Denoising-Enhanced Coarse-to-Fine Infrared Small Target Detection with Attention Prior-Guided Knowledge Distillation

Houzhang Fang[1], Ruixuan Huang[1(✉)], Qiuhuan Chen[1], Xiaolin Wang[1], Yi Chang[2], and Luxin Yan[2]

[1] Xidian University, Xi'an, China
houzhangfang@xidian.edu.cn, {hrx,cqh,wxl}@stu.xidian.edu.cn
[2] Huazhong University of Science and Technology, Wuhan, China
{yichang,yanluxin}@hust.edu.cn

Abstract. Infrared small target detection (IRSTD) in high-resolution images is crucial for many practical applications, such as surveillance of unmanned aerial vehicles (UAVs) and UAV-based ground monitoring. However, IRSTD remains challenging due to the small size and weak features of targets, as well as significant interference from complex dynamic backgrounds. Existing detection methods often suffer from redundant computations on non-target background regions and insufficient exploitation of target context information, which limits their performance in complex backgrounds. To address these issues, we propose an efficient coarse-to-fine infrared small target detection framework with attention prior-guided knowledge distillation, termed ECFNet. In the coarse stage, we design a region binary classification network (RBCN) on grid-based multi-scale feature maps to efficiently recognize target-containing context region proposals. Moreover, we introduce a novel denoising-assisted training strategy that incorporates noisy ground-truth (GT) masks into RBCN feature maps. By training the network to reconstruct the GT masks via a denoising task, this strategy encourages the model to explicitly capture target-background context, thereby enhancing its ability to distinguish target proposals from background regions. In the fine stage, we customize a lightweight target detector to the coarse stage's region proposals for balancing accuracy and efficiency. Furthermore, we propose a knowledge distillation strategy guided by the teacher-student cross-attention prior. This mechanism directs the student to focus on critical target regions, thereby enhancing the discriminative feature representation for infrared small targets. Extensive experiments on three real infrared datasets demonstrate that our method outperforms both existing single-stage and two-stage approaches while maintaining high real-time processing efficiency.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-032-37135-5_9.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
P. Favaro et al. (Eds.): ECCV 2026, LNCS 17072, pp. 150–168, 2026.
https://doi.org/10.1007/978-3-032-37135-5_9

Keywords: Infrared small target detection · Coarse-to-fine framework · Denoising-assisted training · Prior-guided knowledge distillation

1 Introduction

Infrared small target detection (IRSTD) in high-resolution images is an important perception technology in many applications, such as surveillance of unmanned aerial vehicles (UAVs) [6–10,34], UAV-based ground surveillance [17], and remote sensing [28,46]. IRSTD remains highly challenging due to the small size and weak features of targets, as well as severe interference from complex dynamic backgrounds. These factors often result in both detection failures and false alarms [8,9,28].

Advanced detection methods have been introduced to address this challenge. Current methods can be broadly categorized into single-stage [6,9,10,12,35,36,38,40,45] and two-stage approaches [17,32,37]. Single-stage approaches focus on designing complex network architectures to extract infrared small target (IRST) features. However, IRSTs occupy only a small portion of the image, and thus most redundant feature computations are wasted on non-target background regions [17]. This not only limits detection accuracy but also hampers real-time performance, especially when processing high-resolution images (Fig. 1(a)).

Two-stage approaches typically first leverage dense prediction [21,25,32] (*e.g.*, segmentation) to generate initial mask proposals, which are subsequently refined through lightweight detectors [17] (Fig. 1(b)). However, dense prediction suffers from two primary limitations. First, during the initial stage, they rely predominantly on target-specific features for target proposal extraction, often neglecting crucial contextual cues. Considering the inherently subtle feature representation of small targets, this lack of context severely restricts the prediction precision. Second, while existing approaches [17,27,37,42] employ specialized network architectures of different complexities to boost performance, they lack an explicit mechanism to mitigate the severe interference from background clutter inherent in infrared images, inevitably leading to both false alarms and missed detections. Knowledge distillation (KD) is an effective strategy for enhancing feature representation in lightweight detectors. However, conventional KD methods typically focus on minimizing global feature discrepancies between teachers and students while overlooking the prior knowledge associated with small target regions, leading to limited performance gains.

To address the above issues, we propose an efficient coarse-to-fine detection framework (ECFNet) to effectively detect IRSTs (Fig. 1(c)). Specifically, in the coarse stage, we construct a region binary classification network (RBCN) to reformulate the problem into patch-level binary recognition on non-overlapping grid region-based multi-scale feature maps, identifying context region proposals that contain the targets. This significantly reduces the number of predictions required compared to dense prediction [17,37]. Recognizing that infrared small targets are often visually indistinguishable from complex backgrounds, we propose a denoising-assisted training (DAT) strategy. By injecting target-like noise

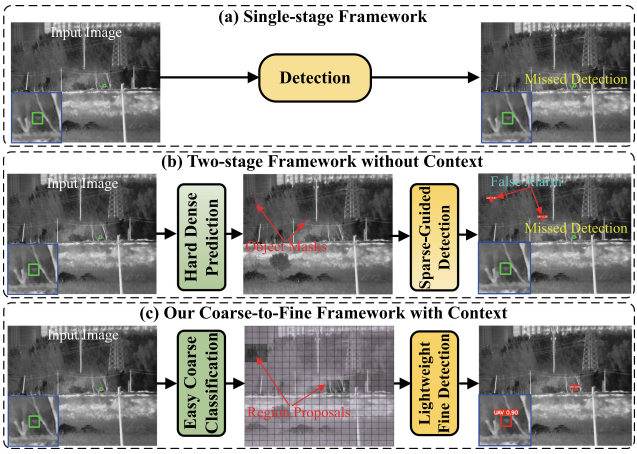


Fig. 1. Three main categories of methods for IRSTD. (a) Single-stage framework: detection models are directly applied to infrared images. (b) Two-stage framework without context: dense prediction (*e.g.*, segmentation) produces a target mask, followed by sparse detection on the mask. (c) Coarse-to-fine framework with context: the coarse stage generates target region proposals with context, and the fine stage applies a lightweight detector on the proposals. Our approach leverages contextual information to enhance the semantic understanding of the proposals, thereby improving target proposal recognition and providing valuable proposals for the fine detection.

to corrupt the ground-truth (GT) masks and integrating them into the RBCN feature space, DAT drives the network to reconstruct the original GT masks through a denoising task. This strategy encourages the network to explicitly model contextual target–background relationships, thereby improving its ability to discriminate target region proposals from complex backgrounds. It significantly improves the quality of region proposals and suppresses complex backgrounds. Moreover, this auxiliary denoising task stabilizes early optimization and effectively accelerates the convergence. To handle the issue of irregular targets being divided across multiple grid regions, we introduce ExpSlicer, which aggregates multi-scale pyramid region feature maps from the RBCN output to construct a complete target region. This approach ensures that the target is fully contained within a large single region patch while preserving critical contextual cues.

In the fine stage, we tailor a lightweight target detector to the coarse stage's region proposals for reliable IRSTD. Unlike mask-based refinement [17,32], our method leverages the complete target region during the fine stage, enhancing robustness to region recognition errors from the coarse stage. Additionally, our lightweight detector processes only a small local region instead of the whole image, achieving more accurate and efficient target detection.

Since lightweight detectors typically suffer from limited representation capability, we propose an attention prior-guided knowledge distillation (APKD) strat-

egy, leveraging cross-attention to propagate the teacher's spatial attention priors over critical target regions to the student. In this work, we first calculate the global cross-attention between the teacher and student features to model their semantic dependencies, enabling the student to benefit from the teacher's holistic understanding of the scene. We then modulate the student's features with the cross-attention weights, guiding the student detector to focus on the critical target regions emphasized by the teacher. The APKD between the teacher's features and the modulated student features further enhances the student's accurate perception of key target regions. Experimental results on infrared UAV, Car and IRSTD-1k [45] datasets demonstrate that our framework ECFNet outperforms state-of-the-art (SOTA) single-stage and two-stage methods while maintaining high detection efficiency.

Our contributions are summarized in three-fold:

- We propose a novel coarse-to-fine target detection framework that significantly enhances the detection performance of IRSTs under complex backgrounds while efficiently reducing computational costs, particularly in detecting high-resolution images.
- We propose a novel denoising-assisted training strategy that reconstructs GT masks from noisy counterparts while explicitly modeling target–background context, improving discrimination in complex backgrounds and accelerating convergence for IRST detection.
- We introduce a novel teacher-student cross-attention-guided KD mechanism that guides the student model to focus more effectively on critical target regions, thereby enhancing its ability to represent discriminative features of IRSTs.

2 Related Work

2.1 Infrared Small Target Detection

Recent years have seen significant progress in IRSTD, with many one-stage detectors proposed to enhance feature extraction and improve localization accuracy. These methods typically process the entire image in a dense manner, aiming to capture target-related information directly from complex backgrounds [1,16,18,36,38,45]. However, due to the low signal-to-clutter ratio and small target size in infrared images, one-stage detectors often incur substantial computational cost on irrelevant background regions. To mitigate this inefficiency, we propose a coarse-to-fine two-stage framework that focuses computational resources on target-relevant regions while maintaining high detection performance.

2.2 Two-Stage Detection Methods

Recently, advanced two-stage detection architectures [21,25,27,32,42] have been developed. ESOD [17] employs a single depth-wise convolution [2] to perform

dense predictions in the first stage, generating target masks as region proposals, followed by a sparse detection head in the second stage. However, the dense prediction is computationally expensive and challenging to optimize, while the sparse detection often discards critical contextual information, leading to increased missed detections and false alarms. QueryDet [37] dynamically selects the most appropriate feature level for each object through a query mechanism, enhancing overall small targets detection performance. While these methods prioritize either global search or local refinement, they often overlook the computational overhead of dense prediction [30] or the context loss in sparse cropping. Therefore, we construct an efficient coarse-to-fine architecture, using a lightweight detection model while retaining key contextual information.

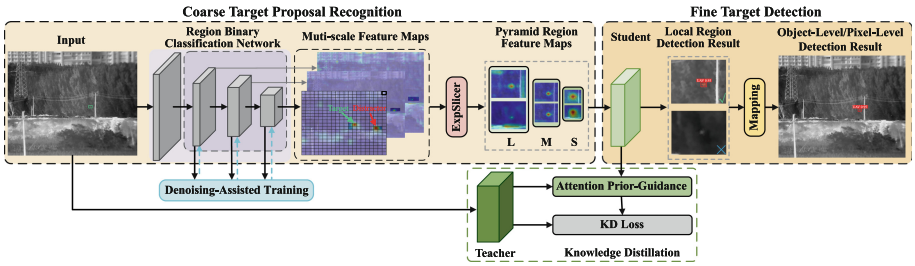


Fig. 2. The overall pipeline of ECFNet. ECFNet first employs the RBCN to generate context region proposals in the feature domain, followed by slicing these regions into square patches using the ExpSlicer. We leverage the DAT to further enhance the ability of RBCN to distinguish target proposals from complex backgrounds. In the fine stage, we customize a lightweight object-level or pixel-level detector in conjunction with an attention prior-guided knowledge distillation strategy to achieve more precise localization. Finally, we map the local region detection result back to the original image.

2.3 Knowledge Distillation

In recent years, feature domain KD [26] has been exploited to improve the performance of lightweight student models by transferring the knowledge of the teacher model. DCSF [3] adopts a cross-layer feature fusion strategy to bridge multi-scale semantic differences. Fang et al. [9] proposed scale-specific KD to enhance multi-scale target feature representation by concatenating the student feature map and the teacher feature map after channel attention modulation. However, such conventional distillation strategies frequently overlook the intrinsic semantic misalignment and representational discrepancy between the teacher and student models. Hence, we propose a dynamic knowledge distillation strategy based on cross-attention [20], which explicitly models the semantic interaction between teacher and student features, enhancing target perception while suppressing background interference.

3 Coarse-to-Fine Detection Method

3.1 Overall Architecture

In this section, we propose ECFNet, an efficient coarse-to-fine infrared small target detection framework, as illustrated in Fig. 2. In the coarse stage, the RBCN distinguishes foreground regions from background, and the ExpSlicer selects high-confidence patches for refinement. In the fine stage, a lightweight IRSTD model performs precise localization on the sliced patches, whose results are then mapped back to the input image. Furthermore, the DAT and APKD modules are incorporated into the coarse and fine stages, respectively, to further enhance detection performance.

3.2 Region Binary Classification Network (RBCN)

Conventional two-stage detection frameworks typically adopt a dense prediction step (*e.g.*, semantic segmentation) followed by sparse convolution-based detection. However, pixel-level classification introduces substantial computational overhead, making it difficult to balance accuracy and efficiency. Moreover, object masks lack essential contextual information. To overcome these limitations, we propose the RBCN, as illustrated in Fig. 3(a), which reformulates pixel-wise prediction as a grid-level binary classification task, thereby simplifying the detection process and reducing computational cost. The backbone of RBCN adopts a four-stage hierarchical architecture. The spatial resolution is progressively reduced while the channel width increases across stages to capture higher-level semantic representations. In the second to the fourth stages, three BasicBlocks are employed to facilitate efficient feature propagation and improve representation capacity. Specifically, the input image is partitioned into $N \times N$ non-overlapping grids, and RBCN outputs a probability mask M_{pred}, where each value denotes the likelihood of a target existing within the corresponding grid. A binary coding strategy is then applied, assigning a value of 1 to grids exceeding a predefined threshold (we set 0.5 here) and 0 otherwise. Additionally, multi-scale feature maps are extracted for downstream processing. For the coarse stage, we design a joint loss function combining binary cross-entropy (BCE) [44] and IoU loss [41]: $\mathcal{L}_{\text{coarse}} = \mathcal{L}_{\text{BCE}}(M_{pred}, M_{GT}) + \lambda_1 \cdot \mathcal{L}_{\text{IoU}}(M_{pred}, M_{GT})$, where λ_1 denotes a learnable weight initialized to 1.0 that balances the two terms.

3.3 Denoising-Assisted Training (DAT)

IRSTs often exhibit high spectral similarity with background clutter, posing significant challenges for effective feature discrimination. To provide high-quality region proposals for the fine detection stage, we introduce a DAT module in the coarse stage, as illustrated in Fig. 3(b). To guide the model in suppressing target-like distractors, DAT synthesizes noise N_i^l based on a 2D Gaussian function [39]: $N_i^l(x, y) = \exp\left(-\frac{(x-x_i)^2+(y-y_i)^2}{2\sigma^2}\right)$, where (x_i, y_i) denotes the center of the i-th synthetic distractor and σ controls the spatial spread. The superscript

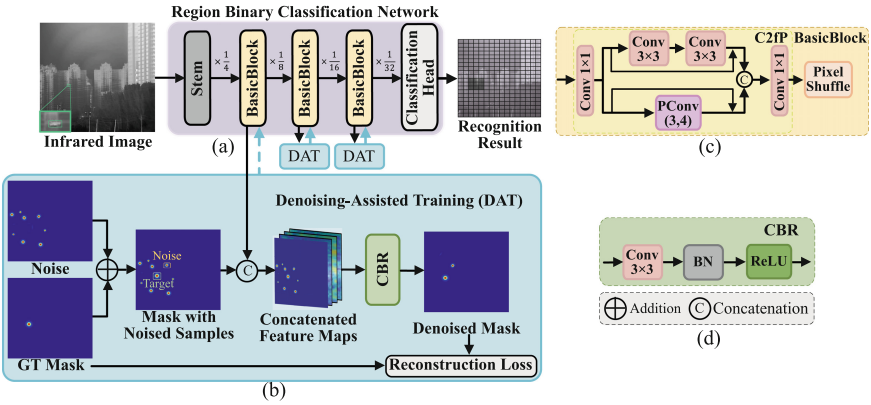


Fig. 3. Details of DAT. Multi-scale denoising prediction and GT are utilized for loss calculation. C2fP BasicBlock is a C2f-style module that mainly fuses a two-3×3-conv residual path with a PConv residual path. CBR denotes the Conv–BN–ReLU block. PConv(3, 4) [38] means the first PConv kernel is 3 and the second is 4.

$l \in \{1, 2, 3\}$ corresponds to multi-scale feature levels produced by the RBCN. These noises are added element-wise to the GT mask M^l_{GT} and concatenated with RBCN features F^l_{RBCN}: $F^l_{\mathrm{DAT}} = \mathrm{cat}\left(F^l_{\mathrm{RBCN}},\ N^l + M^l_{\mathrm{GT}}\right)$, where $\mathrm{cat}(\cdot)$ denotes channel-wise concatenation. The obtained features are sent to the subsequent auxiliary branch to obtain the denoising mask M^l_{DAT}. The DAT module is supervised by resolution-aligned GT masks GT^l, with the reconstruction loss computed as: $\mathcal{L}_{\mathrm{DAT}} = \sum_{l=1}^{3} \left(\mathcal{L}_{\mathrm{BCE}}(M^l_{\mathrm{DAT}}, M^l_{\mathrm{GT}}) + \lambda_2 \cdot \mathcal{L}_{\mathrm{IoU}}(M^l_{\mathrm{DAT}}, M^l_{\mathrm{GT}})\right)$, where λ_2 is a learnable weighting coefficient that balances the reconstruction objectives. By injecting structured target-like noise into the feature space, DAT simulates realistic distractors. Since real and synthetic targets share similar local features, the network is forced to exploit surrounding context rather than isolated responses, thereby improving target-background discrimination.

Discussion. Unlike DN-DETR [15], which introduces denoising queries for matching stabilization, DAT operates on region features by injecting target-like perturbations to simulate infrared clutter, thereby enhancing target-background discriminability under GT-mask supervision.

3.4 Context-Preserving Region Expslicer

After obtaining the high-quality region proposals M_{pred}, the customized ExpSlicer crops multi-scale feature maps F^l_{RBCN} into fixed-size rectangular patches that preserve complete targets, as shown in Fig. 4.

Given the multi-scale feature maps $F^l_{RBCN} \in \mathbb{R}^{B\times C\times H\times W}$ and $M_{pred} \in \{0,1\}^{B\times 1\times N\times N}$ generated by RBCN, we slice patches of feature maps centered on the positive positions of M_{pred} based on target coverage. B, C, H, and W denote the batch size, channels, height, and width, respectively. Each spatial patch in M_{pred} corresponds to a size of $h_p \times w_p$ in F^l, such that: $h_p = H/N, w_p =$

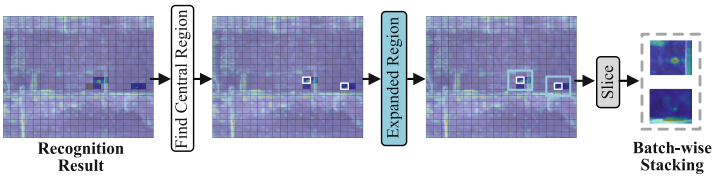


Fig. 4. ExpSlicer selects grids with the highest predicted target coverage (highlighted in white) from RBCN outputs as region centers and extracts 3×3 areas (shown in blue box) around them. (Color figure online)

W/N. For every valid position (g_x, g_y) such that $M_{pred}[b, 0, g_x, g_y] = 1$, we compute its center in the feature map as: $(x_c, y_c) = (g_x \cdot w_p,\ g_y \cdot h_p)$. Then, a region of size $(k \cdot w_p) \times (k \cdot h_p)$, centered at (x_c, y_c), is extracted, where k denotes the side length of the patch centered at (x_c, y_c). The hyperparameter k is set to 3 and N is set to 20 in our experiments. The cropped areas will be marked to prevent repeated cropping in subsequent operations. Let A be the total number of activated positions in M_{pred}. All patches in the same layer $\{P_i\}_{i=1}^{A} \in \mathbb{R}^{C\times(k\cdot h_p)\times(k\cdot w_p)}$ are stacked along the batch dimension to construct an aggregated representation $F^l_{\text{agg}} \in \mathbb{R}^{(A\cdot C)\times(k\cdot h_p)\times(k\cdot w_p)}$. Global non-maximum suppression eliminates redundant predictions arising from overlapping regions and closely targets. For boundary targets, we use zero-padding to ensure the same size of the patches.

Discussion. Such slicing strategy has two advantages. First, it ensures that the target is completely fed into the subsequent lightweight detection model without affecting the detection accuracy. Second, unlike fixed mask-based cropping strategies, ExpSlicer dynamically expands around predicted target regions, ensuring complete target coverage while preserving contextual information. To handle varying proposal counts, we set the proposal batch size to the average proposal number plus one, ensuring that over 50% of images fully contribute to training.

3.5 Attention Prior-Guided Knowledge Distillation (APKD)

In this work, we customize a lightweight detector in the fine stage, to balance accuracy and efficiency by incorporating channel-adaptive scaling and fewer detection heads. To further transfer the teacher's attention priors over critical target regions, we propose an APKD, as shown in Fig. 5. Mainstream conventional attention distillation methods typically

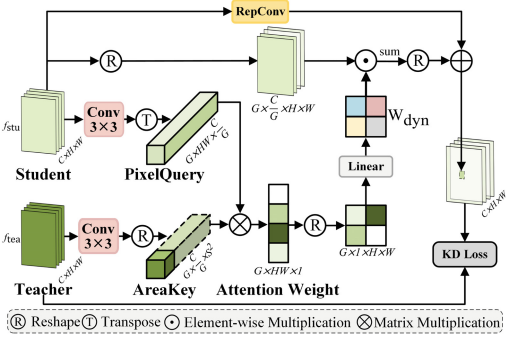


Fig. 5. Overview of our APKD.

flatten feature maps into 1D tokens, which weakens the preservation of spatial locality. To address this limitation, APKD performs teacher-student attention directly over multi-dimensional spatial features, explicitly maintaining the fine-grained spatial structure of IRSTs. We adopt YOLOv12 [29] as the teacher model.

Given the student feature map $f_{\text{stu}} \in \mathbb{R}^{B\times C\times H\times W}$, we apply a 3×3 convolution to divide it into G groups of lower-dimensional channels and obtain PixelQuery $\in \mathbb{R}^{B\times G\times \frac{C}{G}\times HW}$: $\text{PixelQuery} = \text{Conv}_{3\times 3}(f_{\text{stu}})$. This operation encodes local spatial details and fine-grained cues from the student. Meanwhile, the teacher's feature map f_{tea} is processed similarly generating $S \times S$ AreaKey: $\text{AreaKey} = \text{Reshape}\left(\text{Conv}_{3\times 3}(f_{\text{tea}})\right) \in \mathbb{R}^{B\times G\times \frac{C}{G}\times S^2}$, where $S = \sqrt{H \cdot W}$.

Next, we compute their attention weight, denoted as A, to global average representations of the feature map: $A = \text{Softmax}(\text{PixelQuery}^{\top} \cdot \text{AreaKey})$, where A encodes the spatial similarity between input tokens and globally pooled contextual representations from the teacher model. Subsequently, the affinity map is projected into convolution-like attention weights: $W_{\text{dyn}} = \text{Linear}(\text{Reshape}(A))$.

To further improve real-time performance and detection accuracy, we introduce three customized modifications. First, the channel dimensions of the neck network are adaptively aligned with the RBCN outputs to maintain feature consistency across stages and improve computational efficiency. Second, the original detection head is replaced with the stronger YOLOX head [11], where sequential $4\times$ and $2\times$ channel compression is adopted instead of YOLO11's uniform $4\times$ compression, reducing computational cost while improving detection performance. Third, for IRSTD, the number of detection heads is reduced from three to two, with the two shallowest feature maps used as inputs. More details are provided in supplementary material.

Unlike traditional knowledge distillation that passively aligns student features with those of the teacher, our approach actively modulates the student's perception via feature-level attention weight, making the student aware of semantically meaningful regions. This modulation is further enhanced through a context-guided residual gating generated by re-parameterize convolution (RepConv) [5]: $\text{output} = \text{Reshape}(\text{sum}(f_{\text{stu}} \odot W_{\text{dyn}})) + \text{RepConv}(f_{\text{stu}})$, where $\text{sum}(\cdot)$ denotes summation over the spatial dimensions of each channel. We define KD loss as $\mathcal{L}_{\text{KD}} = \sum_{i=1}^{L} w_i \cdot \mathcal{D}_{\cos}(F_s^i, F_t^i)$, where F_s^i and F_t^i denote the student and teacher features at the i-th level, respectively, and w_i is the weight corresponding to that level, set to $\frac{1}{i}$. $\mathcal{D}_{\cos}(\cdot)$ represents the cosine similarity.

Conventional cross-attention distillation methods typically flatten feature maps into 1D tokens, which weakens the preservation of spatial locality. To address this limitation, APKD performs attention directly over multi-dimensional spatial features, explicitly maintaining the fine-grained spatial structure of IRSTs. By using student features as queries and teacher features as keys, APKD guides the student to actively focus on teacher-highlighted critical regions. This teacher-student cross-attention is especially effective for lightweight detectors, improving target awareness and background suppression without increasing inference cost.

4 Experiment

4.1 Datasets and Evaluation Metrics

Datasets. We validate our method's effectiveness on three real IRST datasets, including UAV, Car, and IRSTD-1k [45], under complex backgrounds. Our custom-built UAV and Car datasets comprise 98,420 and 18,230 infrared images, respectively, all with a spatial resolution of 640×512 pixels. Our datasets are divided into training, validation, and test sets in a ratio of 7:2:1. More details of three datasets (*e.g.*, data source, target scale distribution, and background characteristics) are provided in the supplementary material.

Evaluation Metrics. To evaluate object-level detection performance, we adopt precision (P(%)), recall (R(%)), and AP_{50}(%) [9]. For pixel-level evaluation, we utilize Pd(%), Fa(10^{-6}), and nIoU(%) [45]. To evaluate the computational cost, we compute the floating-point operations (FLOPs, in G), the number of parameters (Params, in M), and the frames per second (FPS) with a batch size of 1 for a consistent comparison. We evaluate existing infrared segmentation methods by converting their predicted masks into detection boxes via the minimum bounding rectangles to calculate P, R, and AP [22,36,44]. Furthermore, our method concentrates on coarsely recognizing the target-containing region and locating them in the fine stage, two specific metrics are employed for the ablation study, Coarse Recognition Precision (CRP) and Coarse Recognition Recall (CRR):

$$\mathrm{CRP} = \frac{1}{M}\sum_{j=1}^{M} \mathbb{1}\left\{ \max_{i=1,\ldots,I} \left(\frac{|\mathrm{region}_{k\times k}^{j} \cap \mathrm{box}^{i}|}{|\mathrm{region}_{k\times k}^{j}|} \right) \geq \tau \right\}, \tag{1}$$

$$\mathrm{CRR} = \frac{1}{I}\sum_{i=1}^{I} \mathbb{1}\left\{ \max_{j=1,\ldots,M} \left(\frac{|\mathrm{region}_{k\times k}^{j} \cap \mathrm{box}^{i}|}{|\mathrm{box}^{i}|} \right) \geq \tau \right\}. \tag{2}$$

Here, M and I denote the total number of predicted patches and GT boxes, respectively. The selected regions after applying ExpSlicer and GT boxes are denoted as $\mathrm{region}_{k\times k}$ and box. The thresholds τ are empirically set to 0.9 in our experiments to ensure high spatial alignment between patches and targets. These metrics effectively evaluate the coarse stage: a high CRP indicates accurate region selection, while a high CRR reflects comprehensive coverage of true target regions, preserving essential contextual information.

4.2 Implementation Details

We train our ECFNet with Adam using a learning rate of 0.01. Both coarse-stage and fine-stage training lasts for 100 epochs, with a weight decay of 10^{-1} and a batch size of 12. Additionally, we employ dictinct object-level and pixel-level heads to evaluate our framework on different tasks (More details are provided in the supplementary material). We use random cropping, horizontal flipping, and brightness adjustment for augmentation. Following standard KD protocols, the teacher is trained with the same data, augmentations, and input resolution

as the student. During distillation, the teacher remains frozen and only provide supervision signals. Our experiments are conducted on an NVIDIA RTX 4090 with CUDA 12.4 and PyTorch 2.5. We select YOLOv12-L [29], YOLO11-L [13] as the convolution-based detection methods, D-FINE-L [24], DQ-DETR [12] and DINO [43] are the representative Transformer-based detection methods, ESOD-L [17] and QueryDet [37] are two-stage detection methods, EFLNet [36], DAGNet [9], PConv (MSHNet) [38], GSFANet [4], MSHNet [18], LESPS [40], UIU-Net [35], ISNet [45], DNANet [14] are designed for IRSTD.

4.3 Quantitative Results

As shown in Table 1, the object-level detection model D-FINE-L achieves relatively high P and R, yet it has a high computational cost. In contrast, the two-stage detection method ESOD-L reduces computational overhead, but it has lower P and R. Our ECFNet, while maintaining on par with other SOTA methods in terms of P, achieves the highest values for R and AP_{50}. Compared with methods in Table 1, our model achieves an approximately 60% reduction in computational cost under comparable detection performance, and yields +9.68 and +8.58 improvements in P and R, under similar computational complexity. For pixel-level detection, ECFNet gains a +2.30 and +1.92 improvements in Pd and nIoU, compared to the second-best method.

Table 1. Performance comparison against SOTA detectors on real infrared UAV, Car and IRSTD-1k datasets. The best and second-best results are highlighted in **Bold** and underline within each category (Object-Level and Pixel-Level) respectively.

Object-Level	Pub'Year	UAV			Car			IRSTD-1k			FLOPs↓	Params↓	FPS↑
		P↑	R↑	AP_{50} ↑	P↑	R↑	AP_{50} ↑	P↑	R↑	AP_{50} ↑			
ESOD-L	TIP'25	86.28	83.05	85.63	94.75	91.73	93.32	78.59	76.37	78.33	38.1	72.8	86
YOLOv12-L	NeurIPS'25	95.92	87.91	93.12	96.09	82.87	91.94	87.26	72.78	81.61	88.9	26.4	76
D-FINE-L	ICLR'25	95.27	91.32	93.18	94.38	91.63	93.68	85.67	81.56	83.53	90.7	30.6	40
YOLO11-L	2024	95.53	89.85	95.39	96.04	85.35	92.33	84.53	77.12	82.19	86.6	25.3	78
DQ-DETR	ECCV'24	89.84	78.23	84.22	88.83	74.75	83.15	77.56	75.28	76.93	578.5	58.7	18
DINO	ICLR'23	94.36	83.28	91.27	95.62	87.13	92.46	79.66	72.47	74.28	398.6	47.0	26
QueryDet	CVPR'22	64.59	77.16	64.16	75.81	80.28	76.37	61.92	60.93	61.69	888.4	35.3	9
EFLNet	TGRS'24	93.28	86.89	91.92	93.18	90.53	92.92	87.03	81.70	84.32	102.2	38.3	69
DAGNet	TII'23	88.72	80.34	85.31	83.59	86.43	84.86	84.59	78.82	81.23	170.7	150.5	40
Ours	-	**95.96**	**91.63**	**95.87**	**96.55**	**92.16**	**96.06**	**90.18**	**84.71**	**89.23**	**31.7**	**22.7**	**90**
Pixel-Level	Pub'Year	Pd↑	Fa↓	nIoU↑	Pd↑	Fa↓	nIoU↑	Pd↑	Fa↓	nIoU↑	FLOPs↓	Params↓	FPS↑
PConv (MSHNet)	AAAI'25	81.26	12.11	54.23	83.26	14.66	55.98	92.20	10.70	67.93	47.1	15.6	72
GSFANet	TGRS'25	82.73	14.81	51.96	81.86	19.25	48.87	92.92	20.17	60.52	63.6	8.8	62
MSHNet	CVPR'24	77.22	17.07	45.65	82.74	11.74	51.67	93.88	15.03	67.16	38.2	15.5	78
LESPS	CVPR'23	75.63	23.77	35.68	70.16	23.55	44.57	89.85	14.22	53.19	88.5	**4.7**	71
UIU-Net	TIP'23	79.75	18.54	40.37	76.54	18.82	46.51	92.93	26.87	61.11	140.6	54.5	51
ISNet	CVPR'22	80.66	19.22	41.58	80.37	16.99	49.85	92.33	18.57	64.23	185.4	9.7	63
DNANet	TIP'22	74.69	18.83	33.69	69.13	16.55	43.23	93.27	17.61	62.73	89.3	**4.7**	47
Ours	-	**83.56**	**11.69**	**56.11**	**85.45**	**10.52**	**58.21**	**94.91**	**9.94**	**69.67**	**34.3**	24.1	**85**

We observe that reducing the theoretical computation of general detectors does not inherently yield higher FPS (e.g., D-FINE: 40 FPS with 90.7 FLOPs). In contrast, some infrared detectors sustain higher FPS despite larger computation (e.g., UIU-Net: 51 FPS with 140.6 FLOPs). This difference lies in: general box-based detectors, equipped with stronger feature extraction and multi-scale fusion modules, incur higher inference latency despite lower theoretical complexity. Conversely, infrared detectors are typically optimized for hardware-friendly and tuned for pixel-level enhancement on particular infrared datasets. Such designs enable higher runtime efficiency without sacrificing performance.

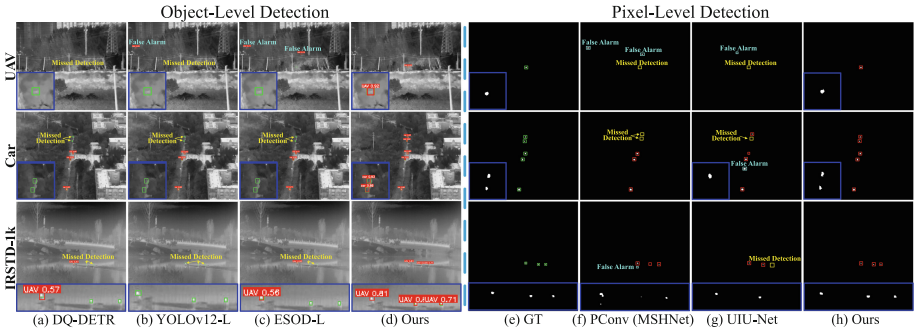


Fig. 6. Qualitative comparisons between our ECFNet and other SOTA methods on the UAV, Car and IRSTD-1k dataset. The green and red boxes denote the GT and the detection results. Closed-up views are shown in the left bottom corner and bottom. (Color figure online)

4.4 Qualitative Results

From Fig. 6, our method effectively detects IRSTs in three representative challenging backgrounds. Even in the presence of challenging background interference, our ECFNet accurately detects targets without false alarms. This is attributed to the DAT in the coarse stage for improved target–background discrimination, and the incorporation of APKD in the fine stage, which guides the student with attention priors from the teacher. By effectively suppressing irrelevant background information before fine stage detection and retaining critical context information, our method ensures high precision and robustness for such scenes. The difficulty in suppressing background clutter in infrared images stems from their low texture and contrast, which limits the effectiveness of general-purpose detectors. Meanwhile, the sparse detection of two-stage frameworks like ESOD leads to insufficient contextual understanding, resulting in false alarms. Other methods tend to produce missed detections or false alarms.

4.5 Ablation Study

Extensive ablation experiments are conducted to validate the effectiveness of our method. More ablation studies and details of baselines in experiments (Table 2, Table 7, Table 6, Table 7 and Table 8) are provided in the supplementary material.

Table 2. Ablation of the main components of our method. **Bold** denotes the best result.

Model		Metrics		
DAT	APKD	P↑	R↑	AP_{50} ↑
-	-	88.23	87.32	87.26
-	✓	89.17(+0.94)	89.56(+2.24)	89.30(+2.10)
✓	-	93.45(+5.22)	90.32(+3.00)	92.71(+5.45)
✓	✓	**95.96**(+7.73)	**91.63**(+4.31)	**95.87**(+8.61)

Table 3. Ablation study on the C2fP module in our RBCN. **Bold**: best, underline: second best.

Module	CRP↑	CRR↑	FLOPs↓	Params↓
Conv	91.20	93.51	11.3	3.1
DWConv	90.06	92.57	**1.8**	**0.4**
DCNv3	92.17	94.42	9.8	0.9
SwinTransformer	**96.22**	95.41	34.5	4.1
C2fP	94.13	**95.69**	8.9	2.6

Impact of Proposed Main Modules. In Table 2, integrating either DAT or APKD consistently improves performance over our self-built baseline (row 1). Incorporating APKD yields a +2.10 gain in AP_{50}, while introducing DAT brings a larger improvement of +5.45, highlighting the importance of high-quality coarse detection. When both modules are applied, performance peaks with increases of +7.73, +4.31, and +8.61 in P, R, and AP_{50}. These two components not only enhance their corresponding detection stages but also complement each other for optimal results. Specifically, DAT injects pseudo-sample noise to improve background–target discrimination, according to Fig. 7(a), while APKD transfers spatial attention from a stronger teacher, guiding the lightweight student to focus on targets and suppress background interference as shown in Fig. 7(b).

Impact of the Main Module in RBCN. In RBCN, the C2fP is the main feature extractor. As shown in Table 3, it achieves the highest CRR with competitive CRP and moderate computational cost. Although SwinTransformer [19] attains slightly higher CRP, it incurs the highest FLOPs. Since the fine stage refines region proposals, the coarse stage can tolerate a few false alarms as long as it preserves high CRR. Unlike modules that improve accuracy through heavier computation, the proposed C2fP enhances directional awareness and spatial modeling. This design enriches feature representation with negligible overhead, achieving better accuracy–efficiency trade-offs than both heavy (DCNv3 [33], SwinTransformer) and lightweight (Conv, DWConv [2]) counterparts.

Impact of Patch Resolution Predicted by RBCN. As shown in Table 4, we compare different RBCN patch resolutions under the same input size. The 10×10 grid achieves CRP and CRR close to the 20×20 setting, showing the robustness of RBCN. However, larger patches introduce more background clutter to the fine detector and slightly reduce AP_{50}, while the finer 40×40 grid lacks

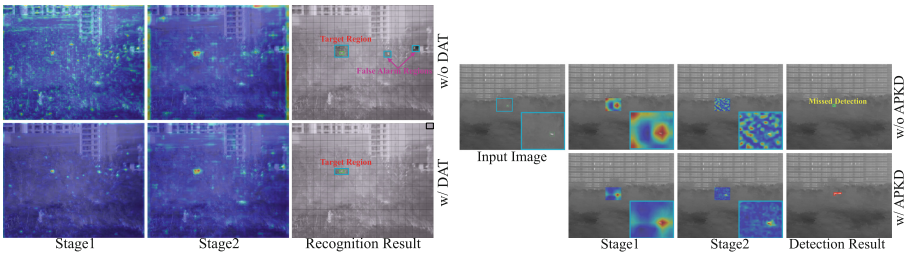


Fig. 7. (a) Comparison of feature maps and results from the RBCN at different stages with and without DAT. Blue boxes denote the recognition result, and green boxes indicate the GT box. (b) Comparison of feature maps and results of the lightweight fine detection model at different stages with and without APKD. Blue boxes denote the ExpSlicer' (Color figure online)s result.

Table 4. Ablation study on patch resolution predicted by RBCN. CRP and CRR evaluate region proposal quality, while AP_{50} reflects its impact on fine detection. **Bold** indicate the best results.

Patches	CRP ↑	CRR ↑	AP_{50} ↑	FLOPs (G) ↓	Params (M) ↓
10 × 10	93.42	95.57	93.84	**26.4**	15.9
20 × 20	**94.13**	**95.69**	**95.7**	26.9	15.9
40 × 40	82.64	79.43	81.62	28.5	15.9

Table 5. Ablation study of DAT. **Bold** and underline indicate the best and the second best results.

Method	CRP↑	CRR↑
-	89.38	93.45
+Loss	92.80	94.35
+GT	92.62	94.11
+Noise	**94.13**	**95.69**

Table 6. Ablation study on inserting DAT into different basic blocks of RBCN. **Bold**: best, underline: second best.

Blocks			Metrics		Blocks			Metrics	
1	2	3	CRP↑	CRR↑	1	2	3	CRP↑	CRR↑
-	-	-	89.38	93.45	✓	✓	-	92.62	93.57
✓	-	-	92.67	93.52	✓	-	✓	93.80	94.72
-	✓	-	91.19	91.36	-	✓	✓	93.83	94.06
-	-	✓	91.77	93.35	✓	✓	✓	**94.13**	**95.69**

sufficient context for weak IRSTs and increases prediction complexity. Thus, we adopt the 20×20 resolution as a balanced setting that preserves semantic context while suppressing clutter.

Impact of DAT. As shown in Table 7, we evaluate the effectiveness of DAT under four configurations. Moreover, compared with directly computing the loss on the feature map, simply concatenating the GT mask even leads to a performance drop. This is because the direct concatenation of the GT mask weakens the model's learning capability [15] and fails to optimize the backbone feature extraction effectively. In contrast, incorporating our DAT during training

achieves the best overall performance. As illustrated in Fig. 8(a), the inclusion of DAT significantly accelerates the training convergence speed and achieves a higher performance ceiling compared to the baseline.

Impact of DAT Insertion Locations. As detailed in Table 6, although applying DAT to individual blocks provides consistent gains, deploying it globally across all RBCN blocks yields the best results.

Impact of High-Resolution Images on Complexity and Accuracy. In Fig. 8(b), we compare the FLOPs and AP of PConv (MSHNet), ESOD-S, and our ECFNet on high-resolution images. It demonstrates that with the same input resolution, our method consistently reduces the FLOPs by up to about 40% while achieving better performance in IRSTD.

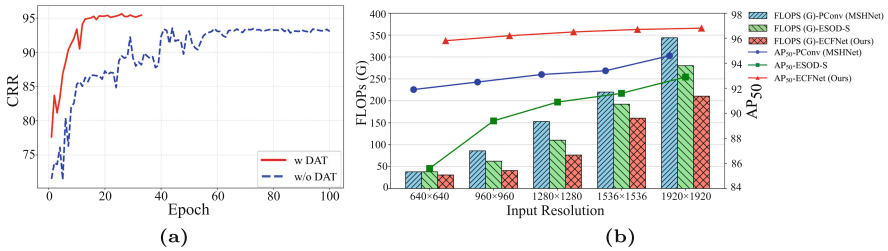


Fig. 8. (a) Convergence curves of the RBCN with and without the DAT module. (b) Performance comparison among PConv (MSHNet), ESOD-S, and our ECFNet. The average computation is reported only for cases with coarse-stage results.

Impact of Knowledge Distillation Strategy. As shown in Table 7, we conduct experiments using several knowledge distillation methods on our ECFNet. It can be seen that our APKD achieves the best results on all evaluated metrics, surpassing the second-best approach by up to 4.72 in AP_{50} and 1.31 in R. Furthermore, as shown in Table 8, three distinct teacher models with up to a 7× difference in complexity all yield performance gains exceeding 2.15 in AP_{50}. This demonstrates APKD's robustness to teacher selection.

Table 7. Ablation study of APKD. **Bold** and underline indicate the best and the second best results.

Method	P↑	R↑	AP_{50} ↑
-	93.45	90.32	92.71
Plain [26]	95.04	85.72	91.15
DRKD [23]	95.17	88.51	90.13
CrossKD [31]	93.82	84.14	87.70
DCSF [3]	95.82	86.73	89.29
APKD (Ours)	**95.96**	**91.63**	**95.87**

Table 8. Ablation study of APKD's teacher model. **Bold**: best, underline: second best. $*$ denotes the default teacher in our APKD.

Teacher	P↑	R↑	AP_{50}↑	FLOPs↓
-	93.45	90.32	92.71	-
PConv (MSHNet)	94.82	90.63	93.72	**12.4**
ESOD-L	94.08	90.89	93.32	38.1
YOLO11-L	95.39	91.27	95.43	86.6
ISNet	94.31	90.28	92.54	185.4
YOLOv12-L*	**95.96**	**91.63**	**95.87**	88.9

5 Conclusion

In this paper, we propose ECFNet, an efficient coarse-to-fine infrared small target detection framework. In the coarse stage, we introduce RBCN to identify target region proposals, while DAT enhances target-background discrimination through a denoising task and accelerate convergence. In the fine stage, we tailor a lightweight detector to efficiently detect the target, and APKD guides it toward critical target regions via cross-attention, thereby enhancing discriminative feature representation for IRSTs. Experimental results demonstrate that ECFNet outperforms SOTA methods under complex backgrounds while maintaining high efficiency, making it suitable for embedded deployment.

Acknowledgments. The authors thank the Area Chair and reviewers for their valuable time and insightful comments. This work was supported by the Open Research Fund of the National Key Laboratory of Multispectral Information Intelligent Processing Technology under Grant 61421132301, and in part by the National Natural Science Foundation of China under Grants 62671485 and 62301228.

References

1. Chen, S., Ji, L., Zhu, S., Ye, M., Ren, H., Sang, Y.: Toward dense moving infrared small target detection: new datasets and baseline. IEEE Trans. Geosci. Remote Sens. **62**, 1–13 (2024)
2. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807 (2017)
3. Dai, T., Lin, Y., Guo, H., Wang, J., Zhu, Z.: DCSF-KD: dynamic channel-wise spatial feature knowledge distillation for object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 2627–2635 (2025)
4. Deng, C., Zhao, Z., Xu, X., Xia, Y., Li, J., Plaza, A.: GSFANet: global spatial-frequency attention network for infrared small target detection. IEEE Trans. Geosci. Remote Sens. **63**, 1–17 (2025)

5. Ding, X., et al.: UniRepLKNet: a universal perception large-kernel convnet for audio video point cloud time-series and image recognition. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5513–5524 (2024)
6. Fang, H., Ding, L., Wang, L., Chang, Y., Yan, L., Han, J.: Infrared small UAV target detection based on depthwise separable residual dense network and multiscale feature fusion. IEEE Trans. Instrum. Meas. **71**, 1–20 (2022)
7. Fang, H., Guo, S., Chen, Q., Chang, Y., Yan, L.: Spatio-temporal context learning with temporal difference convolution for moving infrared small target detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3840–3848 (2026)
8. Fang, H., et al.: DANet: multi-scale UAV target detection with dynamic feature perception and scale-aware knowledge distillation. In: Proceedings of the 31st ACM International Conference on Multimedia (ACM MM), pp. 2121–2130 (2023)
9. Fang, H., Liao, Z., Wang, X., Chang, Y., Yan, L.: Differentiated attention guided network over hierarchical and aggregated features for intelligent UAV surveillance. IEEE Trans. Industr. Inf. **19**(9), 9909–9920 (2023)
10. Fang, H., et al.: Detection-friendly nonuniformity correction: a union framework for infrared UAV target detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11898–11907 (2025)
11. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: exceeding YOLO series in 2021 (2021). arXiv preprint arXiv:2107.08430
12. Huang, Y.X., Liu, H.I., Shuai, H.H., Cheng, W.H.: DQ-DETR: DETR with dynamic query for tiny object detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 290–305 (2024)
13. Jocher, G., Qiu, J.: Ultralytics YOLO11. Software (2024). Version 11.0.0
14. Li, B., et al.: Dense nested attention network for infrared small target detection. IEEE Trans. Image Process. **32**, 1745–1758 (2023)
15. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: DN-DETR: accelerate DETR training by introducing query denoising. IEEE Trans. Pattern Anal. Mach. Intell. **46**(4), 2239–2251 (2024)
16. Li, T., et al.: Pseudo visible feature fine-grained fusion for thermal object detection. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6710–6719 (2025)
17. Liu, K., et al.: ESOD: efficient small object detection on high-resolution images. IEEE Trans. Image Process. **34**, 183–195 (2025)
18. Liu, Q., Liu, R., Zheng, B., Wang, H., Fu, Y.: Infrared small target detection with scale and location sensitivity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17490–17499 (2024)
19. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9992–10002 (2021)
20. Lou, M., Yu, Y.: OverLoCK: an overview-first-look-closely-next convnet with context-mixing dynamic kernels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 128–138 (2025)
21. Meethal, A., Granger, E., Pedersoli, M.: Cascaded zoom-in detector for high resolution aerial images. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2046–2055 (2023)
22. Murrugarra-Llerena, J., Kirsten, L., Jung, C.R.: Can we trust bounding box annotations for object detection? In: Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 4812–4821 (2022)
23. Ni, Z.L., Yang, F., Wen, S., Zhang, G.: Dual relation knowledge distillation for object detection. In: Elkind, E. (ed.) Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI), pp. 1276–1284 (2023)
24. Peng, Y., Li, H., Wu, P., Zhang, Y., Sun, X., Wu, F.: D-FINE: redefine regression task of DETRs as fine-grained distribution refinement. In: Proceedings of the International Conference on Learning Representations (ICLR), pp. 1–13 (2025)
25. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 91–99 (2015)
26. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: FitNets: hints for thin deep nets. In: Proceedings of the International Conference on Learning Representations (ICLR), pp. 1–10 (2015)
27. Sharma, A.: DM-EFS: dynamically multiplexed expanded features set form for robust and efficient small object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 24569–24579 (2025)
28. Shi, Z., et al.: HS-FPN: high frequency and spatial perception FPN for tiny object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 6896–6904 (2025)
29. Tian, Y., Ye, Q., Doermann, D.: YOLOv12: attention-centric real-time object detectors. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 78433–78457 (2025)
30. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 6000–6010 (2017)
31. Wang, J., Chen, Y., Zheng, Z., Li, X., Cheng, M.M., Hou, Q.: CrossKD: cross-head knowledge distillation for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16520–16530 (2024)
32. Wang, K., Du, S., Liu, C., Cao, Z.: Interior attention-aware network for infrared small target detection. IEEE Trans. Geosci. Remote Sens. **60**, 1–13 (2022)
33. Wang, W., et al.: InternImage: exploring large-scale vision foundation models with deformable convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14408–14419 (2023)
34. Wang, X., Fang, H., Li, Q., Wang, L., Chang, Y., Yan, L.: Blur-robust detection via feature restoration: an end-to-end framework for prior-guided infrared UAV target detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 10181–10189 (2026)
35. Wu, X., Hong, D., Chanussot, J.: UIU-Net: U-Net in U-Net for infrared small object detection. IEEE Trans. Image Process. **32**, 364–376 (2023)
36. Yang, B., Zhang, X., Zhang, J., Luo, J., Zhou, M., Pi, Y.: EFLNet: enhancing feature learning network for infrared small target detection. IEEE Trans. Geosci. Remote Sens. **62**, 1–11 (2024)
37. Yang, C., Huang, Z., Wang, N.: QueryDet: cascaded sparse query for accelerating high-resolution small object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13668–13677 (2022)
38. Yang, J., Liu, S., Wu, J., Su, X., Hai, N., Huang, X.: Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 9202–9210 (2025)

39. Yang, X., Yan, J., Liao, W., Yang, X., Tang, J., He, T.: SCRDet++: detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. IEEE Trans. Pattern Anal. Mach. Intell. **45**(2), 2384–2399 (2023)
40. Ying, X., et al.: Mapping degeneration meets label evolution: learning infrared small target detection with single point supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15528–15538 (2023)
41. Yu, J., Jiang, Y., Wang, Z., Cao, Z., Huang, T.: UnitBox: an advanced object detection network. In: Proceedings of the 24th ACM International Conference on Multimedia (ACM MM), pp. 516–520 (2016)
42. Yuan, X., Cheng, G., Yan, K., Zeng, Q., Han, J.: Small object detection via coarse-to-fine proposal generation and imitation learning. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6294–6304 (2023)
43. Zhang, H., et al.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In: Proceedings of the International Conference on Learning Representations (ICLR), pp. 1–12 (2023)
44. Zhang, M., Ouyang, Y., Gao, F., Guo, J., Zhang, Q., Zhang, J.: MOCID: motion context and displacement information learning for moving infrared small target detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 10022–10030 (2025)
45. Zhang, M., Zhang, R., Yang, Y., Bai, H., Zhang, J., Guo, J.: ISNet: shape matters for infrared small target detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 877–886 (2022)
46. Zhang, N., Liu, Y., Liu, H., Tian, T., Ma, J., Tian, J.: DTNet: a specialized dual-tuning network for infrared vehicle detection in aerial images. IEEE Trans. Geosci. Remote Sens. **62**, 1–15 (2024)