arXiv preprint arXiv:2409.07813, 2024.
[20] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.
[21] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[22] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
[23] Howard A G, Zhu M, Chen B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[J]. 2017,arXiv preprint arXiv:1704.04861.
[24] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.
[25] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional Block Attention Module[J]. European Conference on Computer Vision (ECCV), 2018: 3-19.
[26] Suo J, Wang T, Zhang X, et al. HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection[J]. Scientific Data, 2023, 10(1): 227.