

ABFINet: A Lightweight Adaptive Bidirectional Feature Interaction Network for Infrared Target Detection

GuoPing Wang

Xi'an Polytechnic University

Bo Zhang

bozhang@xpu.edu.cn

Xi'an Polytechnic University

KeKe Jing

Xi'an Polytechnic University

WenZhuo Gao

Xi'an Polytechnic University

Research Article

Keywords: Infrared target detection, adaptive bidirectional feature interaction, lightweight network, mbedded deployment

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10566888/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

ABFINet: A Lightweight Adaptive Bidirectional Feature Interaction Network for Infrared Target Detection

GuoPing Wang^{1,2}, Bo Zhang^{1,2}, KeKe Jing¹, WenZhuo Gao^{1,2},
Xin Ren^{1,2}

^{1*}School of Electronics and Information, Xi'an Polytechnic University,
Xi'an, 710600, Shaanxi, China.

²Intelligent Information Processing Laboratory, Xi'an Polytechnic
University, Xian, 710600, Shaanxi, China.

Contributing authors: 2410741294@qq.com; bozhang@xpu.edu.cn;

Abstract

To address the challenges of infrared images under complex backgrounds, including low contrast, strong background interference, and the limitations of embedded deployment, this paper proposes ABFINet, an infrared target detection network based on adaptive bidirectional feature interaction. Specifically, a lightweight multi-scale convolution module is designed to enhance the feature extraction capability of the backbone network while reducing the number of model parameters. Furthermore, an adaptive bidirectional feature interaction fusion module is introduced to establish bidirectional interaction between the spatial boundary information in deep semantic features and the fine-grained target features in shallow layers, thereby improving the feature fusion ability for targets at different scales. Meanwhile, an adaptive fine-grained channel attention mechanism is incorporated to highlight target-related information and suppress background interference. In addition, a dynamic task-aligned detection head is designed to improve target localization and classification performance. Finally, the LAMP pruning strategy is employed to further compress model complexity while preserving detection accuracy as much as possible. Experimental results show that, compared with the baseline model, ABFINet improves mAP@0.5 by 6.2% and 4.9%, and AP_{small} by 7.8% and 8.0% on the FLIR and M3FD datasets, respectively. Moreover, the number of model parameters is reduced by 72.5%, and the model is successfully deployed on an embedded platform, achieving a real-time detection speed of 24 FPS.

Keywords: Infrared target detection, adaptive bidirectional feature interaction, lightweight network, mbedded deployment

1 Introduction

As one of the core technologies in the field of target detection, infrared target detection is widely used in night monitoring, border and coastal security, UAV inspection, automatic driving assistance and other scenes. Different from visible-light imaging systems that depend on external illumination, infrared imaging systems generate images by sensing the infrared radiation emitted from object surfaces. As a result, they are free from the constraints of lighting conditions and are capable of providing clear imaging under all-weather, extreme-weather, and nighttime environments [1]. However, under complex background conditions, infrared images generally suffer from low contrast and weak texture information. In particular, for small-scale and multi-scale targets, discriminative features are highly susceptible to background interference, which often leads to missed detections and false alarms [2]. In addition, embedded deployment places stringent demands on the computational complexity and real-time performance of detection models. Therefore, how to improve the robustness and accuracy of infrared target detection in complex backgrounds while simultaneously achieving lightweight design and real-time inference has become a critical issue in current research.

In recent years, deep learning has developed rapidly in the field of infrared target detection. Single-stage and two-stage models have achieved remarkable success in visible-light detection tasks and have been widely introduced into infrared detection tasks. To address the challenges caused by low contrast and complex background interference, Li et al. [3] employed a target enhancement module to exploit the intrinsic prior knowledge of infrared images, amplifying the grayscale gradients at the boundaries between small targets and the background while suppressing gradient variations within the background. Zhang et al. [4] proposed a dual-layer omnidirectional target enhancement module to alleviate the low-contrast problem. Chen et al. [5] introduced a dual-branch fusion framework to enhance infrared target feature representation in complex scenes. Wu et al. [6] improved infrared image contrast and reduced noise by incorporating an image enhancement module based on improved Gaussian filtering together with a hierarchical feature extraction network. Although these methods have achieved certain improvements in enhancing infrared image contrast, strengthening small-target features, and suppressing partial background interference, they still suffer from relatively heavy module designs, high computational and storage costs, and insufficient adaptive regulation for targets at different scales and under complex backgrounds. As a result, their performance remains limited in terms of real-time embedded deployment and robustness in complex scenarios.

To meet the requirements of embedded and edge devices in terms of real-time performance, power consumption, and storage, researchers have developed various lightweight improvements for infrared target detection. Yan et al. [7] reduced the number of model parameters and improved the detection efficiency of dim small targets in complex backgrounds by refining lightweight modules and network structures.

Zhao et al. [8] proposed a lightweight shape-texture mutual-guided predictor-corrector mechanism, which enhances local texture retention and target structure representation under a lightweight design, thereby alleviating localization deviations caused by weak contrast and missing details. Li et al. [9] proposed a lightweight real-time detection model, termed TA-YOLOv11, which employs an improved bidirectional attention aggregation block to achieve cross-level feature fusion and noise suppression, and has been successfully deployed on the Jetson Orin Nano hardware platform. Although these methods can reduce the complexity of network models to a certain extent, structural compression may weaken feature extraction capability, which in turn increases the rates of missed detections and false alarms in scenarios with strong noise, thermal interference, or complex background textures.

To address the trade-off between detection accuracy and resource efficiency, this paper proposes an infrared target detection network based on adaptive bidirectional feature interaction, termed ABFINet. The main contributions of this study are summarized as follows:

1. Efficient Multi-Scale Convolution Module(EMSC): We propose a novel lightweight multi-scale convolution module, which enhances the modeling capability of contextual information at different scales through grouped multi-scale convolutions. By expanding the effective receptive field while reducing the number of parameters and computational cost, the proposed module improves the backbone network’s multi-scale representation ability for infrared targets.
2. Adaptive Bidirectional Feature Interaction Module(ABFI): An adaptive bidirectional feature interaction fusion module is proposed to enhance cross-scale feature fusion by establishing a bidirectional interaction mechanism between high-resolution and low-resolution features.
3. Dynamic Task-Aligned Detection Head(DATA): We design a Dynamic Task-Aligned Detection Head, which enhances the feature interaction between classification and regression through a task-aligned mechanism, while adopting shared convolutions to reduce redundant parameters and computational overhead in the detection head.
4. Layer-Adaptive Magnitude Pruning(LAMP):The LAMP pruning strategy is employed to compress the model, further reducing the number of parameters and computational cost while preserving detection accuracy, and the compressed model is subsequently deployed and validated on an embedded hardware platform.

2 Related work

2.1 Traditional methods for infrared target detection

Traditional infrared target detection methods mainly focus on manually constructing features based on the differences between targets and backgrounds in terms of grayscale intensity, local contrast, and related characteristics in infrared images, and then achieve target localization and recognition through threshold-based decision strategies and filtering-based background suppression. Ma et al. [10] proposed an improved

weighted local contrast method, in which the ratio information between the target and the local background is employed as an enhancement factor, and a spatial-difference-based weighting strategy is further introduced to highlight the target while suppressing the background. Finally, adaptive thresholding is adopted to accomplish target detection, thereby reducing false alarms and improving overall performance in complex low-altitude backgrounds. Wang et al. [11] performed candidate screening and coarse scale estimation through a three-layer sliding-window scheme, and further introduced an improved dual local contrast measure together with variance-based weighting coefficients in the enhancement stage to suppress clutter. This design improves detection performance while reducing the computational overhead caused by exhaustive multi-scale traversal. Shi et al. [12], from the perspective of filtering-based segmentation, designed a transformed Gaussian filtering kernel to achieve background suppression and normalization, and then combined threshold-based preliminary screening with a clustering strategy to obtain the final target regions, emphasizing a simple processing pipeline and improved detection accuracy under relatively low computational cost. He et al. [13], following a low-rank and sparse modeling paradigm, proposed a learning-based background recovery framework. By integrating a local sparse dictionary with balanced structural texture representation, their method separates background and target through global low-rankness and local sparsity, while further enhancing small-target responses via structural texture modeling, thus adapting to targets under different brightness and scale conditions.

Although the above methods introduce more effective contrast measures, weighting suppression strategies, modeling decomposition schemes, and spatiotemporal search mechanisms at different stages, thereby improving the detection accuracy of traditional algorithms to a certain extent, most of them still rely on prior assumptions regarding target morphology and background statistics. As a result, their performance is highly sensitive to scene variations, and they are prone to false alarms and missed detections in scenarios involving complex background interference, strong noise, dense targets, and occlusion. Consequently, their cross-scene generalization ability and robustness remain limited.

2.2 Deep learning methods for infrared target detection

In recent years, deep-learning-based infrared target detection methods have achieved significant improvements over traditional algorithms in terms of feature representation and cross-scene generalization, thereby alleviating the limitations of conventional methods in complex background suppression and insufficient feature expression. In infrared detection models based on deep learning, feature extraction and feature fusion play crucial roles. Therefore, to enhance the feature extraction and fusion capability of infrared targets, Zhang et al. [14] introduced contrastive vision-language pre-training into infrared small target segmentation, where textual semantic prompts were used to improve target discriminability and generalization in complex environments. Yue et al. [15] proposed the IR-DETR framework and designed an efficient backbone network to achieve effective representation and detection of infrared small targets in a lightweight manner. Wang et al. [16] proposed a lightweight and efficient backbone network for UAV infrared scenarios, which improves the robustness and accuracy of

infrared target detection by suppressing background interference and enhancing target responses through multi-scale attention. Jiang et al. [17] developed a lightweight improved network, which enhances small-target representation through a lightweight backbone and fusion enhancement strategy, while further improving training and localization performance by incorporating an improved regression loss. Xu et al. [18] proposed a multi-domain feature fusion infrared detector and introduced a feature injection module in the fusion stage to strengthen the interaction between channel-wise and spatial information. Zhang et al. [19] enhanced feature representation through attention-guided contextual modeling and pyramid-based multi-scale context fusion, and further preserved small-target information by employing asymmetric fusion during the upsampling stage. Chung et al. [20] introduced residual attention blocks into the feature fusion network to strengthen feature interaction, while simultaneously exploiting information from different layers through deep supervision. Gu et al. [21] achieved the fusion of shallow spatial features and deep semantic features through a coarse-to-fine spatial and channel reconstruction convolution strategy. Khan et al. [22] proposed an innovative hexagonal attention framework that integrates multiple attention mechanisms for feature refinement. Yuan et al. [23] proposed a full-scale dynamic pyramid fusion network to enhance target edge information. This method adopts a two-stage strategy consisting of initial coarse extraction, dynamic refinement, and feature enhancement to improve detection performance.

Although substantial progress has been made in infrared target detection, existing methods still suffer from insufficient detection performance and limited robustness in both backbone design and feature fusion. In addition, some multi-module fusion schemes further increase model complexity, which restricts their application in resource-constrained scenarios.

3 Proposed methodology

The lightweight infrared target detection network ABFINet mainly consists of a backbone network, a neck network, and a detection head, as illustrated in Fig. 1. To enhance the feature extraction capability of the backbone network while reducing the number of network parameters, a lightweight multi-scale convolution module, termed EMSC, is introduced. To improve the fusion capability of multi-level feature maps, an adaptive bidirectional feature interaction module, termed ABFI, is designed to effectively enhance the feature representation ability of the network. To further strengthen the classification and localization performance for small targets, a dynamic task-aligned detection head, termed DTAD, is proposed, thereby improving the overall detection performance of the model.

3.1 Efficient multi-scale convolution module

In infrared target detection, targets usually exhibit smooth edge transitions and weak texture contrast, resulting in insufficiently salient feature responses. Meanwhile, small targets are prone to detail loss and feature dilution during deep downsampling, whereas targets with large scale variations rely more heavily on multi-scale contextual information. These factors collectively lead to insufficient feature representation in network

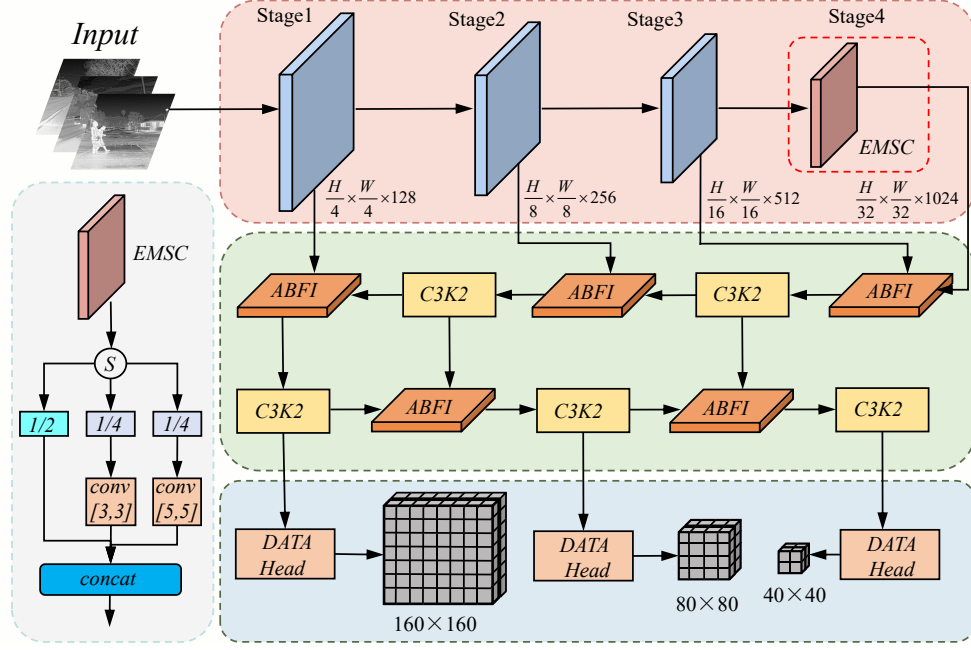


Fig. 1 ABFINet structure diagram

structures. To address this issue, this paper aims to enhance the multi-scale modeling capability of the backbone network with lower computational overhead while maintaining detection accuracy. Inspired by the efficient feature generation mechanism of GhostNet [24], a lightweight multi-scale convolution module, termed EMSC, is proposed, and its architecture is illustrated in Fig. 2.

Let the input feature map be denoted by $X \in \mathbb{R}^{B \times C \times H \times W}$, where B , C , H , and W denote the batch size, the number of channels, and the height and width of the feature map, respectively. First, the input feature map X is divided into two parts along the channel dimension:

$$[X_1, X_2] = \text{Split}(X), \quad (1)$$

where $X_1 \in \mathbb{R}^{B \times \frac{C}{2} \times H \times W}$ and $X_2 \in \mathbb{R}^{B \times \frac{C}{2} \times H \times W}$. Among them, X_1 is directly preserved and delivered to the output fusion stage through an identity mapping without convolutional computation. In this way, only half of the channels participate in the subsequent convolutional operations, which reduces the computational cost while retaining weak infrared texture information. Subsequently, a channel-wise splitting operation is further applied to the learnable branch X_2 :

$$[X_i, X_j] = \text{Split}_c(X_2), \quad (2)$$

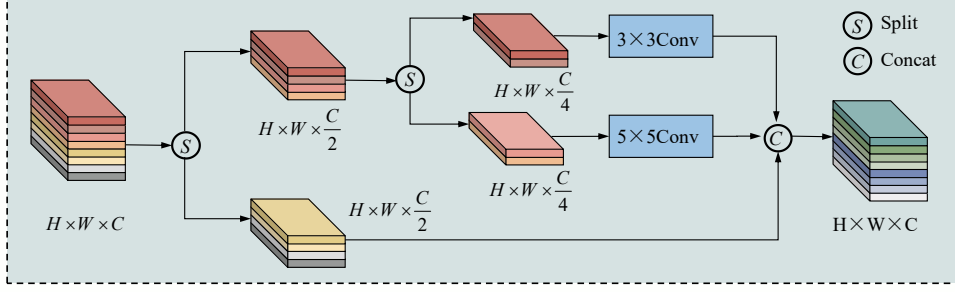


Fig. 2 EMSC model structure

where X_i and X_j denote two sub-branches, with $X_i \in \mathbb{R}^{B \times \frac{C}{4} \times H \times W}$ and $X_j \in \mathbb{R}^{B \times \frac{C}{4} \times H \times W}$.

Finally, the two sub-branches are processed by independent convolutional layers with different kernel sizes. The corresponding computations can be formulated as

$$Y_i = \text{Conv}_{3 \times 3}(X_i; W_i), \quad Y_j = \text{Conv}_{5 \times 5}(X_j; W_j), \quad (3)$$

where $\text{Conv}_{k \times k}(\cdot)$ denotes a two-dimensional convolution operation, and W_i and W_j are the corresponding convolution kernel parameters. The 3×3 branch is used to capture the local fine-grained features of infrared targets, while the 5×5 branch is employed to extract more global feature information of infrared targets.

Finally, the original identity branch and the two multi-scale convolution branches are concatenated along the channel dimension. This design not only reduces the number of model parameters and computational cost, but also more effectively enhances the backbone network's feature representation capability for infrared targets in scenarios with weak texture, blurred boundaries, and significant scale variations.

3.2 Adaptive bidirectional feature interaction module

In infrared feature fusion, shallow features contain abundant detail and boundary information and provide relatively accurate localization, but their semantic representation is weak. In contrast, deep features encode stronger semantic information, yet they tend to lose target contours and fine details due to the reduction in spatial resolution. Traditional pyramid structures generally adopt a unidirectional top-down fusion strategy and directly perform cross-level addition, which easily leads to mismatches between semantic and detail information, introducing redundant and inconsistent features. This issue is particularly pronounced in infrared scenarios characterized by weak texture.

To address this problem, this paper proposes an Adaptive Bidirectional Feature Interaction Module, whose architecture is illustrated in Fig. 3. Through a bidirectional fusion mechanism between high-resolution and low-resolution features, the proposed module enables semantic information and detailed information to be sufficiently circulated and complemented across multiple scales, thereby improving the consistency

and effectiveness of cross-scale features and enhancing the fused representation capability for infrared targets.

First, the shallow and deep features are fed into two Re-calibration Attention Unit

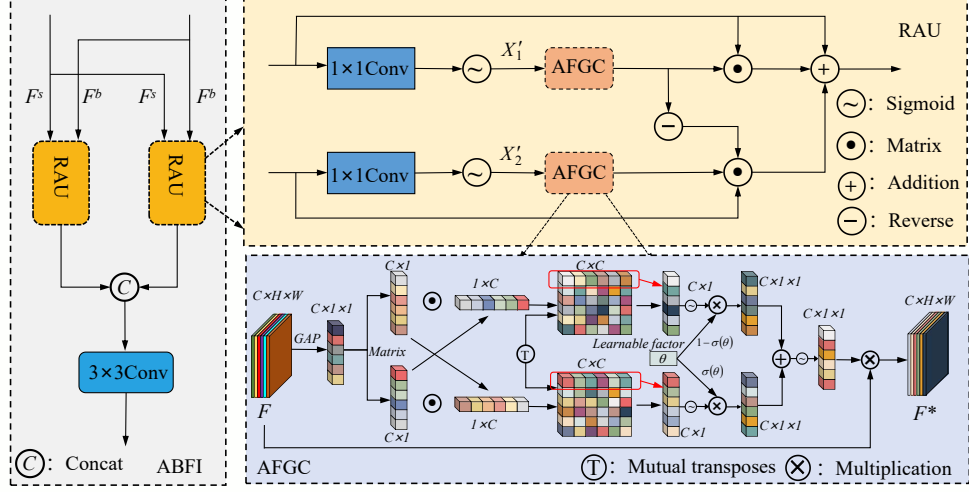


Fig. 3 ABFI model structure

(RAU) modules in a complementary cross-input manner. The outputs of the two RAU branches are then concatenated and passed through a 3×3 convolution, followed by batch normalization and ReLU activation, to achieve effective fusion and enhancement of cross-level features. The RAU module can be expressed as

$$X'_1 = W_\theta(X_1), \quad X'_2 = W_\phi(X_2), \quad (4)$$

$$\text{RAU}(X_1, X_2) = X'_1 \odot X_1 + X'_2 \odot X_2 \odot \Theta(X'_1) + X_1, \quad (5)$$

where X_1 and X_2 are the two input features, respectively; $W_\theta(\cdot)$ and $W_\phi(\cdot)$ denote the learnable transformations for the corresponding inputs; X'_1 and X'_2 denote the transformed feature maps; and $\odot$ represents element-wise multiplication. The symbol $\Theta(\cdot)$ denotes the inverse operation, which is implemented by subtracting the feature X_i . In this work, a 1×1 convolution is used as the linear mapping operation. Based on the above design, the AFBIM can be expressed as

$$Y = \text{Conv}_{3 \times 3} \left(\text{Concat}(\text{RAU}(F^s, F^b), \text{RAU}(F^b, F^s)) \right), \quad (6)$$

where $\text{Conv}_{3 \times 3}(\cdot)$ denotes a 3×3 convolution followed by batch normalization and ReLU activation. F^s and F^b denote the shallow detailed features and deep semantic features extracted from the backbone network, respectively.

Infrared images usually suffer from strong noise and low contrast, and cross-level

feature interaction tends to introduce interference while weakening target responses. To address this issue, an Adaptive Fine-Grained Channel Attention (AFGC) mechanism is introduced before the gating operation [25]. By constructing channel correlation matrices via cross-correlation, the proposed mechanism captures inter-channel dependencies and integrates multi-granularity associations between global semantic information and local detailed information, thereby achieving adaptive channel recalibration. Specifically, the AFGC module first extracts global and local channel features and then obtains the corresponding channel correlation weights through cross-correlation operations. Subsequently, a learnable fusion factor is introduced to adaptively combine the global and local channel weights. The computation process is formulated as

$$W = \sigma(\sigma(\theta) \odot \sigma(U_{gc}^w) + (1 - \sigma(\theta)) \odot \sigma(U_{lc}^w)), \quad (7)$$

$$F^* = W \otimes F, \quad (8)$$

where U_{gc}^w and U_{lc}^w denote the global and local channel weights, respectively; θ denotes the learnable fusion factor; and F and F^* represent the input and output feature maps, respectively. With this design, the AFGC module reduces redundant correlation computation between global and local information, enhances target-related discriminative representations, and suppresses infrared noise interference, thus providing more reliable weight guidance for subsequent cross-level feature interaction.

3.3 Dynamic task-aligned detection head

In complex infrared scenarios, constrained by low contrast and small-scale imaging characteristics, the regression learning of the localization branch is more likely to become unstable. In addition, conventional detection heads usually adopt independent classification and localization branches, which leads to insufficient interaction between the two tasks and increases the rates of false alarms and missed detections for infrared small targets in complex backgrounds. To address these issues, this paper designs a Dynamic Task-Aligned Detection Head (DATA). Based on a shared feature extractor, the proposed detection head learns task interaction information from multi-level convolutional features and constructs a joint feature representation. Its overall architecture is illustrated in Fig. 4.

Specifically, DATA first feeds multi-scale feature maps from different levels into two groups of shared 3×3 convolutions for feature enhancement and normalization, thereby reducing the computational complexity of the model. Then, inspired by TOOD [26], the concatenated multi-scale feature F is fed into the Task-Aligned Dynamic (TAD) module, which explicitly captures the interaction between classification and localization tasks and outputs task-interactive features. The task interaction process is formulated as

$$\omega = \sigma(F_{c2}(\delta(F_{c1}(x^{inter})))), \quad (9)$$

where ω denotes the learned attention weight, F_{c1} and F_{c2} are two fully connected layers, $\delta(\cdot)$ denotes the activation function, and x^{inter} is obtained by applying average pooling to X^{inter} , with X^{inter} representing the concatenated feature.

In the localization branch, through interactive feature learning, tensors corresponding to the predicted offset and mask information are first generated. These tensors

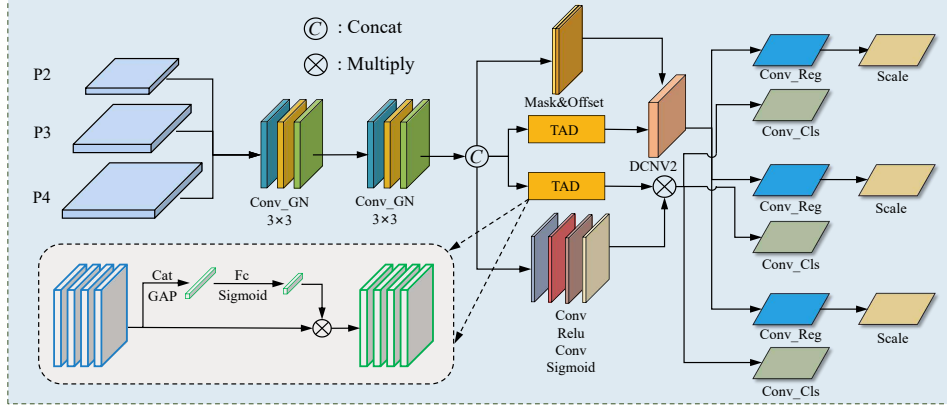


Fig. 4 DATA head structure

are then fed into DCNV2, where the learned offset and mask information are used to perform spatial transformation and feature fusion on the input features, thereby producing aligned features that better conform to the target geometry and effective regions. In the classification branch, attention weights are generated through convolution and normalization operations, and are then multiplied element-wise with the features, enabling dynamic selection of channel-wise and spatial information and yielding feature representations that are more beneficial for classification. Finally, the aligned and filtered features are fed into each detection head, respectively. Meanwhile, to address the inconsistency of target scales across different detection layers, a learnable scale factor, namely Scale, is introduced to calibrate different scale levels.

3.4 Layer-adaptive magnitude pruning

To reduce model size and computational overhead while preserving detection accuracy, this paper adopts the layer-adaptive magnitude-based pruning strategy, namely LAMP, to perform structured sparsification and compression. The core idea of LAMP [27] is to operate within a global magnitude pruning framework while introducing layer-wise adaptive reweighting of weight magnitudes within each layer, thereby automatically allocating pruning ratios across layers and avoiding over-pruning or redundant retention caused by fixed thresholds or manually specified layer-wise pruning ratios.

Specifically, during LAMP pruning, the weights of each prunable layer are first measured according to their magnitudes. For the weight tensor of the l -th layer, denoted by $W^{(l)}$, it is flattened into a vector and sorted in ascending order based on absolute value, resulting in the ordered sequence $W^{(l)}[u]$. A LAMP score is then computed for each weight within the layer. After collecting the scores from all layers, a global ranking is performed, and the weights with lower scores are pruned under a given overall sparsity constraint. The pruned model is subsequently fine-tuned to recover its

performance. The LAMP score is formulated as

$$\text{score}(u; W) = \frac{(W[u])^2}{\sum_{v \geq u} (W[v])^2}, \quad (10)$$

where $W[u]$ denotes the u -th weight in the sequence sorted according to $|W|$, and the denominator represents the remaining weight energy from position u onward within the same layer. Therefore, the score reflects the relative importance of the current weight with respect to the remaining weight energy, which enables adaptive sparsity allocation across layers and reduces the performance loss caused by over-pruning.

After the LAMP scores are obtained, all model weights are globally ranked according to their scores, and the smallest portion is removed until the target sparsity is reached. Owing to the layer-wise normalization mechanism, the sparsity ratio of each layer can be allocated automatically, thus avoiding the need to manually specify pruning ratios for different layers.

4 Experimental analysis

4.1 Dataset

This study uses two publicly available infrared image datasets, i.e., the FLIR dataset [28] and the M3FD dataset [29], for training and evaluation. The FLIR dataset was collected on streets and highways in Santa Barbara, California, USA, using a roof-mounted thermal imaging camera. The M3FD dataset, released by Dalian University of Technology for infrared object detection, contains 4,200 infrared images. The training and test splits of the FLIR and M3FD datasets are presented in Table 1.

To further assess the performance of the proposed method on targets of different scales, the target size definitions in the COCO dataset [30] are adopted. Specifically, targets smaller than 32×32 pixels are regarded as small targets, those between 32×32 and 96×96 pixels are regarded as medium targets, and those larger than 96×96 pixels are regarded as large targets.

Table 1 Statistics of the FLIR and M3FD datasets

Name	Categories	Training set	Validation set	Test set	Image size
FLIR	3	7156	1022	2045	640×51
M3FD	6	2940	420	840	1024×768

4.2 Experimental environment

To ensure the comparability and reliability of the experimental results, all experiments were conducted on a platform equipped with n Intel(R) Core(R) i7-12700 CPU and an NVIDIA GeForce RTX 3060 GPU with 12 GB memory. The operating system was

Ubuntu 20.04.5 LTS, and PyCharm was used as the integrated development environment. The software environment included CUDA 12.2, PyTorch 2.2, and Python 3.10. The detailed hyperparameter settings are presented in Table 2.

Table 2 Hyperparameter settings

Dataset	Weight decay	Optimizer	Momentum	Learning rate	Batch size	Epochs
FLIR	0.0005	SGD	0.937	0.01	16	200
M3FD	0.0005	SGD	0.937	0.01	16	300

4.3 Evaluation metrics

To comprehensively evaluate the detection performance and computational efficiency of the proposed model, this paper adopts Precision (P), Recall (R), mean Average Precision (mAP), the number of parameters, GFLOPs, and FPS as evaluation metrics.

Precision measures the proportion of true positive targets among all targets predicted as positive by the model. where TP denotes the number of true positive targets, and FP denotes the number of false positive targets, and is defined as follows:

$$P = \frac{TP}{TP + FP} \quad (11)$$

Recall is used to measure the proportion of true targets successfully detected by the model, where FN denotes the number of false negative targets, and is defined as follows:

$$R = \frac{TP}{TP + FN} \quad (12)$$

mAP is one of the most commonly used comprehensive evaluation metrics in object detection, and is employed to measure the overall detection performance of a model at different recall levels. where N denotes the number of target categories. mAP@0.5 represents the mean Average Precision when the IoU threshold is set to 0.5, It is defined as follows:

$$mAP = \frac{\sum_{i=1}^N \int_0^1 P(R) dR}{N} \quad (13)$$

FPS is used to measure the processing speed of the model during the inference stage, indicating the number of image frames that can be processed per unit time. where N denotes the number of images processed by the model during testing, and T denotes the corresponding total inference time. It is defined as follows:

$$FPS = \frac{N}{T} \quad (14)$$

GFLOPs is used to measure the computational complexity of the model, representing the number of billions of floating-point operations required in a single forward inference pass.

Parameters is used to measure the model size, representing the total number of learnable parameters contained in the model.

4.4 Comparison with other algorithms

To further verify the effectiveness of ABFINet for infrared target detection, comparative experiments are conducted on the FLIR and M3FD infrared multimodal datasets. The comparison models include general-purpose object detectors such as YOLOv11n, together with advanced detectors such as MMI-Det that are specifically designed for infrared or multimodal scenarios. Through these comparisons, the proposed method can be more objectively evaluated in terms of detection accuracy, lightweight design, and real-time performance. The quantitative comparison results are shown in Table 3.

Table 3 presents the performance comparison results between the proposed method and several mainstream detection models on the FLIR and M3FD datasets. Overall,

Table 3 Performance comparison of different models on the FLIR and M3FD datasets

Dataset	Method	Input	mAP@0.5	Params (M)	FLOPs (G)	FPS
FLIR	YOLOv8n[31]	IR	77.6	3.0	8.1	278
	RTDETR-r18[32]	IR	72.3	32.0	103.4	50
	YOLOv9-tiny[33]	IR	75.1	2.6	10.7	185
	YOLOv10n[34]	IR	76.6	2.7	8.2	357
	YOLOv11n[35]	IR	76.4	2.58	6.3	330
	YOLOv13n[36]	IR	75.4	2.45	6.2	217
	MMI-Det[37]	RGB+IR	79.8	207.6	229.2	–
	ICAFusion[38]	RGB+IR	79.2	120.2	192.7	41.4
	LRAF-Net[39]	RGB+IR	80.5	18.8	40.5	46.8
	ABFINet (Ours)	IR	82.3	0.71	6.6	148
M3FD	YOLOv8n[31]	IR	73.2	3.0	8.1	278
	RTDETR-r18[32]	IR	73.2	32.0	103.4	50
	YOLOv9-tiny[33]	IR	69.5	2.6	10.7	185
	YOLOv10n[34]	IR	72.5	2.7	8.2	357
	YOLOv11n[35]	IR	76.7	2.58	6.3	327
	YOLOv13n[36]	IR	72.5	2.45	6.2	227
	MMI-Det[37]	RGB+IR	76.6	207.6	229.2	–
	ICAFusion[38]	RGB+IR	71.9	120.2	192.7	–
	DAMSDet[40]	RGB+IR	80.2	–	–	–
	ABFINet (Ours)	IR	80.9	0.71	6.6	139

ABFINet achieves superior detection results on both datasets, while maintaining relatively low parameter count and computational cost, demonstrating a favorable balance between accuracy and efficiency.

On the FLIR dataset, ABFINet improves $mAP@0.5$ by 4.7% and 10% compared with YOLOv8n and RTDETR-r18, respectively. These results indicate that the proposed model has stronger feature representation capability for infrared targets under complex backgrounds, and can better balance detection accuracy and recall performance.

On the M3FD dataset, ABFINet improves $mAP@0.5$ by 7.7% compared with YOLOv8n. The results further demonstrate that the proposed model maintains strong robustness and adaptability in complex multimodal scenarios, thereby effectively improving infrared target detection performance.

From the perspectives of model complexity and inference speed, ABFINet contains only 0.71M parameters and requires 6.6 GFLOPs, which are significantly lower than those of advanced multimodal detection models such as MMI-Det, and also lower than those of most lightweight detection models. In terms of inference speed, ABFINet achieves FPS values of 148 and 136 on the FLIR and M3FD datasets, respectively, indicating that it maintains favorable real-time performance while achieving higher detection accuracy.

4.5 Ablation experiment

To further evaluate the detection performance and real-time capability of ABFINet, ablation experiments were conducted on the FLIR and M3FD datasets for each proposed module. $mAP@0.5$, average precision for small objects (AP_s), and the number of parameters were adopted as evaluation metrics to assess the effectiveness of each improved module, where “ $\sqrt{}$ ” indicates that the corresponding module is included. The ablation results are presented in Table 4. According to the ablation results in Table 4, first, after introducing the EMSC module on the basis of the baseline model, the number of parameters and computational cost are reduced by 0.1M and 0.1G, respectively, while $mAP@0.5$ and AP_s are improved by 0.3% and 0.4%, respectively. These results indicate that the EMSC module enhances the backbone network’s feature extraction capability for multi-scale infrared targets while reducing model complexity.

Second, after incorporating the ABFI module, the detection performance of the model is improved most significantly. On the FLIR dataset, $mAP@0.5$ and AP_s are increased by 5.1% and 6.9%, respectively; on the M3FD dataset, $mAP@0.5$ and AP_s are improved by 3.2% and 5.8%, respectively. These results demonstrate that the ABFI module can effectively strengthen cross-level feature interaction and fusion, thereby improving the model’s representation capability for small targets and weak-texture targets.

After introducing the DATA module, the number of model parameters is further reduced, while AP_s on the FLIR and M3FD datasets increases to 36.5% and 26.0%, respectively. This indicates that the DATA module improves small-target detection performance while simultaneously reducing model complexity.

Finally, after applying the LAMP pruning strategy to the complete model, the number of parameters is further compressed from 1.69M to 0.71M, and the FLOPs

are reduced from 8.7G to 6.6G. Meanwhile, the FPS on the FLIR and M3FD datasets increases to 148 and 139, respectively. At the same time, $mAP@0.5$ decreases only slightly, from 82.6% to 82.3% and from 81.3% to 80.9%, respectively. These results indicate that the LAMP pruning strategy can effectively compress model size and computational cost while maintaining high detection accuracy, thereby further improving the deployability of the model.

Table 4 Ablation experiment of ABFINet on the FLIR and M3FD datasets

Dataset	EMSC	ABFI	DATA	LAMP	P	R	mAP_{50}	AP_s	Param. (M)	FLOPs (G)	FPS
FLIR					81.9	67.3	76.4	34.2	2.58	6.3	347
	✓				82.7	67.2	76.8	34.5	2.48	6.2	352
		✓			84.0	71.4	81.5	41.1	1.99	9.6	148
			✓		84.4	68.6	78.4	36.5	2.20	7.9	296
	✓	✓			84.8	71.2	81.8	41.3	1.94	9.6	156
	✓		✓		83.3	69.3	78.6	36.8	2.15	7.8	285
		✓	✓		84.1	73.2	82.3	42.1	1.74	8.7	137
	✓	✓	✓		84.3	73.6	82.6	42.4	1.69	8.7	142
✓	✓	✓	✓	83.8	73.4	82.3	41.8	0.71	6.6	148	
M3FD					83.7	69.3	76.7	24.5	2.58	6.3	323
	✓				84.3	70.2	77.2	24.9	2.48	6.2	326
		✓			84.4	73.5	79.9	30.3	1.99	9.6	137
			✓		83.6	71.4	77.3	26.0	2.20	7.9	269
	✓	✓			84.9	72.8	80.1	30.4	1.94	9.6	142
	✓		✓		83.6	71.8	77.6	26.2	2.15	7.8	267
		✓	✓		82.8	74.8	80.4	31.6	1.74	8.7	128
	✓	✓	✓		85.9	74.1	81.3	33.6	1.69	8.7	134
	✓	✓	✓	✓	85.7	73.6	80.9	32.5	0.71	6.6	139

4.5.1 Effectiveness study of lightweight multi-Scale convolution

To investigate the influence of different convolution kernel sizes on the performance of the EMSC module for infrared target detection, experiments were conducted using different combinations of convolution kernel sizes. The corresponding results are presented in Table 5.

According to the experimental results in Table 5, when the convolution kernel sizes are set to 3 and 5, respectively, the EMSC module achieves the best feature representation capability for infrared target detection, leading to enhanced perception of infrared target features. Meanwhile, detection accuracy is improved while the number of model parameters and computational cost are reduced.

To further verify the effectiveness of the EMSC module, the receptive fields before and after the improvement are compared in this paper. To ensure the fairness of the experiment, feature maps from the same layer before and after the improvement are selected for comparison, as shown in Fig. 5. It can be observed that the improved receptive field is more concentrated within the medium-response intensity range, while preserving more contextual information over a wider region. This is beneficial for distinguishing targets from complex backgrounds and improving feature representation capability.

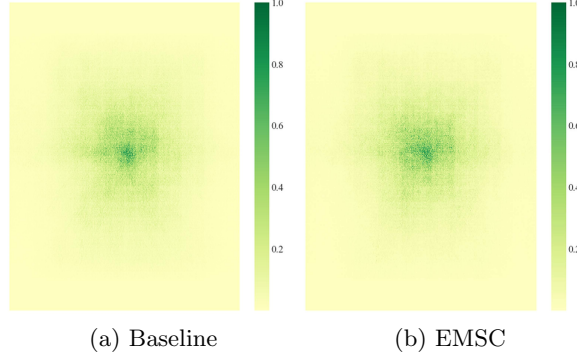


Fig. 5 Comparison of receptive field visualization before and after improvement.

4.5.2 Effectiveness study of adaptive bidirectional feature interaction fusion

To validate the multi-scale fusion capability of the proposed ABFI module in the feature fusion network, this paper compares it with several mainstream feature fusion networks, including BiFPN [41], Gold-YOLO [42], AFPN [43], and RepGFPN [44]. The corresponding experimental results are presented in Table 6.

According to the experimental results in Table 6, after introducing the ABFI module, the proposed fusion network improves AP_s by 6.2%, 17.5%, 7.2%, and 6.4%

Table 5
Comparison of detection performance with different convolution kernel sizes on the FLIR dataset

Dataset	Model	P	R	$mAP@0.5$	Params (M)	FLOPs (G)
FLIR	Baseline	81.9	67.2	76.4	2.58	6.3
	[1,3]	82.7	67.2	76.3	2.46	6.2
	[1,5]	82.3	66.5	76.4	2.48	6.2
	[1,7]	81.7	66.8	75.9	2.50	6.3
	[3,5]	83.4	67.3	76.9	2.48	6.2
	[3,7]	82.9	66.3	76.4	2.51	6.3
	[5,7]	81.8	66.9	76.4	2.52	6.3

compared with BiFPN, Gold-YOLO, AFPN, and RepGFPN, respectively. Meanwhile, AP_m also achieves the best result. These results demonstrate that the ABFI module can effectively promote the interaction and fusion of infrared target information at different scales, thereby enhancing the model’s multi-scale feature fusion capability for infrared targets.

Table 6 Performance comparison of different fusion networks on the FLIR dataset

Dataset	Model	AP_S	AP_M	AP_L	$mAP@0.5$	Params(M)	FLOPs(G)
FLIR	BiFPN	34.9	65.0	79.7	75.4	1.92	6.3
	Goldyolo	23.6	55.2	66.7	77.2	5.90	9.2
	AFPN	33.9	65.7	80.4	75.8	2.65	8.8
	RepGFPN	34.7	66.1	82.8	76.5	3.66	8.2
	ABFI (ours)	41.1	66.5	81.2	81.5	1.99	9.6

To further verify the effectiveness of the proposed AFGC mechanism, a visualization analysis of feature responses before and after introducing the module is conducted. As shown in Fig. 6, compared with the case without AFGC, the high-response regions in the target area become more continuous and concentrated after introducing AFGC, and the feature representation of the target body is significantly enhanced. Meanwhile, some irrelevant responses in the background regions are suppressed. These results indicate that AFGC can improve target feature representation to a certain extent and enhance the discriminability between the target and the background.

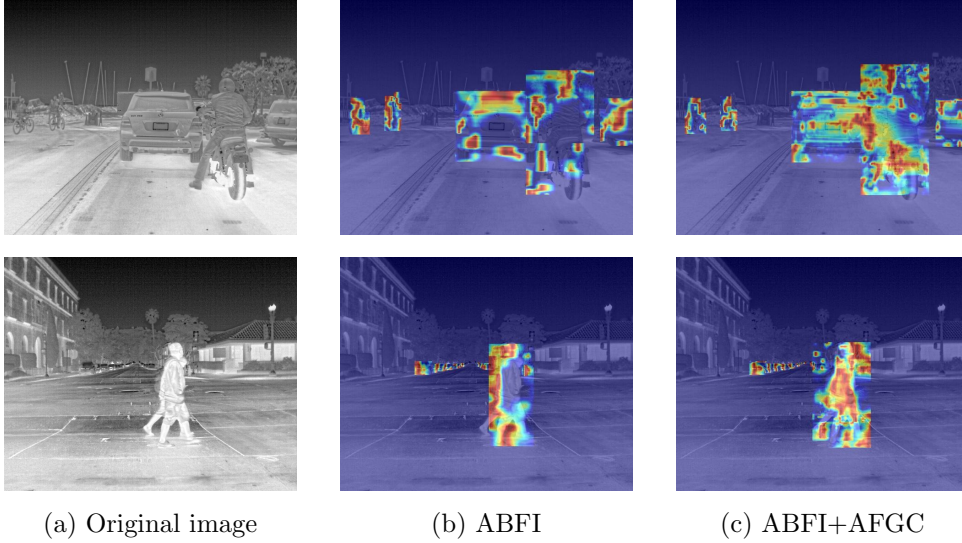
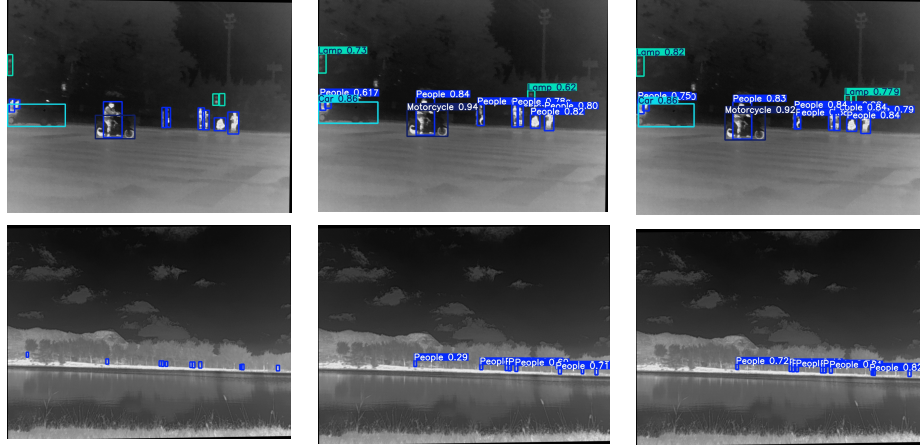


Fig. 6 Heat map comparison chart

4.5.3 Effectiveness study of the dynamic task-aligned detection head

To further verify the effectiveness of the proposed DATA detection head, qualitative detection comparisons are conducted on selected scenes from the FLIR and M3FD datasets. The selected examples include dense-target and small-target scenarios, and the corresponding results are shown in Fig. 7. As can be observed, the DATA detection head improves infrared target detection accuracy in low-contrast and target-dense scenes by leveraging multi-level task feature interaction. In addition, it enhances boundary perception and localization regression for dim small targets, thereby effectively reducing missed detections in small-target scenarios.



(a) Ground Truth

(b) Baseline

(c) DATA

Fig. 7 Comparison of detection results of before and after introducing DATA

4.6 Analysis of visualization

To more intuitively analyze the detection performance of different models under complex infrared scenarios, this paper conducts qualitative comparison experiments on infrared images selected from different scenes in the FLIR and M3FD datasets, including target occlusion scenarios, low-contrast scenarios, target-dense scenarios, and small-target scenarios. The corresponding visual comparison results are shown in Fig. 7.

Fig. 8(a) compares the detection results of different algorithms in target occlusion scenarios. It can be observed that, under strong interference from complex background information, ABFINet is able to detect the targets more accurately, while also achieving the best confidence performance. Fig. 8(b) presents the comparative detection results of different algorithms in low-contrast scenarios. The results show that, by enhancing the feature extraction capability of the backbone network, ABFINet can achieve stable and accurate detection performance in low-contrast scenes, while

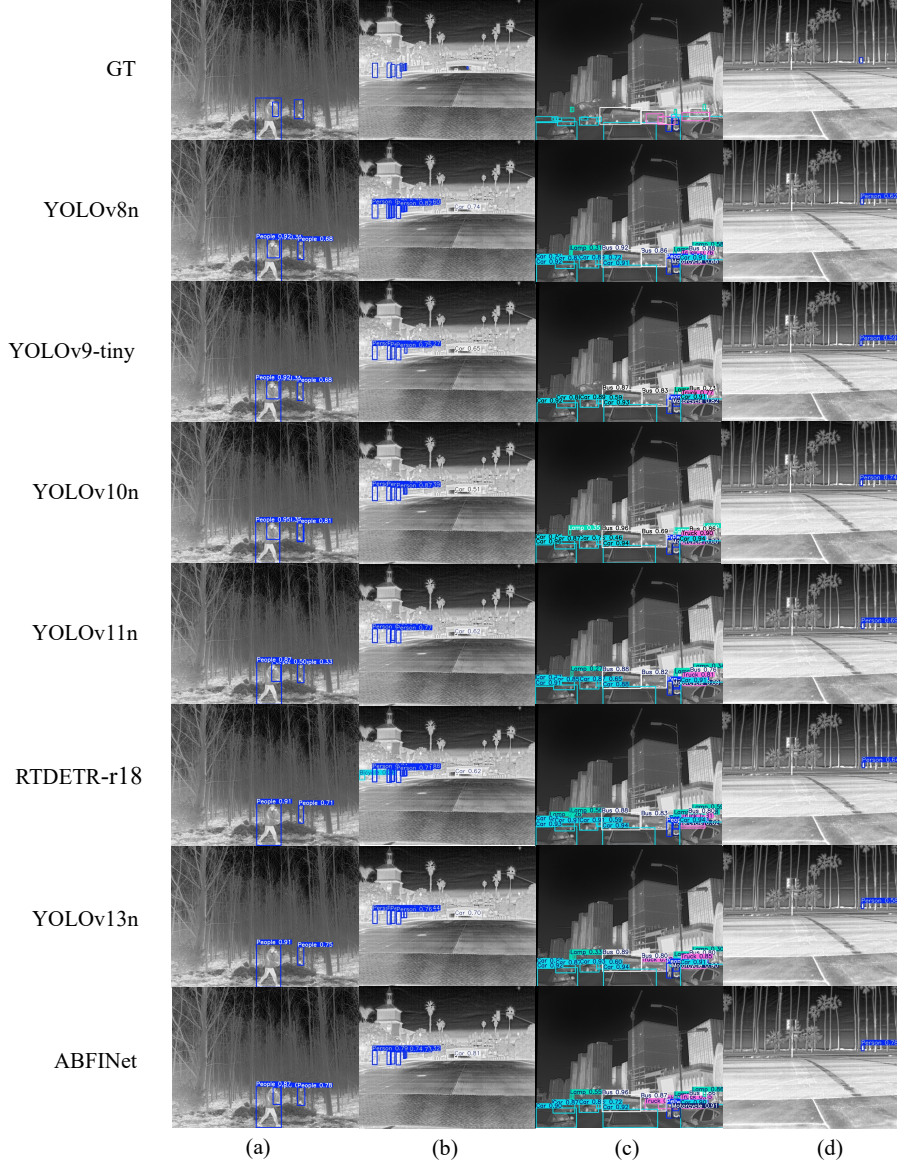


Fig. 8 Comparison of detection results of the FLIR and M3FD dataset under different models

also obtaining the highest confidence scores. Fig. 8(c) illustrates the detection comparison results in target-dense scenarios. Benefiting from the adaptive bidirectional feature interaction mechanism, ABFINet can effectively fuse multi-scale target features, thereby reducing false detections in dense scenes. Fig. 8(d) compares the detection results of different algorithms in small-target scenarios. It can be clearly seen that, by adopting the dynamic task-aligned detection head, ABFINet achieves more

accurate localization and classification for small targets, with the highest confidence scores, outperforming the other detection models.

In summary, compared with other detection models, ABFINet exhibits more outstanding performance in infrared image detection tasks and demonstrates clear advantages in overall detection performance.

4.7 Embedded deployment

To verify the deployment feasibility and practical application effectiveness of the proposed infrared target detection algorithm on an embedded platform, an infrared intelligent target detection system based on the RK3588 platform is designed. The RK3588 is a high-performance development board developed by Rockchip, featuring an architecture with $4 \times$ Cortex-A76 and $4 \times$ Cortex-A55 cores, and providing up to 6 TOPS of computing power. Owing to its suitability for deploying lightweight detection models, the RK3588 is selected as the hardware platform in this study.

To fully exploit the NPU acceleration capability of the RK3588 platform, the proposed detection model needs to be converted into the RKNN format supported by NPU inference, and model deployment and inference optimization are then carried out on this basis. Specifically, the trained weight file is first converted into an open neural network model in ONNX format, followed by model quantization. Subsequently, the converted ONNX model is transformed into an RKNN model supported by the RK3588 embedded platform. Finally, inference is performed by invoking the on-board NPU API to obtain the detection results. The overall deployment process is illustrated in Fig. 9.

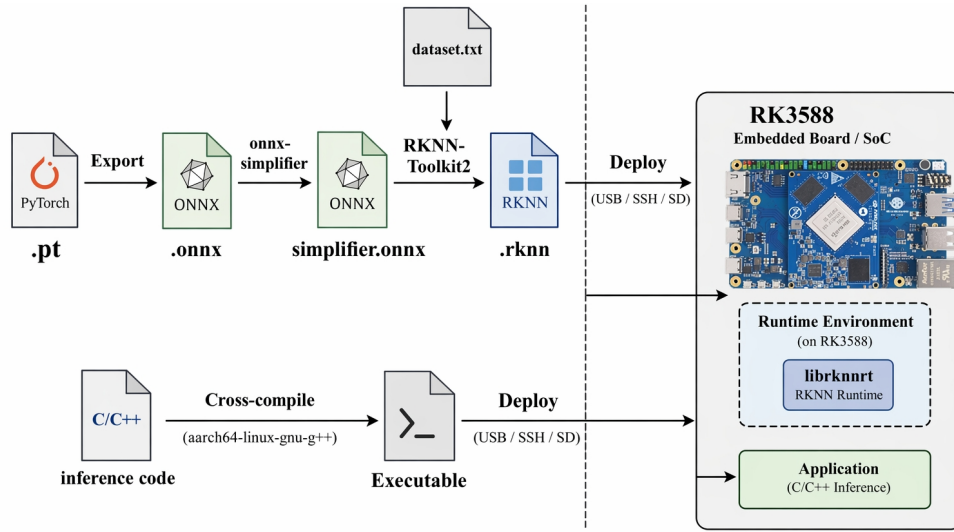


Fig. 9 Embedded device deployment process

Specifically, the trained weight file is first converted into an open neural network model in ONNX format, and the model is subsequently quantized. The converted ONNX model is then transformed into an RKNN model supported by the RK3588 embedded platform. Finally, inference is performed by invoking the on-board NPU API, and the corresponding detection results are obtained. The real-time inference results are shown in Fig. 10.

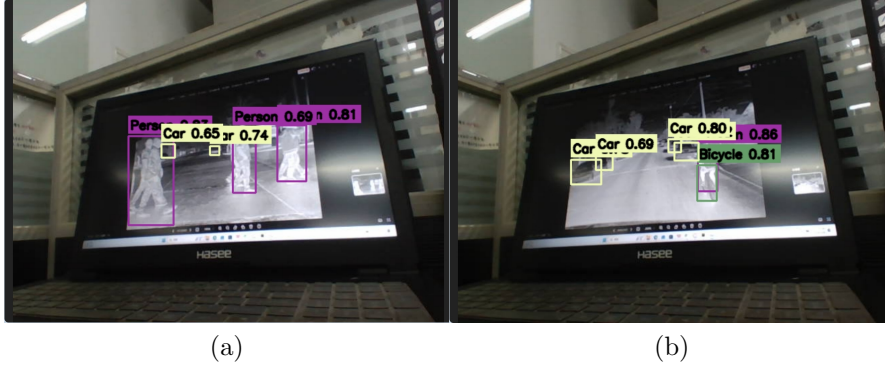


Fig. 10 Inference on the RK3588 Hardware Platform

To evaluate the performance of the proposed model on hardware platforms, experiments are conducted on both the RTX 3060Ti GPU and the RK3588 platform. Comparative analysis is performed in terms of computational cost, the number of parameters, memory usage, inference speed, and power consumption. The corresponding results are presented in Table 7. Before deploying the model on the RK3588 platform, quantization is applied to reduce the computational burden and system memory access overhead. As can be seen from Table 7, although the GPU achieves higher precision and faster inference speed, it is accompanied by higher energy consumption. In contrast, the RK3588 exhibits lower computational load, memory access overhead, and power consumption than the GPU. These results demonstrate that the proposed detection algorithm is more advantageous for embedded deployment.

Table 7 Analysis of test results on different devices

Device	$mAP@0.5$	FLOPs/G	Memory/MB	Frame rate/FPS	Power/W
GPU	82.3	6.6	553	148	82
RK3588	79.7	6.3	231	24	3.6

5 Conclusion

This paper addresses the challenges of weak texture, low contrast, missed detection of small targets, and deployment constraints in infrared target detection under complex

backgrounds, and proposes an infrared target detection network based on an adaptive gating mechanism, termed ABFINet. The proposed method is collaboratively designed from four aspects, namely feature extraction, cross-level fusion, detection head optimization, and model compression. By introducing the EMSC, ABFI, and DATA modules, together with the LAMP pruning strategy, the proposed method improves detection performance while meeting the requirement of model lightweights. Experimental results show that ABFINet achieves favorable detection performance on the public FLIR and M3FD datasets, demonstrating that the proposed method can effectively enhance the feature representation capability of infrared targets under complex backgrounds and improve the detection performance for small targets. On this basis, the model is further deployed and validated on an embedded hardware platform. The results show that the proposed method maintains satisfactory detection accuracy while also achieving favorable real-time inference capability, thereby demonstrating its feasibility for practical application scenarios.

Nevertheless, the proposed method still has room for further improvement in scenarios involving extremely small targets, severely deficient target textures, and strong background interference. Meanwhile, the adaptability and acceleration performance of the pruned model on different hardware platforms still require further optimization. Future work will focus on improving robustness in complex scenarios and optimizing cross-platform deployment, to further improve the practical applicability of the model.

Author Contributions GuoPing Wang: conceived the research, designed the proposed method, conducted the main experiments, analyzed the experimental results, and wrote the original manuscript. Bo Zhang: supervised the research, provided guidance on the methodology, and reviewed and revised the manuscript. KeKe Jing: assisted with data organization, experimental validation, and visualization. WenZhao Gao :assisted with result analysis, figure preparation, and manuscript proofreading. All authors discussed previous versions of the manuscript and approved the final manuscript.

Data Availability The datasets used in this study are publicly available, including the FLIR dataset and the M3FD dataset.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Yang K, Xiang W, Chen Z, Zhang J, Liu Y (2024) A review on infrared and visible image fusion algorithms based on neural networks. *Journal of Visual Communication and Image Representation* 101:104179
2. Ma T, Yang Z, Liu B, Sun S (2023) A lightweight infrared small target detection network based on target multiscale context. *IEEE Geosci Remote Sens Lett* 20:1–5

3. Li Y, He F, Zhang Q, Zhang W (2025) Lightweight Target Omni-Directional Enhancement Network for infrared small target detection. *Infrared Phys Technol* 151:106058
4. Zhang P, Jing Y, Liu G, Chen Z, Wu X, Sasaki O, Pu J (2025) Infrared dim tiny-sized target detection based on feature fusion. *Scientific Reports* 15:4355
5. Chen H, Yuan P, Liu W, Li F, Wang A (2025) A novel object detection algorithm combined YOLOv11 with dual-encoder feature aggregation. *Sensors* 25(23):7270
6. Wu P, Zuo Z, Su S, Zhao B (2025) Infrared target detection based on image enhancement and an improved feature extraction network. *Drones* 9(10):695
7. Chang Y, Ma D, Ding Y, Chen K, Zhou D (2024) Lightweight design for infrared dim and small target detection in complex environments. *Remote Sensing* 16(20):3761
8. Zhao Z, Ling H, Qin X, Xiao H, Wu J, Liu C (2025) STMG-PCNet: A lightweight shape-texture mutual-guided predictor-corrector network for infrared small target detection. *IEEE J Sel Topics Appl Earth Observ Remote Sens.* <https://doi.org/10.1109/JSTARS.2025.3646741>
9. Li J, Zhang W, Zhang X Y, Guo P, Mu C, Sun C (2025) A lightweight model for infrared photovoltaic panel defect detection via UAV-based inspection systems. *IEEE Trans Instrum Meas* 74:5050412
10. Ma P, Wang J, Pang D, Shan T, Sun J, Jin Q (2024) Improved weighted local contrast method for infrared small target detection. *Journal of Beijing Institute of Technology* 33(1):19–27
11. Wang H, Hu Y, Wang Y, Cheng L, Gong C, Huang S, Zheng F (2024) Infrared small target detection based on weighted improved double local contrast measure. *Remote Sensing* 16(21):4030
12. Shi S, Song Y (2024) A novel filtering-based detection method for small targets in infrared images. *Computers, Materials & Continua* 81(2):2911–2934
13. He Y, Zhang R, Xi C, Zhu H (2024) Learning background restoration and local sparse dictionary for infrared small target detection. *Optoelectronics Letters* 20(7):437–448
14. Zhang M, Li X, Gao F, Guo J, Gao X, Zhang J (2025) SAIST: Segment Any Infrared Small Target Model Guided by Contrastive Language-Image Pretraining. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 9549–9558

15. Yue X, Liu L, Du Y (2025) IR-DETR: An efficient detection transformer with multi-layer feature fusion for infrared small targets. *Infrared Phys Technol* 150:105926
16. Wang H, Ma K, Yue J, Li Y, Huang J, Liu J, Li L, Wang X, Cai N, Gao S (2025) Small-target detection based on improved YOLOv8 for infrared imagery. *Electronics* 14(5):947
17. Jiang L, Shen Y, Da M, Hu J, Zhang Z (2025) GML-YOLO: a lightweight infrared small target detection algorithm. *Meas Sci Technol* 36(8):085202
18. Xu W, Yin Q, Xu C, Zhao Z, Li Y, Huang D (2025) MFFOD: multidomain feature fusion object detector for infrared images. *Meas Sci Technol* 36(3):035401
19. Zhang T, Li L, Cao S, Pu T, Peng Z (2023) Attention-guided pyramid context networks for detecting infrared small target under complex background. *IEEE Trans Aerosp Electron Syst* 59(4):4250–4261
20. Chung W Y, Lee I H, Park C G (2023) Lightweight infrared small target detection network using full-scale skip connection U-Net. *IEEE Geosci Remote Sens Lett* 20:7000705
21. Gu Y, Guo Y, Xie W, Wu Z, Dong S, Xie G, Xu W (2025) MDSF: A plug-and-play block for boosting infrared small target detection in YOLO-based networks. *IEEE Trans Geosci Remote Sens* 63:1–14
22. Khan H, Usman M T, Koo J (2025) Bilateral feature fusion with hexagonal attention for robust saliency detection under uncertain environments. *Information Fusion* 121:103165
23. Yuan X, Yang C, Chen Y, Zhang Y, Zhong A, Zheng Q (2025) FDPF-Net: A full-scale dynamic pyramid fusion network for infrared small target detection. *IEEE Trans Geosci Remote Sens* 63:5623615, 1–15
24. Han K, Wang Y, Tian Q, Guo J, Xu C, Xu C (2020) GhostNet: More features from cheap operations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 1580–1589. <https://doi.org/10.1109/CVPR42600.2020.00165>
25. Sun H, Wen Y, Feng H, Zheng Y, Mei Q, Ren D, Yu M (2024) Unsupervised bidirectional contrastive reconstruction and adaptive fine-grained channel attention networks for image dehazing. *Neural Networks* 176:106314
26. Feng C, Zhong Y, Gao Y, Scott M R, Huang W (2021) TOOD: task-aligned one-stage object detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp 3490–3499. <https://doi.org/10.1109/ICCV48922.2021.00349>

27. Lee J, Park S, Mo S, Ahn S, Shin J (2020) Layer-adaptive sparsity for the magnitude-based pruning. arXiv preprint [arXiv:2010.07611](https://arxiv.org/abs/2010.07611)
28. FLIR (2018) Free Teledyne FLIR thermal dataset for algorithm training. Available online: <https://www.flir.com/oem/adas/adas-dataset-form/>
29. Liu J, Fan X, Huang Z, Wu G, Liu R, Zhong W, Luo Z (2022) Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 5802–5811
30. Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick C L (2014) Microsoft COCO: common objects in context. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
31. Jocher G, Chaurasia A, Qiu J (2023) Ultralytics YOLOv8. Available online: <https://github.com/ultralytics/ultralytics>
32. Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, Liu Y, Chen J (2024) DETRs beat YOLOs on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 16965–16974 <https://doi.org/10.1109/CVPR52733.2024.01605>
33. Wang C Y, Yeh I H, Liao H Y M (2024) YOLOv9: learning what you want to learn using programmable gradient information. arXiv preprint [arXiv:2402.13616](https://arxiv.org/abs/2402.13616)
34. Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, Ding G (2024) YOLOv10: real-time end-to-end object detection. Adv Neural Inf Process Syst 37:107984–108011. <https://doi.org/10.52202/079017-3429>
35. Jocher G, Qiu J, Chaurasia A (2024) Ultralytics YOLO11. Available online: <https://github.com/ultralytics/ultralytics>
36. Lei M, Li S, Wu Y, Hu H, Zhou Y, Zheng X, Ding G, Du S, Wu Z, Gao Y (2025) YOLOv13: real-time object detection with hypergraph-enhanced adaptive visual perception. arXiv preprint [arXiv:2506.17733](https://arxiv.org/abs/2506.17733)
37. Zeng Y, Liang T, Jin Y, Li Y (2024) MMI-Det: exploring multi-modal integration for visible and infrared object detection. IEEE Trans Circuits Syst Video Technol 34(11):11198–11213
38. Shen J, Chen Y, Liu Y, Zuo X, Fan H, Yang W (2024) ICAFusion: iterative cross-attention guided feature fusion for multispectral object detection. Pattern Recognit 145:109913

39. Fu H, Wang S, Duan P, Xiao C, Dian R, Li S, Li Z (2024) LRAF-Net: long-range attention fusion network for visible–infrared object detection. *IEEE Trans Neural Netw Learn Syst* 35(10):13232–13245
40. Guo J, Gao C, Liu F, Meng D, Gao X (2025) DAMSDet: dynamic adaptive multispectral detection transformer with competitive query selection and adaptive feature fusion. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp 464–481. https://doi.org/10.1007/978-3-031-73383-3_27
41. Tan M, Pang R, Le Q V (2020) EfficientDet: scalable and efficient object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 10778–10787. <https://doi.org/10.1109/CVPR42600.2020.01079>
42. Wang C, He W, Nie Y, Guo J, Liu C, Han K, Wang Y (2023) Gold-YOLO: efficient object detector via gather-and-distribute mechanism. In: *Advances in Neural Information Processing Systems* 36, pp 51094–51112. <https://doi.org/10.52202/075280-2224>
43. Yang G, Lei J, Zhu Z, Cheng S, Feng Z, Liang R (2023) AFPN: asymptotic feature pyramid network for object detection. In: *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp 2184–2189. <https://doi.org/10.1109/SMC53992.2023.10394415>
44. Xu X, Jiang Y, Chen W, Huang Y, Zhang Y, Sun X (2022) DAMO-YOLO: a report on real-time object detection design. *arXiv preprint arXiv:2211.15444*