Journal of Beijing Institute of Technology, 2026, Vol. 35, No. 5

Enhanced Infrared Target Detection: Integrating Anti-Interference and Multi-Scale Perception

Meijing Gao[1,2]✉, Bingzhou Sun[1], Sibo Chen[1,3]✉, Huanyu Sun[1], Kunda Wang[1], Xu Chen[1], Yufan Du[1]

(1. School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing 100081, China; 2. State Key Laboratory of Environment Characteristics and Effects for Near-Space, Beijing 100081, China; 3. Beijing Institute of Aerospace Intelligence and Information, Beijing 100854, China)

Abstract: Infrared target detection using unmanned aerial vehicles (UAVs) presents a frontier in intelligent infrared recognition with diverse applications. However, challenges such as complex noise interference and drastic target scale variations hinder practical implementation. This study introduces an infrared target detection algorithm that integrates anti-interference and multi-scale perception mechanisms. The algorithm features a vision Transformer (ViT)-based anti-interference feature encoder and a multi-scale feature encoder with deformable attention. Experimental results on the HIT-UAV and VT-MOT datasets demonstrate a 4.5 and 5.3 percentage point improvement in small-target detection accuracy, respectively, highlighting its robust anti-interference capabilities and high practical applicability.

Keywords: infrared target detection; anti-interference mechanism; multi-scale perception; edge awareness; gated fusion network

1 Introduction

Driven by the rapid advancement of infrared imaging and deep learning, deep learning-based infrared target detection has been widely applied across diverse platforms, particularly on unmanned aerial vehicles (UAVs), which has become a research frontier.

UAV-based infrared target detection confronts two critical challenges. First, infrared imaging is highly vulnerable to complex noise interferences, including environmental and thermal radiation noise. These factors blur target edges and diminish thermal contrast in infrared images, resulting in poor robustness of existing detection algorithms under noisy conditions. Second, fluctuations in UAV flight altitude and attitude cause abrupt changes in target scale within images. Modeling this dynamic scale variation effectively remains challenging for detection algorithms, thereby compromising detection accuracy and adaptability.

To improve robustness under noisy infrared imaging, we build an anti-interference branch based on a vision Transformer (ViT) backbone and design two task-specific auxiliary pre-training tasks: interference-aware learning and edge-aware learning. Specifically, Poisson-Gaussian noise, atmospheric turbulence, and motion blur are used to guide noise-invariant feature learning, and edge-enhanced views are introduced to strengthen weak boundary perception.

To address target scale variation, we develop

Manuscript received Apr. 23, 2026; revised Jun. 10, 2026; accepted Jun. 21, 2026. The associate editor coordinating the review of this manuscript was Dr. Hao Huan. This work was supported by the National Natural Science Foundation of China (No. 62471034), and the Sichuan Science and Technology Program (No. 2025NSFSC2078)

✉ Corresponding author. Email: gaomeijing@126.com, sibochen@yeah.net

DOI: 10.15918/j.jbit1004-0579.2026.031

a multi-scale feature encoder by adapting RepGFPN (reparameterized generalized feature pyramid network), SimAM (simple, parameter-free attention module), and deformable attention to UAV infrared detection. These existing components are used to enhance cross-scale feature interaction, salient-region response, and small-target representation. To integrate the anti-interference and multi-scale features, we design a spatially aligned gated fusion module, where the ViT output sequence is aligned with the C3 feature map and adaptively fused through position-wise and channel-wise gates. The fused features are then decoded by a Transformer decoder with content-position decoupling.

The main contributions of this work are summarized as follows. First, building upon ViT and contrastive learning, we design an infrared interference-aware learning task for UAV infrared detection. By constructing noisy positive views with Poisson-Gaussian noise, atmospheric turbulence, and motion blur, the encoder is encouraged to learn noise-invariant target representations. Second, based on edge enhancement and suppression operations, we introduce an edge-aware learning task to enhance the perception of weak target boundaries in low-contrast infrared scenes. Third, we propose a spatially aligned gated fusion module to integrate anti-interference features and multi-scale features. The gate assigns position-wise and channel-wise fusion weights, enabling the detector to balance noise robustness and scale perception under complex UAV imaging conditions.

2 Related Work

Infrared target detection, as a crucial sub-discipline of general object detection, has followed a similar evolutionary path to mainstream object detection techniques. Modern object detection algorithms can be broadly categorized into two frameworks: convolutional neural networks (CNNs) and Transformer-based architectures. The CNN-based framework is further divided into two-stage and one-stage algorithms. Representative two-stage methods include Fast R-CNN (region-based convolutional neural networks) [1], Faster R-CNN [2], and Mask R-CNN [3]. However, their inferior inference efficiency renders them ill-suited for UAV-based detection tasks.

One-stage methods are further categorized into anchor-based and anchor-free paradigms. Representative anchor-based methods include the YOLO (you only live once) series [4–8] and SSD (single shot multibox detector) [9]. Although these methods have demonstrated remarkable performance, their dependency on anchor box priors and complex hyperparameter optimization constrains their adaptability. Consequently, anchor-based methods have been increasingly marginalized in UAV infrared target detection. To streamline the pipeline, anchor-free methods have emerged, encompassing keypoint-based and center-based approaches. Keypoint-based representatives include CornerNet [10], ExtremeNet [11], and CenterNet [12]; center-based representatives include DenseBox [13], GARPN [14], FCOS [15], and FoveaBox [16]. These anchor-free approaches eliminate the need for predefined anchor boxes, reduce hyperparameter dependency, enhance flexibility in handling scale variation and irregular targets, and enable faster inference. However, they often exhibit compromised localization accuracy in dense prediction scenarios, require meticulous design of positive/negative sample assignment strategies, and face challenges in small-object detection.

In recent years, numerous YOLO-based improved algorithms for UAV infrared target detection have been proposed, yielding promising results. Aiming to tackle challenges like low accuracy, large model footprint, and sluggish inference in UAV infrared scenarios, Cao et al. [17] proposed YOLO-TSL, a lightweight detection algorithm. Tang et al. [18] proposed IRSTD-

YOLO for small infrared target detection in low-contrast, complex-background environments. However, these CNN-based methods are inherently constrained by CNNs' limitations—specifically, limited receptive fields and inadequate global semantic modeling—leading to subpar anti-interference performance.

To overcome these limitations, numerous algorithms have embraced the Transformer architecture, initially proposed by Vaswani et al. [19] in 2017. Dosovitskiy et al. [20] introduced the vision Transformer (ViT), which revolutionized the modeling of global image relationships, yielding superior performance on various vision tasks. Carion et al. [21] further proposed DETR (detection Transformer), the first object detection model to integrate the Transformer architecture. However, both ViT and DETR suffer from high computational complexity, slow training convergence, and poor real-time applicability to UAV scenarios.

To address these challenges, Zhu et al. [22] proposed deformable DETR, which mitigated global attention's computational overhead via deformable attention mechanisms and incorporated multi-scale feature maps to enhance small-object detection—demonstrating efficacy in UAV infrared scenarios. Meng et al. [23] proposed conditional DETR, which decouples decoder query content and positional information to expedite training. In 2024, Zhao et al. [24] proposed RT-DETR, the first real-time DETR-based detector, which substantially enhanced multi-scale detection performance. Building on RT-DETR's superior detection performance, numerous recent approaches have extended it for UAV infrared detection. Wang et al. [25] proposed PHSI-RTDETR, which integrates a lightweight RPConv-Block module to boost detection accuracy while substantially reducing model complexity and computational overhead. Additionally, they designed the AIFI-HiLo module to enhance dense-scene performance and minimize miss and false detections. Xue et al. [26] proposed FECI-RTDETR, which integrates the EA-AIFI module to boost small-object detection performance while cutting computational costs.

Despite significant advancements in UAV infrared target detection, existing algorithms generally lack efficient mechanisms for suppressing infrared noise, resulting in subpar anti-interference capabilities. Furthermore, although Transformer-based detectors like deformable DETR enable multi-scale feature interaction, their performance still lags behind traditional feature pyramid networks, rendering them ill-suited for UAV detection tasks. Thus, this study focuses on enhancing anti-interference mechanisms and multi-scale perception in infrared scenarios, aiming to improve robustness and adaptability for UAV operations in complex environments.

3 Methods

3.1 Anti-Interference Feature Encoder

Fig. 1 depicts the overall architecture of the proposed algorithm. The input is a raw infrared image, processed by two specially designed encoder branches: the multi-scale feature encoder and the ViT-based anti-interference encoder. The multi-scale feature encoder comprises a YOLOv11 backbone, a RepGFPN multi-scale fusion network, and a deformable attention mechanism. The ViT-based anti-interference encoder is grounded in a vision Transformer backbone, pre-trained via two auxiliary learning tasks —interference-aware learning and edge-aware learning. The multi-scale encoder generates multi-scale feature sequences, whereas the ViT encoder produces anti-interference feature sequences. These two sequences are first aligned and then fused via a gated fusion network, with the fused feature sequence fed into the decoder. The decoder employs a content-position decoupling mechanism to decode and ultimately output target positions and categories.

The structure of the ViT anti-interference feature encoder is illustrated in Fig. 2. The input

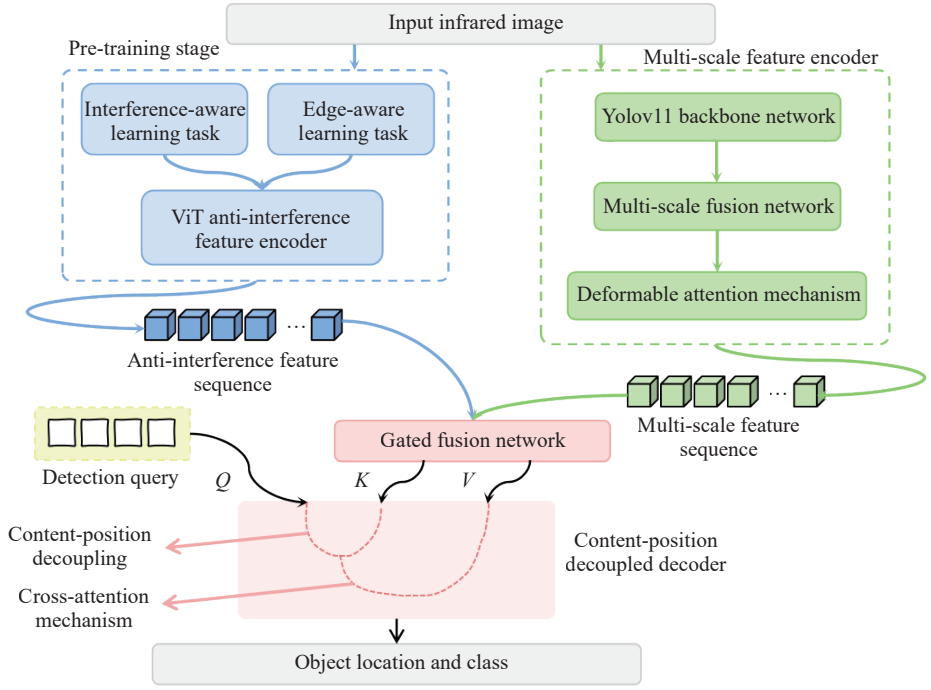


Fig. 1 Structure diagram of infrared target detection algorithm integrating anti-interference mechanism and multi-scale fusion

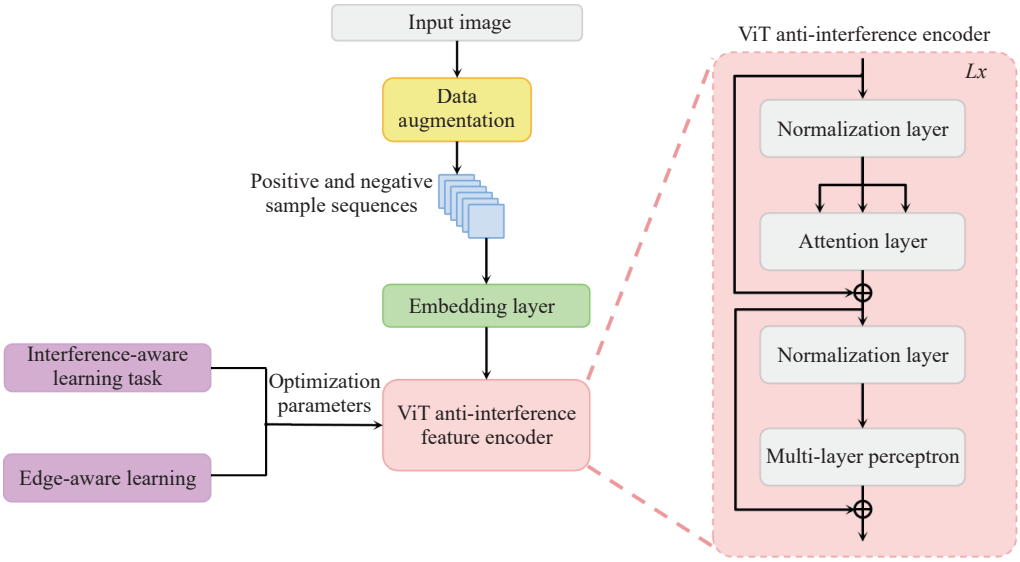


Fig. 2 Structure diagram of ViT anti-interference feature encoder

is the raw infrared image, which undergoes an augmented data generation process using manually designed strategies to construct positive and negative sample sequences. These positive and negative sample sequences are then passed through the ViT embedding layer, converting the image data into sequence data. The sequence data are processed by the ViT anti-interference feature encoder to generate encoded image feature sequences. These feature sequences are further optimized through two downstream tasks: interference-aware learning and edge-aware learning, which fine-tune the ViT pre-trained network parameters.

3.2 Design of the Interference-Aware Learning Task

The interference-aware learning task is pivotal for UAV infrared target detection. It empowers the backbone network to discriminate between target intrinsic features and noise interference,

thereby constraining the network's lower layers to filter out noise-induced pseudo-features and enabling the model to focus on authentic target features. The concept of interference-aware learning originates from deep metric learning, aiming to map inputs to an embedding space through a neural network. In this embedding space, the inter-sample distance directly reflects their semantic similarity. The specific design approach will be elaborated on from three dimensions: positive/negative sample construction, model architecture, and optimization objectives.

1) Positive and Negative Sample Construction: Positive samples denote instances from the same category or target, whereas negative samples originate from distinct categories or targets. Through contrasting positive/negative samples, the model minimizes intra-class distances and maximizes inter-class distances in the feature space, thereby constructing a discriminative embedding representation.

For the original input image, three types of synthetic noise —Poisson-Gaussian noise, atmospheric turbulence noise, and motion blur —are applied. These three noise types are selected because they simulate common interferences in infrared imaging systems: sensor noise, atmospheric transmission interference, and platform motion interference. This combination encompasses the primary interference sources in UAV infrared imaging, enabling the model to learn feature representations with enhanced generalization capability. Specifically, Poisson-Gaussian noise simulates the inherent noise characteristics of infrared sensors, atmospheric turbulence noise represents long-distance imaging air disturbance, and motion blur represents image degradation caused by vibrations from the UAV platform. After adding interference to the original image, the positive sample P is defined as follows

$$\begin{aligned} P(x) &= \left[T_1(x),\, T_2(x),\, T_3(x)\right] \\ T_i &\in \{T_{\text{noise}},\, T_{\text{turb}},\, T_{\text{blur}}\} \end{aligned} \tag{1}$$

where T_i corresponds to Poisson–Gaussian noise, atmospheric turbulence, and motion blur, respectively.

Negative samples $N(x)$ are obtained through in-batch negative sampling, as shown in

$$N(x) = \left\{x_j \mid j \neq i,\ x_j \in B\right\} \tag{2}$$

where B represents the current batch. The positive samples are generated as $y^+ = T(x)$, while the negative samples $\{y_k\}$ are collected from other images in the same mini-batch.

2) Model Architecture: The model employs a dual-branch architecture. The original image branch focuses on extracting target intrinsic features, whereas the interference sample branch specializes in modeling noise patterns. The model structure is illustrated in Fig. 3.

The model inputs comprise the original image, positive samples, and negative samples. After traversing the lower layers of the ViT anti-interference feature encoder, both branches leverage SENet (squeeze and excitation networks) to process the encoded feature sequences. SENet dynamically recalibrates channel feature importance, augmenting the model's sensitivity to critical features.

Finally, global average pooling (GAP) is applied to generate feature vectors. The GAP operation collapses spatial dimensions, yielding the original image feature vector $\boldsymbol{Z}_1$, positive sample feature vector $\boldsymbol{Z}_2$, and negative sample feature vector $\boldsymbol{Z}_3$. These feature vectors are then utilized for subsequent loss computation. Model parameters of the ViT anti-interference encoder are optimized via loss gradient backpropagation.

3) Optimization Objective: Guided by the principles of deep metric learning, the model's optimization objective is to maximize the mutual information between the original image and positive samples. Mutual information is a measure of the statistical dependence between two random variables, and in this case, the InfoNCE (information noise-contrastive estimation) loss function is used. The final loss function is defined as

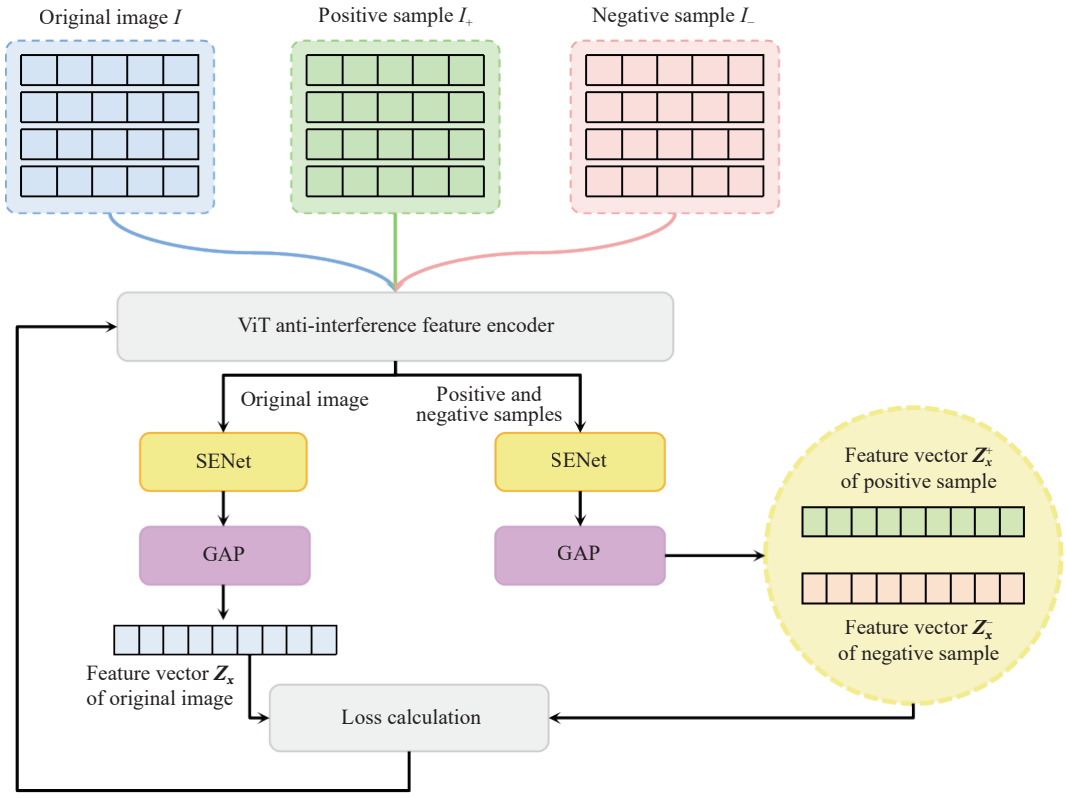


Fig. 3 Structure diagram of interference-aware learning network

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}\left[\lg \frac{\exp(\text{sim}(\boldsymbol{z}_x, \boldsymbol{z}_{y^+})/\tau)}{\sum_{k=0}^{K}\exp(\text{sim}(\boldsymbol{z}_x, \boldsymbol{z}_{y_k})/\tau)}\right] \quad (3)$$

where $\text{sim}(\cdot)$ denotes the similarity function for measuring feature space similarity (cosine similarity is applied here), τ is the temperature coefficient, and $\boldsymbol{z}_x = f(x)$ and $\boldsymbol{z}_y = f(y)$ are the encoded feature vectors. The denominator comprises the sum of similarities between one positive sample and K negative samples.

3.3 Edge-Aware Learning Task Design

In infrared scenes, target edges serve as critical features for discriminating targets from the background. The proposed edge-aware learning enhances the model's ability to perceive high-frequency information in the image through explicit feature crossovers, enabling more effective target localization. The following subsections elaborate on the design approach across three dimensions: positive/negative sample construction, model architecture, and optimization objectives.

1) Positive and Negative Sample Construction: The core principle of positive/negative sample construction is to separate target high-frequency edge information from image low-frequency noise interference. Positive samples are generated via the Perona-Malik diffusion equation (PMD) to emphasize target edges, as defined as

$$\frac{\partial I}{\partial t} = \text{div}\left(c(|\nabla I|)\nabla I\right) \quad (4)$$

where I denotes the input image, ∇I denotes the image gradient magnitude, and the diffusion coefficient is defined as

$$c(|\nabla I|) = \frac{1}{1 + (|\nabla I|/K)^2} \quad (5)$$

Parameter K governs edge sensitivity. When the gradient surpasses K, the diffusion coefficient approaches 0, suppressing diffusion; conversely, when the gradient is below K, the coefficient approaches 1, enhancing diffusion. The image processed by the PMD is presented in Fig. 4(b). The edge-enhanced image is obtained

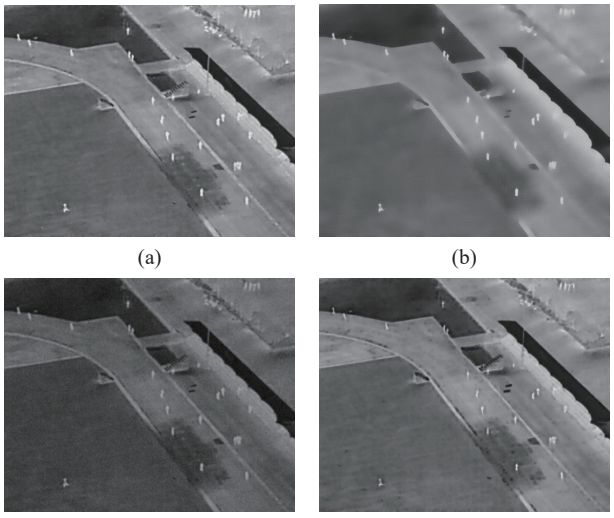


Fig. 4 Examples of generated views for auxiliary learning tasks: (a) original image; (b) edge-enhanced image generated by PMD; (c) noisy image generated by Poisson–Gaussian degradation; (d) edge-suppressed image generated by grayscale erosion

by applying PMD to the original infrared image. PMD smooths homogeneous background regions while preserving high-gradient structures, thereby strengthening weak target boundaries. The edge-suppressed image is generated by applying a grayscale erosion operation to weaken fine boundary responses and small bright structures. In the edge-aware learning task, the pair composed of the original image and the edge-enhanced image is treated as a positive pair, whereas the pair composed of the original image and the edge-suppressed image is treated as a negative pair. This design forces the encoder to distinguish boundary-preserving feature interactions from boundary-suppressed ones.

2) Model Architecture: The network architecture for edge-aware learning is depicted in Fig. 5. The model inputs comprise the original image, edge-enhanced image, and edge-suppressed image. After traversing the lower layers of the ViT anti-interference feature encoder, SENet is leveraged again to recalibrate channel feature importance. A global average pooling (GAP) operation then yields three feature vectors: the

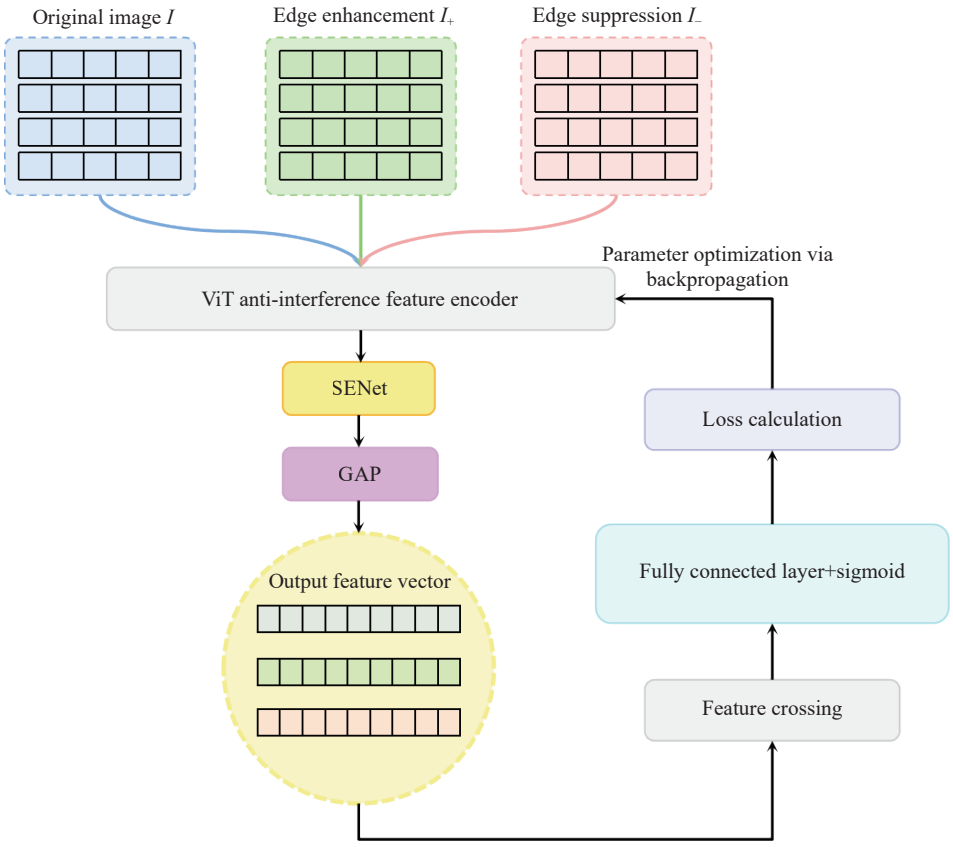


Fig. 5 Structure diagram of edge-aware learning network

original image features Z_X, edge-enhanced features Z_{x+}, and edge-suppressed features Z_{x-}.

The feature crossing layer performs a Hadamard product, where Z_X and Z_{x+} are element-wise multiplied to construct feature interaction terms. Finally, the crossed features are fed into a fully connected layer followed by a sigmoid activation function to produce the output probability, as specified in

$$A_{\text{edge}} = \sigma\big(\text{MLP}(z_x \odot z_y)\big) \tag{6}$$

where $\odot$ denotes the Hadamard product, σ represents the sigmoid activation function, the MLP (multi-layer perceptron) has an output dimension of 1, and A_{edge} denotes the final output probability.

3) Optimization Objective: The feature interaction terms between x and y are treated as one class, whereas those between x and z are considered as another class. A binary cross-entropy loss function is employed to guide model learning, as defined in

$$\mathcal{L}_{\text{BCE}} = -\sum_{i=1}^{N}\Big[y_i \lg\big(A^{+}_{\text{edge},i}\big) + \big(1-y_i\big)\lg\big(1-A^{-}_{\text{edge},i}\big)\Big] \tag{7}$$

This loss function forces the model to accurately discriminate between the interaction terms of positive and negative samples, thus enhancing the model's ability to perceive edge features and further improving the algorithm's anti-interference performance.

3.4 Multi-Scale Feature Encoder Design

The architecture of the multi-scale feature encoder is shown in Fig. 6. After inputting the infrared image, the backbone network first extracts feature maps at three distinct scales. These feature maps are then fused using the reparameterized feature pyramid network (RepGFPN) for multi-scale feature fusion. Finally, the fused features undergo multi-scale feature encoding through a deformable attention mechanism.

The deformable attention mechanism embodies a sparse interaction paradigm, where each query point interacts with a constrained number of sampling points across scales, resulting in inadequate contextual modeling. To tackle this limitation, a RepGFPN is embedded preceding the deformable attention module, facilitating comprehensive multi-scale feature fusion and cross-scale interaction. This effectively compensates for the sparse attention mechanism's inability to model cross-scale global dependencies.

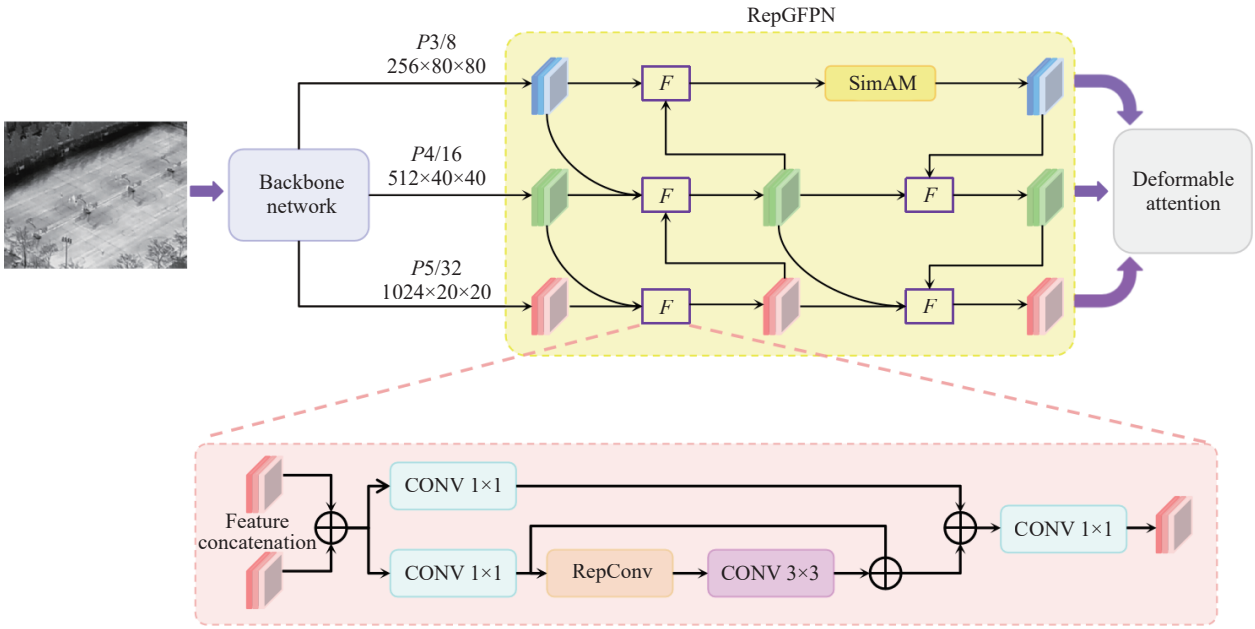


Fig. 6 Structure diagram of multi-scale feature encoder

The fusion module follows a CSP (cross-stage partial) architecture. Building upon this, the RepVGG re-parameterization mechanism is integrated to transform the multi-branch training architecture into an efficient single-path inference topology. This reduces the inference cost while retaining the multi-scale information representation capability.

During multi-scale feature fusion, the low-level *C*3 feature map—characterized by high resolution and rich detail—plays a pivotal role in small-object detection. However, the *C*3 layer is prone to low-level interferences (e.g., background noise, edge blur), which diminishes its semantic representation capability. Thus, the challenge of extracting effective information from the *C*3 layer while suppressing redundant interferences during fusion becomes critical for enhancing multi-scale feature interaction performance.

To solve this, the SimAM (simple attention module) is introduced at the output end of the high-resolution *C*3 feature map to enhance the response of its salient regions. SimAM operates without additional trainable parameters, featuring computational efficiency and high integrability. Applying SimAM to the *C*3 layer not only enhances fine-grained target representation but also prevents substantial computational overhead.

3.5 Gated Fusion Network Module

The anti-interference feature encoder and multi-scale feature encoder output anti-interference and multi-scale feature sequences, respectively. Due to their distinct structures, the two encoders produce different feature sequences. Directly fusing them would result in misalignment of position information and confusion in semantic correspondence. Therefore, spatial dimension alignment of the two feature sequences is required before fusion, realized by adjusting the image patch size in the ViT encoder.

Specifically, the patch size of the ViT embedding layer is set to 8×8, ensuring that the receptive field of each image patch corresponds to the *C*3 feature map from the backbone network. With an input image size of 640×640, the length l of the ViT output sequence matches the length l of the flattened *C*3 multi-scale feature map, ensuring spatial position alignment.

Following feature alignment, the key challenge is to effectively fuse anti-interference and multi-scale features. Traditional fusion methods can be broadly categorized into: 1) Direct addition of feature vectors at corresponding positions, which assumes equal importance for both feature types and hinders dynamic adjustment for diverse scenes. 2) Feature dimension concatenation with joint encoding, which preserves complete information but increases feature dimension and computational overhead.

To solve these problems, this paper proposes a probability-based gated fusion strategy. This strategy employs an MLP (multi-layer perceptron) to generate gating weights, dynamically allocating contribution ratios for features from the two encoders at each spatial position. Compared with traditional methods, the gating mechanism enables automatic adjustment of fusion ratios for semantic features at different positions, enhancing flexibility and adaptability to complex backgrounds and diverse targets.

The aligned anti-interference feature sequence F_{VIT} and multi-scale feature sequence F_{DETR} are concatenated along the channel dimension to derive the fused feature sequence $F_{\text{fusion}} = [F_{\text{VIT}}, F_{\text{DETR}}] \in \mathbb{R}^{L\times 2d}$. An MLP is employed to generate the gating weights $G \in \mathbb{R}^{L\times d}$, which regulate the feature importance at each spatial position. This process is formulated as

$$G = \sigma\left(\text{MLP}\left(F_{\text{fusion}}\right)\right) \tag{8}$$

Here, σ denotes the sigmoid function, and the MLP output dimension is d. Element-wise multiplication of the gating weights G with the original features achieves dynamic fusion, as pre-

sented in

$$F = G \odot F_{\text{VIT}} + \left(1 - G\right) \odot F_{\text{DETR}} \tag{9}$$

The gated fusion weights are optimized jointly with the detection loss. For each spatial position l and channel c, the gradient of the detection loss with respect to the gate can be written as

$$\frac{\partial \mathcal{L}_{\text{det}}}{\partial G_{l,c}} = \frac{\partial \mathcal{L}_{\text{det}}}{\partial F_{l,c}} \left(F_{\text{VIT},l,c} - F_{\text{DETR},l,c}\right) \tag{10}$$

This indicates that the gate is directly supervised by the final detection objective. When the anti-interference feature contributes more to reducing the detection loss, the corresponding gate value is increased; otherwise, the model assigns a larger weight to the multi-scale feature. In this way, the fusion module learns adaptive feature importance in a data-driven manner.

The fused feature F is used as the $C3$-layer feature to replace the original multi-scale feature F_{C3}, resulting in the final output $O_{\text{fused}} \in \mathbb{R}^{L \times d}$ of the integrated encoder. This feature sequence integrates anti-interference and multi-scale information, enabling robust semantic representation.

4 Results and Discussion

To verify the effectiveness of the proposed algorithm, three datasets are used for evaluation: HIT-UAV, Drone-Vehicle, and VT-MOT. Dataset details are presented in Tab. 1. Among them, the input image size is 640×640, the ViT patch size is 8×8, the number of encoder and decoder layers is 3. We use the Adam optimizer during training, where the epoch is configured at 100, the batch size is set to 8, and the learning rate is established at 0.001. Our proposed method is executed on an RTX4090 GPU.

4.1 Downstream Task Performance Analysis Experiments

To evaluate the impact of two downstream tasks (Task 1 and Task 2) on the final detection performance, ablation experiments were conducted on the HIT-UAV and VT-MOT datasets. The results are presented in Tab. 2. Task 1 denotes interference-aware learning, whereas Task 2 corresponds to edge-aware learning.

The results indicate that the baseline model obtains an mAP of 0.468 on the HIT-UAV dataset. Following the integration of Task 1 (interference-aware learning) into the baseline, the mAP increases to 0.492, with notable improvements of 6.2 and 4.3 percentage points in metrics $\text{AP}_{\text{person}}$ and $\text{AR}_{\text{person}}$, respectively. This suggests that contrastive learning between the original infrared images and their synthetically degraded views significantly enhances pedestrian

Tab. 1 Datasets used in experiments

Dataset	Train size	Validation set size	Test size	Classification categories
HIT-UAV	2029	290	579	Pedestrian, Bicycle, Car, Other Vehicles
Drone-Vehicle	17990	1469	8980	Car
VT-MOT	16909	3120	–	Pedestrian, Car

Tab. 2 Ablation study results on HIT-UAV and VT-MOT

Dataset	Ablation method	$\text{AP}_{\text{person}}$	$\text{AR}_{\text{person}}$	AP_{car}	AR_{car}	mAP
HIT-UAV	Baseline (Multi-scale Encoder Only)	0.421	0.575	0.689	0.715	0.468
	Baseline + Task 1	0.483	0.618	0.702	0.790	0.492
	Baseline + Task 2	0.434	0.593	0.692	0.721	0.483
	Baseline + Task 1 + Task 2	0.515	0.659	0.755	0.856	**0.513**
VT-MOT	Baseline (Multi-scale Encoder Only)	0.315	0.709	0.426	0.755	0.384
	Baseline + Task 1	0.349	0.813	0.475	0.802	0.423
	Baseline + Task 2	0.324	0.726	0.445	0.758	0.388
	Baseline + Task 1 + Task 2	0.367	0.852	0.508	0.823	**0.437**

target detection. Upon integrating Task 2 (edge-aware learning) into the baseline, the *mAP* increases to 0.483, representing an improvement of 1.5 percentage points. This result indicates that Task 2 contributes positively to detection performance.

In summary, the ViT-based anti-interference feature encoder, whether used for a single task or both tasks together, consistently improves detection performance. The combination of Task 1 and Task 2 yields particularly prominent results, demonstrating that the two tasks are semantically complementary and can jointly optimize feature representation. This further highlights that feature transfer efficiency can be improved via meticulous encoder design, even in the presence of gradient conflicts in multi-task learning.

4.2 Algorithm Comparison Experiment

To validate the algorithm's effectiveness, a series of comparative experiments were conducted, comprising three components: 1) comparison of multi-scale feature fusion methods; 2) comparison of overall detection performance; and 3) comparison of overall anti-noise performance. This study compares several representative multi-scale fusion networks, including FPN [27], GFPN [28], PAN-FPN [29], and the proposed approach. The comparative results are shown in Tab. 3.

Tab. 3 demonstrates that the proposed method outperforms other networks in detection performance on both HIT-UAV and VT-MOT datasets. Regarding inference speed, the FPS of the proposed method is 39, lagging behind the other three methods. However, this overhead remains within an acceptable threshold. The trade-off between sacrificing some speed for a significant improvement in detection performance reflects a good balance between accuracy and efficiency in the proposed method.

4.3 Comparison of Overall Algorithm Performance

This study employs six representative object detection algorithms as baselines, including YOLOv11, deformable DETR [22], conditional DETR [23], RT-DETR [24], PHSI-RTDETR [25], and FECI-RTDETR [26]. As the state-of-the-art CNN-based detection algorithm, YOLOv11 serves as a benchmark for the performance ceiling of traditional CNN approaches. The other five algorithms are DETR-based variants that share the same architectural backbone as the baseline algorithm in this study, ensuring comparative fairness.

Specifically, PHSI-RTDETR and FECI-RTDETR are specialized variants for UAV infrared target detection, and their comparative results offer direct insights into the proposed algorithm's performance in targeted application scenarios. Evaluation metrics for the seven algorithms across three datasets are presented in Tab. 4.

These experimental results demonstrate the proposed algorithm's performance superiority in complex-scene target detection: First, it achieves state-of-the-art performance in challenging high-

Tab. 3 Comparison experiment results of multi-scale fusion methods

Dataset	Method	AP_{person}	AR_{person}	AP_{car}	AR_{car}	mAP	FPS
HIT-UAV	FPN	0.395	0.552	0.654	0.691	0.440	46
	GFPN	0.513	0.629	0.749	0.836	0.505	41
	PAN-FPN	0.438	0.585	0.688	0.735	0.474	43
	Ours	0.515	0.659	0.755	0.856	**0.513**	39
VT-MOT	FPN	0.292	0.686	0.408	0.734	0.370	46
	GFPN	0.341	0.791	0.481	0.794	0.423	41
	PAN-FPN	0.317	0.715	0.449	0.750	0.390	43
	Ours	0.367	0.852	0.508	0.823	**0.437**	39

Tab. 4 Algorithm test results on the original datasets

Dataset	Method	AP_{person}	AR_{person}	AP_{car}	AR_{car}	mAP	FPS
HIT-UAV	YOLOv11	0.508	0.621	0.752	0.835	0.497	51
	Deformable DETR	0.415	0.553	0.569	0.662	0.327	25
	Conditional DETR	0.367	0.514	0.575	0.671	0.315	19
	RT-DETR	0.504	0.617	0.743	0.806	0.494	44
	PHSI-RTDETR	0.512	0.626	0.751	0.825	0.505	35
	FECI-RTDETR	0.521	0.618	0.763	0.836	0.498	32
	Ours	0.515	0.659	0.755	0.856	**0.513**	39
Drone-Vehicle	YOLOv11	–	–	0.672	0.853	0.672	51
	Deformable DETR	–	–	0.569	0.662	0.327	25
	Conditional DETR	–	–	0.575	0.671	0.575	19
	RT-DETR	–	–	0.629	0.954	0.629	44
	PHSI-RTDETR	–	–	0.751	0.825	0.751	35
	FECI-RTDETR	–	–	0.743	0.836	0.743	32
	Ours	–	–	0.755	0.856	**0.755**	39
VT-MOT	YOLOv11	0.318	0.830	0.541	0.888	0.429	51
	Deformable DETR	0.332	0.802	0.352	0.692	0.342	25
	Conditional DETR	0.289	0.732	0.354	0.643	0.321	19
	RT-DETR	0.356	0.835	0.377	0.695	0.367	44
	PHSI-RTDETR	0.361	0.843	0.428	0.752	0.394	35
	FECI-RTDETR	0.354	0.828	0.459	0.803	0.426	32
	Ours	0.367	0.852	0.508	0.823	**0.437**	39

altitude UAV scenarios, particularly in HIT-UAV pedestrian detection. Second, its high average recall rate indicates strong target retrieval capability under complex conditions. Third, it balances speed and accuracy at 39 FPS —while slower than YOLOv11 and RT-DETR, this speed sacrifice yields significant performance improvements, attributable to two design innovations: 1) The anti-interference feature encoder effectively mitigates noise interference in high-altitude infrared imaging, enhancing feature extraction under complex conditions; 2) The multi-scale feature fusion network enables the detection of targets at various scales, ensuring excellent results for both vehicles and pedestrians.

We further visualize the attention distributions of the content-position decoupling mechanism for vehicle and pedestrian targets, as shown in Fig. 7. For vehicle targets, position attention mainly focuses on bounding-box boundaries and vehicle contours. Although different attention heads show slight variations, their high-response regions are generally distributed around the object shape. By comparison, content attention pays more attention to local semantic details, such as windows and wheels, and maintains consistent responses across visually similar vehicles. For pedestrian targets, position attention is concentrated on human contours, while content attention responds more strongly to discriminative body regions such as the head and legs. These results indicate that position queries emphasize spatial boundary modeling, and content queries capture shared semantic attributes across instances, thereby improving the stability and interpretability of detection.

To evaluate algorithm noise resistance, we added Poisson-Gaussian mixed noise, atmospheric turbulence noise, and motion blur noise to the test sets of HIT-UAV, Drone-Vehicle, and VT-MOT, generating noisy dataset variants. Each image was randomly subjected to one of the three noise types with a one-third probability. The comparison results in Tab. 5 use relative

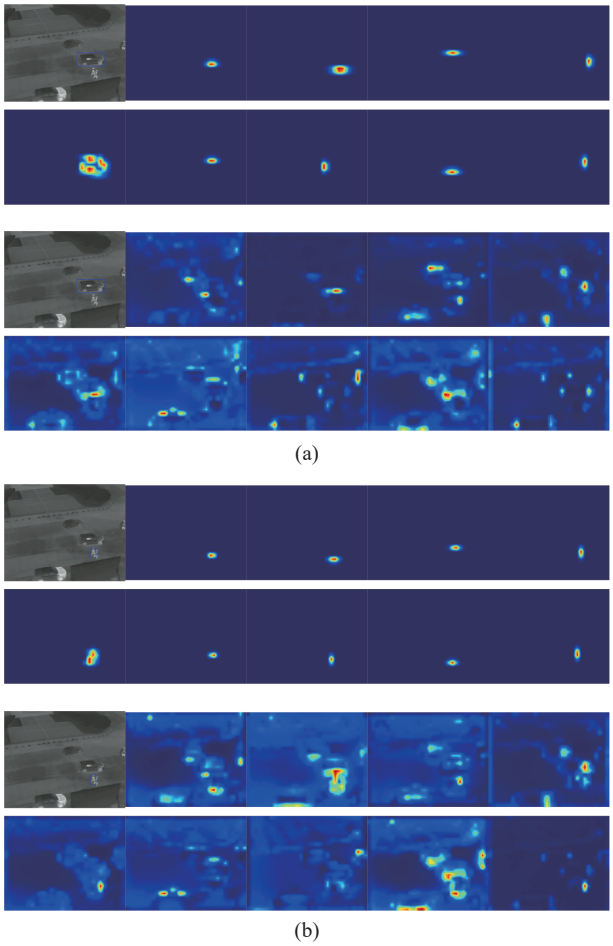


Fig. 7 Attention visualization of the content-position decoupling mechanism: (a) vehicle targets; (b) pedestrian targets (In each group, the first two rows show position attention, and the last two rows show content attention. The image in row 1, column 1 denotes the target image, the image in row 2, column 1 denotes the aggregated attention map, and the remaining columns correspond to individual attention heads)

change as the evaluation metric—i.e., the relative change in each metric of the algorithm under noisy conditions compared to the original noise-free conditions.

These results validate the algorithm's comprehensive advantages in complex scenarios. The results demonstrate that the algorithm maintains stable detection performance across diverse noise interferences, attributed to the pre-trained interference-resistant feature encoder integrated into the multi-scale feature network. This enables the model to extract noise-invariant features and significantly enhances noise robustness.

5 Conclusion

This study systematically investigates the challenges confronting UAV infrared target detection technologies, including complex noise interference and significant target scale variations. These issues impede the real-world application of UAV infrared detection algorithms. To tackle these challenges, we propose an infrared target detection algorithm integrating interference-resistant mechanisms and multi-scale perception. Specifically, to address noise interference in UAV-platform infrared detection, we construct an interference-resistant encoder based on ViT and design two downstream tasks: interference-aware learning and edge-aware learning. Interference-aware learning maximizes mutual information between original and noisy images, enabling the model to learn noise-invariant features and decouple interference in the feature space, thereby enhancing noise robustness. Edge-aware learning, through explicit feature crossing and a binary classification task, helps the model focus on target edges, improving its edge perception for infrared targets. Then, a multi-scale feature fusion network is embedded into the deformable DETR framework. By combining a deformable attention mechanism, we design a multi-scale feature encoder to enhance multi-scale detection capabilities. A gated fusion network is employed to integrate interference-resistant and multi-scale features, while a content-position decoupling mechanism is applied in the decoder to improve decoding efficiency. Visualization, ablation, and comparative experiments demonstrate that the proposed algorithm exhibits strong noise resistance, robust performance, and high adaptability, showing promising potential for real-world infrared target detection scenarios with significant practical value.

Tab. 5 Test results of seven algorithms on the noise-corrupted datasets

Dataset (Noisy)	Method	ΔAP_{person} (%)	ΔAR_{person} (%)	ΔAP_{car} (%)	ΔAR_{car} (%)	ΔmAP (%)
HIT-UAV	YOLOv11	−31.5	−19.4	−28.9	−25.2	−25.0
	Deformable DETR	−13.0	−10.8	−15.7	−14.9	−13.9
	Conditional DETR	−13.9	−12.1	−17.4	−12.4	−14.3
	RT-DETR	−22.1	−18.2	−11.4	−11.1	−9.7
	PHSI-RTDETR	−19.9	−16.0	−11.0	−8.5	−7.4
	FECI-RTDETR	−22.3	−17.6	−11.5	−9.4	−5.0
	Ours	−7.6	−8.6	−8.8	−10.7	−5.7
Drone-Vehicle	YOLOv11	–	–	−7.6	−4.9	−7.6
	Deformable DETR	–	–	−8.6	−10.2	−8.6
	Conditional DETR	–	–	−8.1	−9.2	−8.1
	RT-DETR	–	–	−7.0	−7.9	−7.9
	PHSI-RTDETR	–	–	−6.3	−7.6	−6.3
	FECI-RTDETR	–	–	−6.8	−7.6	−6.8
	Ours	–	–	−5.2	−5.3	−6.2
VT-MOT	YOLOv11	−13.0	−17.9	−17.9	−31.7	−12.4
	Deformable DETR	−18.3	−27.5	−15.5	−17.8	−15.6
	Conditional DETR	−23.5	−37.8	−19.3	−22.7	−18.6
	RT-DETR	−17.0	−31.6	−12.7	−20.1	−14.9
	PHSI-RTDETR	−12.3	−25.1	−8.7	−15.9	−10.1
	FECI-RTDETR	−13.7	−21.7	−6.3	−12.9	−9.2
	Ours	−7.0	−16.7	−6.7	−10.3	−7.8

References:

[1] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440-1448, 2015.

[2] S. Ren, K. He, R. Girshick, and J. Sun,"Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2016.

[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017.

[4] C. Y. Wang, I. H. Yeh, and H. Y. M. Liao, "You only learn one representation: Unified network for multiple tasks," *arXiv preprint*, arXiv: 2105.04206, 2021.

[5] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for realtime object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7464-7475, 2023.

[6] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, and W. Nie, "Yolov6: A single-stage object detection framework for industrial applications," *arXiv preprint*, arXiv: 2209.02976, 2022.

[7] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint*, arXiv: 2107.08430, 2021.

[8] C. Y. Wang, I. H. Yeh, and H. Y. M. Liao, "Yolov9: Learning what you want to learn using programmable gradient information," in *Proceedings of European Conference on Computer Vision*, pp. 1-21, 2024.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proceedings of European Conference on Computer Vision*, pp. 21-37, 2016.

[10] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 734-750, 2018.

[11] S. Pang, X. Liu, S. Mao, H. Jia, and B. Liu, "Advanced-extremenet: Combined with depthwise separable convolution for the detection of steel bars," in *Proceedings of the 2nd International Conference on Artificial Intelligence and Information Systems*, pp. 1-6, 2021.

[12] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6569-6578, 2019.

[13] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "Densebox: Unifying landmark localization with end to end object detection," *arXiv preprint*, arXiv: 1509.04874, 2015.

[14] L. Liu, Y. Zhu, M. R. U. Rahman, P. Zhao, and H. Chen, "Surface defect detection of solar cells based on feature pyramid network and ga-faster-RCNN," in *Proceedings of the 2nd China Symposium on Cognitive Computing and Hybrid Intelligence (CCHI)*, pp. 292-297, 2019.

[15] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9627-9636, 2019.

[16] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi,"Foveabox: Beyound anchor-based object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 7389-7398, 2020.

[17] L. Cao, Q. Wang, Y. Luo, Y. Hou, J. Cao, and W. Zheng,"Yolo-tsl: A lightweight target detection algorithm for uav infrared images based on triplet attention and slim-neck," *Infrared Physics & Technology*, vol. 141, pp. 105487, 2024.

[18] Y. Tang, T. Xu, H. Qin, and J. Li, "Irstd-yolo: An improved yolo framework for infrared small target detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 22, pp. 23-34, 2025.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 12-24, 2017.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint*, arXiv: 2010.11929, 2020.

[21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of European Conference on Computer Vision*, pp. 213-229, 2020.

[22] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint*, arXiv: 2010.04159, 2020.

[23] D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, and J. Wang, "Conditional detr for fast training convergence," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3651-3660, 2021.

[24] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16 965-16 974, 2024.

[25] S. Wang, H. Jiang, Z. Li, J. Yang, X. Ma, J. Chen, and X. Tang,"Phsi-rtdetr: A lightweight infrared small target detection algorithm based on uav aerial photography," *Drones*, vol. 8, no. 6, pp. 240, 2024.

[26] R. Xue, S. Hua, and H. Xu,"Feci-rtdetr a lightweight unmanned aerial vehicle infrared small target detector algorithm based on rt-detr," *IEEE Access*, vol. 13, pp. 9578-9591, 2025.

[27] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117-2125, 2017.

[28] W. Xiao, M. Xu, and Y. Lin, "Global feature pyramid network," *arXiv preprint*, arXiv: 2312.11231, 2023.

[29] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8759-8768, 2018.



Meijing Gao received her Ph.D. degree in Optical Engineering from Beijing Institute of Technology, Beijing, China, in 2008. She is an professor of the School of Integrated Circuits and Electronics at Beijing Institute of Technology since 2021. Her current research interests include image and video processing technology, photoelectric imaging technology and photoelectric detection technology.



Bingzhou Sun was born in 2001. He received his B.S. degree from Tianjin University, Tianjin, China, in 2023. Currently, he is pursuing his M.S. degree at Beijing Institute of Technology, Beijing, China. His current research interests include target detection and recognition.



Sibo Chen was born in 1996. She received the B.S. degree from the Beijing Jiaotong University, Beijing, China, in 2017 and the M.S. degree from the Beijing Jiaotong University, Beijing, China, in 2019, respectively. She is currently pursuing the Ph.D. degree in the Beijing Institute of Technology, Beijing, China. Her current research interests include target detection and recognition.



Huanyu Sun was born in 2002, received the B.S. degree from the School of Computer and Communication Engineering at Northeastern University, Qinhuangdao, China, in 2024. She is currently pursuing the Ph.D. degree at Beijing Institute of Technology, China. Her current research direction focuses on photoelectric imaging target detection and recognition.



Kunda Wang was born in 2000. He received the B.S. degree from the China University of Geosciences, Wuhan, China, in 2022 and the M.S. degree from the Beijing Institute of Technology, Beijing, China., in 2025, respectively. His research interests include target detection and recognition.



Xu Chen was born in 2001. He received the B.S. degree from Hebei University of Technology, Hebei, China, in 2022. Currently, he is pursuing his M.S. degree at Beijing Institute of Technology, Beijing, China. His current research interests include target detection and recognition.



Yufan Du was born in 2001. He received his B.S degree form Beijing University of Technology, Beijing, China, in 2023. Currently, he is pursuing his M.S. degree at Beijing Institute of Technology, Beijing, China. His current research interests include image fusion and super-resolution reconstruction.