Learning with Semantic Priors: Stabilizing Point-Supervised Infrared Small Target Detection via Hierarchical Knowledge Distillation

Yuanhang Yao[1] , **Ping Qian**[1] , **Zhu Liu**[1] , **Long Ma**[1] , **Weimin Wang**[1*]

[1]School of Software Technology, Dalian University of Technology, China
yuanhangyao2027@gmail.com, wangweimin@dlut.edu.cn

Abstract

Single-frame Infrared Small Target Detection (ISTD) aims to localize weak targets under heavy background clutter, yet dense pixel-wise annotations are expensive. Point supervision with online label evolution reduces annotation cost; however, lightweight CNN detectors often lack sufficient semantics, leading to noisy pseudo-masks and unstable optimization. To address this, we propose a hierarchical VFM-driven knowledge distillation framework that uses a frozen Vision Foundation Model (VFM) during training. We formulate point-supervised learning as a bilevel optimization process: the inner loop adapts a VFM-embedded teacher on reweighted training samples, while the outer loop transfers validation-guided knowledge to a lightweight student to mitigate pseudo-label noise and training-set bias. We further introduce Semantic-Conditioned Affine Modulation (SCAM) to inject VFM semantics into CNN features at multiple layers. In addition, a dynamic collaborative learning strategy with cluster-level sample reweighting enhances robustness to imperfect pseudo-masks. Experiments on diverse challenging cases across multiple ISTD backbones demonstrate consistent improvements in detection accuracy and training stability. Our code is available at https://github.com/yuanhang-yao/semantic-prior.

1 Introduction

Single-frame Infrared Small Target Detection (ISTD) is a vital technique in infrared search and tracking systems, owing to its superior capabilities of passive detection and all-weather operation. The goal is to detect faint signals within intricate settings characterized by severe background clutter and real-world degradations [Liu *et al.*, 2025; Wang *et al.*, 2024; Wang *et al.*, 2023], with broad applications in early warning and maritime search and rescue. Despite the success of fully supervised methods, their scalability is severely limited by the prohibitive cost of annotation. Consequently, point supervision, which requires only one coordinate annotation per target, has emerged as a challenging direction.

In the early stage, methods were dominated by handcrafted priors. Representative techniques, such as filter-based methods (e.g., Top-Hat [Bai and Zhou, 2010]) and Human Visual System (HVS)-based methods (e.g., LCM [Chen *et al.*, 2014]), utilize local contrast mechanisms to suppress background clutter. However, these methods rely on the adjustments of hyper-parameters and specific assumptions, limiting generalization in complex scenarios [Liu *et al.*, 2023a]. Recently, with the paradigm shift toward data-driven semantic segmentation, deep learning-based methods [Wang *et al.*, 2025; Wang *et al.*, 2026] have dominated the ISTD field. To preserve the signatures of tiny targets from corrupted features, diverse mechanisms are proposed: multi-scale feature fusion (e.g., UIUNet [Wu *et al.*, 2023b]) effectively integrates high-level semantics with low-level spatial details; contextual modulation schemes (e.g., ACM [Dai *et al.*, 2021]) utilize local contrast and attention to highlight dim targets against complex backgrounds; and dense nested interactions (e.g., DNANet [Li *et al.*, 2023]) enable repetitive feature enhancement to maintain signal strength. However, fully supervised annotation is labor-intensive, limiting scalability.

To address the annotation bottleneck, point-supervised methods have recently emerged as a promising direction. LESPS [Ying *et al.*, 2023] pioneered this paradigm, introducing a label evolution strategy to iteratively update pseudo-masks from point signals. But the unconstrained evolution in LESPS often suffers from optimization instability. To mitigate this, PAL [Yu *et al.*, 2025] proposed an easy-to-hard active learning curriculum to stabilize the training trajectory. Although these methods are effective, a fundamental issue remains unaddressed: the semantic deficiency in lightweight ISTD networks. In the online label evolution framework, the generation of pseudo-labels heavily depends on the feature extraction capability. However, shallow features lack high-level semantic discrimination, making it difficult to distinguish dim targets from high-frequency noise. This semantic gap leads to the generation of noisy pseudo-labels, causing training oscillation and performance degradation.

The emergence of Vision Foundation Model (VFM) [Radford *et al.*, 2021; Kirillov *et al.*, 2023; Oquab *et al.*, 2023] has catalyzed a paradigm shift in ISTD, formally characterized as a foundation-driven paradigm. Capitalizing on this, SAM-

*Corresponding author.

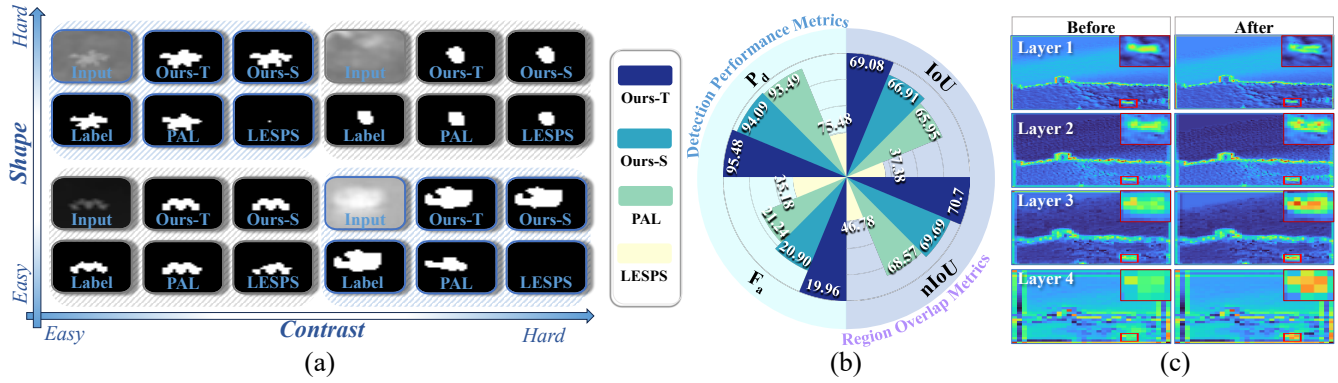

Figure 1: Efficiency overview. (a) Qualitative predictions on four characteristic test subsets (Salient, Filamentary, Faint, Camouflaged), spanning easy-to-hard contrast and shape. (b) Quantitative results on SCTransNet, where we achieve state-of-the-art on both region overlap and detection performance. (c) VFM-guided modulation sharpens target-aware features while suppressing background clutter.

based adaptations like IRSAM [Zhang *et al.*, 2024] exploit tailored prompt encoders and self-prompting mechanisms to transfer open-world segmentation capabilities to the infrared domain. Concurrently, vision-language approaches such as SAIST [Zhang *et al.*, 2025] leverage CLIP to align textual semantics with visual features for enhanced discrimination ability. However, despite their superior representation power, these methods inevitably incur a prohibitive inference burden due to their reliance on heavy, over-parameterized backbones. This heavy-encoder design results in redundant computation and high latency, fundamentally hindering their deployment on resource-constrained edge devices for real-time use.

Driven by these observations, a critical research question arises: How can we leverage the robust semantic priors of VFMs to rectify the feature insufficiency of point supervision, while preserving the inference efficiency? This objective presents a dual challenge. On one hand, the network can effectively distill the rich semantic representation and generalization capability from frozen VFMs to bridge the semantic gap of point-supervised networks, thereby stabilizing the label evolution process. On the other hand, it can achieve a complete decoupling of the heavy parameters from the inference pipeline. Furthermore, direct knowledge distillation straightforwardly minimizes the discrepancy between students and teachers on the training set. However, we argue that this naive alignment creates a generalization bottleneck: the student tends to overfit to the training domain bias.

To address the instability and generalization bottlenecks in online point supervision, we propose a hierarchical VFM-driven knowledge distillation framework. By reformulating the learning paradigm as a hierarchical optimization process, we enable the model to dynamically optimize for representation fitting (lower level) and generalization feedback (upper level). In detail, through explicitly optimizing for performance on an unseen validation distribution in the upper level, we generate VFM-guided feedback that prevents the student network from overfitting to the bias of training set. Furthermore, we design the semantic-conditioned affine modulation to inject robust object-centric semantics from frozen VFMs into the student network, rectifying the semantic deficiency of shallow features. Moreover, to combat overfitting risks, we introduce a dynamic collaborative learning strategy with cluster-level sample reweighting. This mechanism dynamically down-weights outliers while emphasizing informative hard samples, significantly enhancing both the robustness of label evolution and the detection performance in complex scenarios. Our contributions can be summarized as:

- We propose a VFM-driven knowledge distillation framework for point-supervised ISTD by reformulating online point supervision as a bilevel optimization problem. We prioritize validation-based generalization feedback, improving training stability under evolving pseudo-masks.
- To bridge the semantic gap between VFM priors and the fine-grained semantics required by tiny-target localization, we design the semantic-conditioned affine modulation, a lightweight interface that injects VFM semantics into the shallow ISTD network.
- A dynamic collaborative learning strategy is introduced to enhance the robustness of label evolution under noisy pseudo-masks and severe data imbalance. By cluster-level sample reweighting, it automatically down-weights noisy pseudo-labels without manual tuning.
- Experiments on SIRST benchmarks demonstrate consistent improvements across multiple backbones, while keeping the student network efficient at inference.

2 The Proposed Method

2.1 VFM-driven Distillation Framework

As mentioned above, pseudo-masks are inevitably noisy in the early phase of online evolution, and their evolution quality is strongly coupled with the feature extractor. Lightweight CNNs may entangle weak targets with high-frequency clutter, causing unstable mask updates and fragile representations. VFMs possess stronger target-centered semantics and better cross-domain robustness [Radford *et al.*, 2021], making them suitable for addressing the semantic deficiency of point supervision. However, directly deploying VFMs is often impractical due to high inference costs and a mismatch

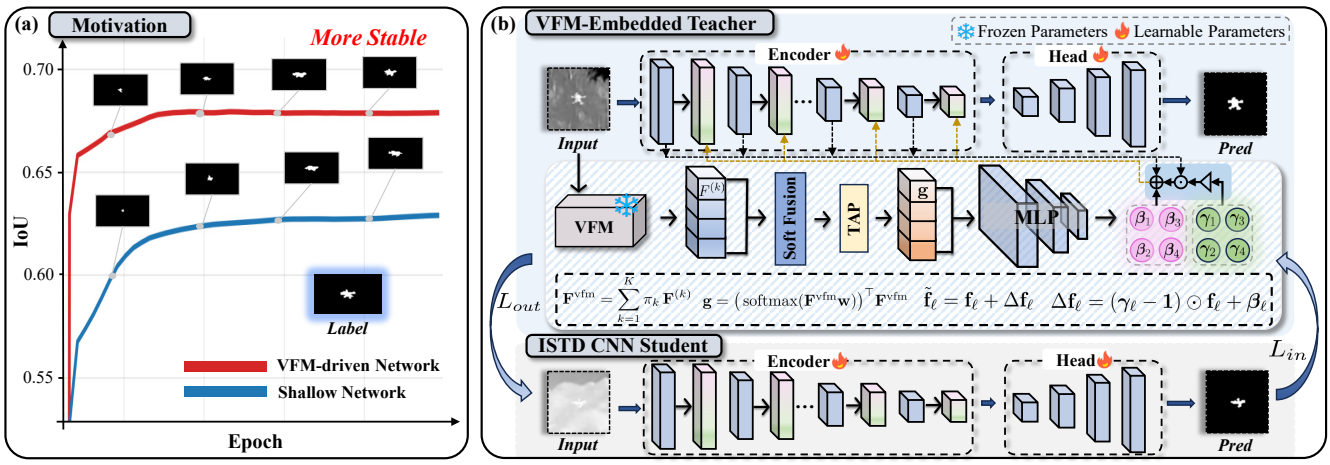

Figure 2: (a) Motivation of introducing VFM semantics to stabilize online label evolution in point-supervised ISTD. (b) The proposed hierarchical VFM-driven distillation framework with a VFM-embedded teacher, SCAM-based modulation, and an ISTD CNN student.

with the semantic granularity of small targets. A naive alternative is to distill students on the training set, but this direct alignment may become a generalization bottleneck: under the induction of evolving pseudo-masks, students tend to imitate the teacher on the training distribution, overfitting to dataset-specific textures instead of inheriting transferable semantics.

Therefore, motivated by recent advances in bilevel optimization [Liu *et al.*, 2023b; Liu *et al.*, 2024c], we propose a hierarchical VFM knowledge distillation framework that uses a frozen VFM only during training and explicitly optimizes for generalization capabilities, which can be formulated as:

$$\begin{aligned} \min_{\theta,\alpha} \quad & \mathbb{E}_{(x,\tilde{y})\sim\mathcal{D}_{\text{val}}}\Big[\, L_{\text{out}}(x,\tilde{y};\theta,\alpha,\phi^\star)\,\Big] \\ \text{s.t.} \quad & \phi^\star = \arg\min_{\phi}\ \mathbb{E}_{(x,\tilde{y})\sim\mathcal{D}_{\text{tr}}}\Big[\, L_{\text{in}}(x,\tilde{y};\theta,\alpha,\phi)\,\Big]. \end{aligned} \tag{1}$$

Specifically, we introduce the base ISTD network, denoted as S_θ, and a VFM-embedded teacher network, denoted as $T_{\theta,\phi}$, where ϕ denotes the teacher-side adaptation parameters, and α denotes the parameters of the sample reweighting function $w_\alpha(\cdot)$. The teacher shares the student weights θ but additionally equips a SCAM modulator (reported in Sec. 2.2) parameterized by ϕ, which injects semantic information into student layers from the VFM. The inner objective L_{in} is to fit a representation on weighted training data and update teacher-side adaptation parameters ϕ, enabling it to generate more reliable soft labels and stabilize label evolution:

$$L_{\text{in}} = w_\alpha(x)\Big(L_{\text{task}}(p_t,\tilde{y}) + \lambda_{\text{in}}\, L_{\text{kd}}(p_t,p_s)\Big) + \lambda_{\text{gate}}\,\mathcal{R}_{\text{gate}}, \tag{2}$$

where $\mathcal{R}_{\text{gate}} = \sum_\ell \|\mathbf{u}_\ell\|_1$ is set to encourage sparsity in the gating variables of semantic modulation $\mathbf{u}_\ell$, so that the teacher selectively injects VFM semantics only when beneficial, yielding a more compressible guidance signal for the CNN student. Here, p_t and p_s denote the teacher and student predictions, respectively. We set $L_{\text{task}}(p,\tilde{y}) = \text{BCE}(p,\tilde{y})$. For knowledge distillation, we use the temperature-scaled sigmoid distributions $q_s = \sigma(z_s/\tau)$ and $q_t = \sigma(z_t/\tau)$ with $\tau = 4$, and define L_{kd} as the mean-squared error between q_s and q_t. λ_{in} controls the inner KD strength and λ_{gate} weights the sparsity regularizer $\mathcal{R}_{\text{gate}}$. We set λ_{in} to 0.1 and λ_{gate} to 5×10^{-3}. The outer objective L_{out} conducts generalization-driven distillation on validation distributions, jointly optimizing the student and sample weights, so that knowledge transfer is more inclined to improve validation performance rather than merely matching the training set:

$$L_{\text{out}} = w_\alpha(x)\Big(L_{\text{task}}(p_s,\tilde{y}) + \lambda_{\text{out}}\, L_{\text{kd}}(p_s,p_t)\Big), \tag{3}$$

where λ_{out} controls the outer KD strength and is set to 1.0 in all experiments. We detach the student prediction in the inner-level KD and detach the teacher prediction in the outer-level KD, since the teacher and student share θ.

2.2 Semantic-Conditioned Affine Modulation

The practical bottleneck of introducing a VFM into point-supervised ISTD lies in the semantic granularity mismatch: Transformer token features have strong semantics but coarse spatial resolution, while infrared small targets rely on fragile local cues and multi-scale CNN priors. Existing fusion methods are either very heavy (e.g., cross-attention) or overly coarse (e.g., concatenation / global pooling), which can easily lead to blurring of small targets, unstable gradients under noisy pseudo-masks, and significantly increase training and inference overhead. Therefore, we aim to find a training-phase mechanism to inject VFM semantics into CNN detectors in a lightweight and controllable manner. We propose Semantic-Conditioned Affine Modulation (SCAM), which converts the frozen VFM features into multi-layer affine parameters for modulating CNN backbone features. For an input image $x \in \mathbb{R}^{H\times W}$, the frozen VFM outputs token embeddings. We denote the token features from the k-th selected transformer block as $\mathbf{F}^{(k)}(x) \in \mathbb{R}^{N\times d}$, where N is the number of tokens, and d is the token embedding dimension. We then learn a soft fusion across depths from K blocks:

$$\mathbf{F}^{\text{vfm}} = \sum_{k=1}^{K} \pi_k\, \mathbf{F}^{(k)}, \tag{4}$$

where $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^{K}$ are learnable weights with $\sum \pi_k = 1$, and K is set to 12 to maximize the number of selected layers.

Inspired by DynamicViT [Rao *et al.*, 2021], we introduce Token-aware Attention Pooling (TAP) to aggregate $\mathbf{F}^{\mathrm{vfm}}$ into a global semantic vector $\mathbf{g}$. Specifically, given $\mathbf{F}^{\mathrm{vfm}} = [t_1, \ldots, t_N] \in \mathbb{R}^{N\times d}$, TAP assigns an attention weight to each token via a lightweight scoring head:

$$a_j = \frac{\exp(\mathbf{w}^\top t_j)}{\sum_{m=1}^{N} \exp(\mathbf{w}^\top t_m)}, \quad \mathbf{g} = \sum_{j=1}^{N} a_j t_j \in \mathbb{R}^d, \tag{5}$$

where $\mathbf{w}$ is a learnable parameter vector and a_j are normalized attention scores. Unlike global average pooling, this design better preserves tokens related to the target, facilitating semantic stability for small targets in complex backgrounds.

Subsequently, SCAM maps $\mathbf{g}$ to the affine parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ of each layer in the CNN. For the feature $\mathbf{f}_\ell$ of the layer ℓ, a compact MLP generates $(\boldsymbol{\gamma}_\ell, \boldsymbol{\beta}_\ell)$ with inputs $\mathbf{g}$, where $\boldsymbol{\gamma}_\ell, \boldsymbol{\beta}_\ell \in \mathbb{R}^{C_\ell}$ represent the channel scale and bias. We introduce a channel gating vector $\mathbf{u}_\ell$ to filter $\boldsymbol{\gamma}_\ell$, enabling selective semantic injection and providing an interface for subsequent sparse regularization. To ensure stable training under online pseudo-masking, we adopt a residual form of scale transformation: $\boldsymbol{\gamma}_\ell \leftarrow 1 + \tanh(\boldsymbol{\gamma}_\ell \odot \mathbf{u}_\ell)$. Finally, after $\boldsymbol{\gamma}_\ell, \boldsymbol{\beta}_\ell$ are broadcast along spatial dimensions, we modulate the features through an affine transformation in a residual form:

$$\tilde{\mathbf{f}}_\ell = \mathbf{f}_\ell + \Delta\mathbf{f}_\ell, \qquad \Delta\mathbf{f}_\ell = (\boldsymbol{\gamma}_\ell - \mathbf{1}) \odot \mathbf{f}_\ell + \boldsymbol{\beta}_\ell, \tag{6}$$

where $\tilde{\mathbf{f}}_\ell$ are modulated CNN features.

Overall, the design adds a small modulator on top of frozen VFM and performs layer-by-layer affine operations on CNN features. Therefore, it has the following advantages: (i) computational overhead is much lower than that of heavy feature fusion; (ii) token scoring and lightweight modulation bring stability; (iii) it is plug-and-play for different CNN backbones, requiring only the acquisition of layer features and reuse of the original detection head.

2.3 Dynamic Collaborative Learning Strategy

ISTD datasets are extremely imbalanced in terms of target size, shape, SNR, and background complexity. If uniform sampling and loss averaging are adopted, the model is prone to overfitting simple samples and reducing robustness. Furthermore, dividing the training into a staged process (first evolving pseudo-labels, then distilling) fails to utilize validation feedback to correct which samples are more suitable for semantic transfer. We propose a dynamic collaborative learning strategy, which couples (i) cluster-level reweighting with (ii) unified bilevel optimization [Liu *et al.*, 2026; Liu *et al.*, 2024b] with Gauss–Newton hypergradient. Specifically, we compute hand-crafted priors (e.g., intensity statistics, texture complexity, spectral sharpness) for each image and cluster the training and validation samples. Each cluster c has a learnable logit α_c, and the weight for each sample i is obtained through softmax within the batch:

$$w_\alpha(x_i) = \frac{\exp(\alpha_{c(i)})}{\frac{1}{|\mathcal{B}|}\sum_{k\in\mathcal{B}} \exp(\alpha_{c(k)})}, \tag{7}$$

Algorithm 1 Dynamic Collaborative Learning Strategy

Input:
Training set $\mathcal{D}_{\mathrm{tr}}$, validation set $\mathcal{D}_{\mathrm{val}}$, frozen VFM;
student parameters θ, SCAM parameters ϕ, cluster logits α
Initialize: pseudo-masks $\tilde{y}$ from point annotations
for each epoch **do**
  Update pseudo-masks $\tilde{y}$ using teacher prediction p_t
  Sample mini-batch $\mathcal{B}_{\mathrm{tr}} \sim \mathcal{D}_{\mathrm{tr}}$ and compute weights w_α
  Inner step: $\phi \leftarrow \phi - \eta\nabla_\phi L_{\mathrm{in}}(\mathcal{B}_{\mathrm{tr}})$
  Sample mini-batch $\mathcal{B}_{\mathrm{val}} \sim \mathcal{D}_{\mathrm{val}}$ and compute weights w_α
  Outer step: update θ, α by $\nabla_{\theta,\alpha} L_{\mathrm{out}}(\mathcal{B}_{\mathrm{val}})$ using Eq. (8)
end for

where $\mathcal{B}$ denotes the current mini-batch and $c(i)$ is the cluster index of sample x_i. Cluster-level learning avoids overfitting to a small number of simple samples and promotes balanced improvement across different levels of scenarios.

To enhance the efficiency of bilevel learning, we adopt a Gauss–Newton style hypergradient approximation based on gradient alignment. Specifically, the hypergradient for α is corrected by an implicit term induced by the inner update:

$$\begin{aligned} \tilde{\nabla}_\alpha L_{\mathrm{out}} &\approx \nabla_\alpha L_{\mathrm{out}} + \eta\,\rho\,\nabla_\alpha L_{\mathrm{in}}, \\ \text{s.t.} \quad \rho &= \frac{\langle \nabla_\theta L_{\mathrm{out}},\, \nabla_\theta L_{\mathrm{in}} \rangle}{\|\nabla_\theta L_{\mathrm{in}}\|_2^2 + \epsilon}, \end{aligned} \tag{8}$$

where η is the inner-step learning rate, $\langle\cdot,\cdot\rangle$ denotes the inner product, and ϵ is a small constant for numerical stability. This Gauss–Newton correction avoids explicit Hessian inversion and maintains stability as the pseudo-mask continuously evolves. The whole algorithm is summarized in Alg. 1.

3 Experiments

3.1 Implementation Configurations

Datasets. The comprehensive dataset SIRST3 [Ying *et al.*, 2023] consists of three sub-datasets SIRST-v1 [Dai *et al.*, 2021], NUDT-SIRST [Li *et al.*, 2023], and IRSTD-1k [Zhang *et al.*, 2022]. Following the LESPS split protocol, all models are trained on SIRST3, with 10% of the samples held out for validation (1402/274/1079 for train/val/test).

We further report results on a test split into Salient, Filamentary, Faint, and Camouflaged, defined by target shape difficulty and target-to-background contrast. Specifically, Salient indicates regular and high-contrast targets, Filamentary indicates morphology-challenging but high-contrast targets, Faint indicates low-contrast but less complex targets, and Camouflaged represents the hardest cases with both challenging morphology and low contrast. This split enables a more interpretable, challenge-aware comparison.

Experimental settings. Unless otherwise specified, we train for 300 epochs with AdamW (batch size 16). The base learning rate is 1×10^{-3} for both inner and outer updates; for backbones that are less stable under weak supervision, we use 5×10^{-4}. We adopt DINOv3 [Siméoni *et al.*, 2025] ViT-S+/16 pretrained on LVD-1689M and keep the VFM backbone frozen throughout. We use PAL [Yu *et al.*, 2025] as the default pseudo-mask evolution engine, while LESPS [Ying *et al.*, 2023] is reported as a baseline. Bilevel updates are

Scheme	Desc.	Overall				Salient				Filamentary				Faint				Camouflaged			
		IoU↑	nIoU↑	P_d↑	F_a↓	IoU↑	nIoU↑	P_d↑	F_a↓	IoU↑	nIoU↑	P_d↑	F_a↓	IoU↑	nIoU↑	P_d↑	F_a↓	IoU↑	nIoU↑	P_d↑	F_a↓
ALCLNet [Yu *et al.*, 2022b]	Full	79.20	82.01	96.81	15.23	80.28	83.93	99.63	14.88	78.13	80.87	98.25	14.81	69.83	75.36	95.54	22.05	82.50	83.13	95.17	13.94
	LESPS	29.12	40.52	68.57	34.33	39.75	57.83	83.21	31.45	31.23	41.34	74.75	35.44	25.32	43.44	66.07	38.86	26.07	29.64	60.14	36.94
	PAL	51.11	59.49	91.43	30.18	56.22	69.46	98.13	33.10	52.30	59.32	94.25	33.22	43.21	61.54	86.61	29.17	49.91	53.81	87.86	33.04
	Ours-T	**54.99**	**61.26**	**93.49**	**27.48**	**61.61**	**70.57**	**98.88**	**25.25**	**55.89**	**60.73**	**96.25**	**27.98**	**47.03**	**62.31**	**91.07**	**23.67**	**53.52**	**55.48**	**90.34**	**25.93**
	Ours-S	**53.00**	**60.24**	**91.50**	**27.60**	**59.62**	**70.25**	**98.88**	**25.16**	**53.14**	**59.58**	**94.50**	**25.98**	**45.42**	**61.65**	**87.50**	**23.85**	**51.89**	**54.63**	**88.55**	**31.90**
GGLNet [Zhao *et al.*, 2023]	Full	80.74	83.63	96.15	11.03	83.22	85.49	100.00	6.69	77.24	81.74	96.50	16.26	74.08	79.10	93.75	9.20	83.16	84.85	94.90	9.71
	LESPS	38.98	49.09	73.02	32.71	49.94	63.90	86.57	29.88	39.42	48.93	77.00	31.35	35.23	54.16	75.00	33.20	36.04	38.57	65.52	35.76
	PAL	58.55	65.62	93.16	27.85	62.49	72.04	98.51	25.15	58.02	64.95	95.25	24.56	47.40	64.17	89.29	29.62	56.32	59.54	88.28	28.25
	Ours-T	**59.14**	**66.03**	**93.69**	**23.68**	**64.07**	**74.34**	**99.25**	**24.29**	**59.76**	**66.09**	**96.25**	**18.14**	**48.45**	**64.60**	**91.07**	**28.22**	**58.77**	**61.55**	**90.90**	**27.02**
	Ours-S	**58.90**	**65.69**	**93.23**	**24.86**	**63.55**	**73.34**	**98.88**	**24.06**	**59.58**	**65.86**	**96.00**	**19.81**	**47.95**	**65.45**	**91.96**	**29.55**	**57.37**	**60.77**	**89.79**	**27.51**
MLCLNet [Yu *et al.*, 2022a]	Full	79.13	82.76	95.88	23.37	79.16	84.96	99.63	20.21	77.33	82.49	97.00	21.26	69.74	73.30	90.18	17.70	80.37	84.66	94.76	18.34
	LESPS	36.54	48.30	86.91	34.18	49.96	65.47	95.52	31.57	37.76	47.51	91.25	32.40	32.77	52.91	80.36	39.29	32.31	36.30	82.34	39.10
	PAL	60.20	66.86	93.29	28.75	63.75	73.80	98.13	28.90	59.02	65.89	95.25	25.23	48.98	65.16	88.96	31.36	60.43	61.31	89.45	35.33
	Ours-T	**65.10**	**69.68**	**94.68**	**27.49**	**69.80**	**75.78**	**98.51**	**28.68**	**63.76**	**67.91**	**96.50**	**24.00**	**56.72**	**69.83**	**92.86**	**17.02**	**65.72**	**66.93**	**92.14**	**23.65**
	Ours-S	**62.63**	**67.28**	**94.49**	**27.22**	**68.33**	**74.25**	**99.63**	**27.93**	**61.75**	**65.97**	**95.75**	**24.08**	**51.54**	**65.54**	**89.61**	**23.60**	**61.95**	**63.94**	**90.34**	**24.41**
MSHNet [Liu *et al.*, 2024a]	Full	80.80	82.67	96.54	11.73	81.27	84.76	99.63	5.34	80.60	82.04	97.75	12.73	70.50	75.64	89.29	6.77	81.86	83.23	95.59	5.63
	LESPS	39.62	51.26	74.62	36.18	50.33	68.06	89.18	34.61	40.79	50.06	79.00	37.14	36.38	58.04	75.89	35.98	36.20	39.25	66.90	35.94
	PAL	61.66	65.26	91.16	22.29	71.00	74.04	99.25	17.80	60.76	64.56	95.00	20.36	55.97	65.88	91.07	22.14	60.07	60.84	89.45	24.86
	Ours-T	**62.24**	**66.50**	**94.02**	**18.00**	**72.73**	**76.04**	**99.63**	**8.49**	**61.27**	**66.31**	**96.50**	**17.97**	**58.07**	**67.76**	**93.75**	**20.48**	**61.79**	**62.57**	**90.21**	**23.62**
	Ours-S	**62.36**	**66.26**	**93.36**	**20.52**	**72.43**	**75.31**	**99.63**	**15.34**	**61.44**	**65.85**	**95.50**	**14.56**	**59.76**	**67.79**	**91.96**	**21.31**	**60.46**	**61.44**	**89.52**	**24.44**
RepISDNet [Wu *et al.*, 2023a]	Full	78.34	82.18	96.08	21.95	77.76	83.47	100.00	15.06	76.17	81.05	97.50	29.51	68.87	76.49	93.75	17.51	81.31	83.59	94.21	27.96
	LESPS	35.85	47.63	72.29	40.76	47.44	64.34	88.43	37.83	37.13	46.89	76.75	32.53	33.77	55.25	73.21	35.89	31.85	35.62	63.72	42.60
	PAL	58.27	62.28	92.77	28.62	63.09	72.22	99.25	25.76	60.01	65.39	96.00	28.51	43.86	60.69	84.82	28.38	56.42	57.58	86.48	25.06
	Ours-T	**64.19**	**69.00**	**95.48**	**24.86**	**68.21**	**74.96**	**99.63**	**18.60**	**63.74**	**67.57**	**97.00**	**25.63**	**53.49**	**66.21**	**87.50**	**17.89**	**63.07**	**63.69**	**91.86**	**15.92**
	Ours-S	**59.47**	**64.56**	**92.96**	**26.99**	**63.88**	**73.90**	**99.63**	**20.51**	**60.31**	**65.40**	**96.50**	**24.67**	**44.25**	**61.30**	**87.50**	**22.06**	**57.59**	**58.82**	**89.52**	**24.62**
DNANet [Li *et al.*, 2023]	Full	82.30	85.38	96.08	7.67	84.66	87.10	99.63	4.96	78.29	84.27	97.75	9.62	74.40	78.43	91.96	18.49	85.28	86.83	94.48	5.45
	LESPS	38.50	49.41	71.96	29.07	51.82	67.29	88.43	32.65	38.10	47.07	75.75	29.03	32.49	54.44	72.32	34.76	35.94	38.37	63.72	35.95
	PAL	62.67	68.22	92.89	17.59	66.37	75.05	98.88	21.75	62.06	68.34	96.25	18.55	49.74	66.67	90.18	21.71	62.71	63.84	88.55	25.39
	Ours-T	**63.09**	**69.42**	**95.15**	**16.18**	**70.91**	**75.56**	**99.63**	**12.26**	**63.74**	**68.67**	**96.50**	**17.85**	**51.18**	**68.32**	**92.86**	**20.09**	**63.08**	**66.39**	**93.79**	**22.28**
	Ours-S	**62.81**	**68.54**	**94.62**	**16.68**	**69.15**	**75.51**	**99.63**	**18.59**	**63.25**	**68.50**	**97.50**	**16.46**	**50.36**	**67.20**	**93.75**	**21.35**	**62.77**	**65.19**	**91.31**	**21.56**
SCTransNet [Yuan *et al.*, 2024]	Full	81.96	84.97	97.21	20.59	83.97	84.64	100.00	13.30	80.43	83.12	98.00	14.20	75.01	81.25	94.64	19.80	84.25	85.33	96.14	13.21
	LESPS	37.38	46.78	75.48	35.18	49.64	63.36	85.45	34.91	39.25	46.56	78.00	30.84	35.37	53.05	75.89	32.73	32.96	34.86	70.34	34.07
	PAL	65.95	68.57	93.49	21.24	72.07	73.09	98.88	25.37	66.81	68.67	97.00	23.06	57.25	68.31	90.18	20.45	64.27	64.88	89.66	24.54
	Ours-T	**69.08**	**70.07**	**95.48**	**19.96**	**75.58**	**75.50**	**99.63**	**18.99**	**68.93**	**69.34**	**97.75**	**22.89**	**65.94**	**69.32**	**93.75**	**19.09**	**67.77**	**67.33**	**93.24**	**18.73**
	Ours-S	**66.91**	**69.69**	**94.09**	**20.90**	**74.72**	**74.72**	**100.00**	**21.29**	**67.01**	**69.30**	**97.25**	**19.90**	**61.32**	**68.82**	**91.96**	**20.11**	**65.92**	**64.93**	**90.34**	**23.78**

Table 1: Performance comparison under different target characteristics. We report IoU (%), nIoU (%), P_d (%), and F_a (10^{-6}).

performed every 5 epochs: we sample mini-batches from the training and validation sets and run 4 Gauss–Newton hypergradient steps to update the ISTD CNN, SCAM parameters, and cluster weights α. We report IoU and nIoU for mask quality, P_d for detection probability (recall), and F_a for false alarm rate, following standard ISTD evaluation.

3.2 Experimental Results

Quantitative results. Table 1 reports results on SIRST3 over the overall test set and four characteristic partitions, including Salient, Filamentary, Faint, and Camouflaged, across seven backbones. Compared with single-point baselines such as LESPS and PAL, our method achieves consistent improvements across metrics, while the deployable student (Ours-S) remains close to the VFM-embedded teacher (Ours-T). PAL is a stronger baseline than LESPS but still suffers from missed targets and higher false alarms, especially on harder subsets, whereas our gains are most evident on Faint and Camouflaged samples, indicating improved robustness under distribution shift. For example, on DNANet, Ours-T improves overall IoU from 62.67 to 63.09 and reduces F_a from 17.59 to 16.18; on SCTransNet, Ours-T further boosts IoU from 65.95 to 69.08 with lower F_a, and achieves a large gain on Faint targets. Consistent improvements are observed across additional backbones, including ALCLNet, GGLNet, and MLCLNet, where both Ours-T and Ours-S improve IoU/nIoU and P_d while keeping F_a competitive. Overall, these results validate that our bilevel, generalization-driven knowledge distillation effectively transfers VFM priors into a lightweight CNN student for robust detection across diverse target characteristics.

Qualitative results. Fig. 3 presents qualitative comparisons produced by different methods under two backbones, DNANet and SCTransNet. Our approach yields consistently more complete and compact target responses with clearer boundaries in both the teacher and the student. Notably, even under extremely low contrast, tiny target size, or heavy background clutter, Ours-S remains capable of activating the true target region while effectively suppressing spurious background responses. In contrast, LESPS and PAL frequently suffer from missed detections, which are marked as Missed Target, or fragmented/shifted predictions across multiple cases. Despite discarding the VFM branch at inference, Ours-S remains close to Ours-T, indicating that our bilevel knowledge distillation effectively transfers VFM generalizable semantics into the lightweight student for robust ISTD.

3.3 Ablation Studies

Training Stability Analysis. ISTD is challenging due to extremely small targets and scarce annotations. Under point-supervised label evolution, the pseudo-mask $\tilde{y}$ is noisy and constantly changing, which often amplifies training instability and overfitting. Our VFM-embedded teacher injects robust global semantics via SCAM, while the bilevel optimization leans more towards transferable cues. Fig. 4 shows higher final IoU on both train and test, with smoother optimization and reduced overfitting. Overall, these observations

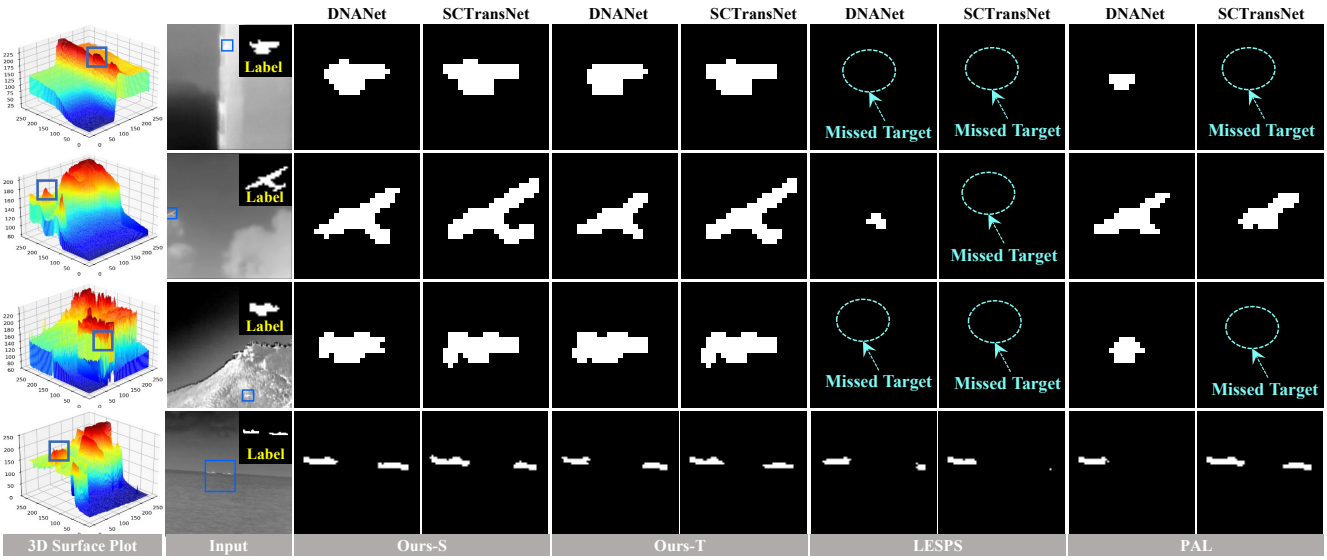

Figure 3: Qualitative comparison on challenging infrared small-target scenarios. We visualize predicted masks from our student (Ours-S) and VFM-embedded teacher (Ours-T) instantiated on DNANet / SCTransNet, and compare them with LESPS and PAL.

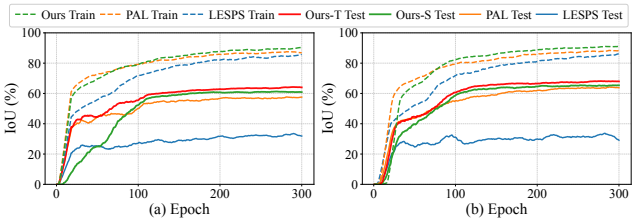

Figure 4: Training stability analysis. IoU curves on the training and test sets for MLCLNet (a) and SCTransNet (b).

Backbone	Params (M)	Student		Teacher		Rel. Overhead	
		Lat.↓	Mem↓	Lat.↓	Mem↓	Lat. (×) ↓	Mem (×) ↓
MLCLNet	0.66	34.54	939.6	54.22	1051.8	1.57	1.12
DNANet	4.70	529.25	1263.0	548.39	1387.3	1.04	1.10
ALCLNet	5.67	30.47	555.5	49.63	681.4	1.63	1.23
SCTransNet	11.33	223.40	1314.1	265.33	1742.6	1.19	1.33

Table 2: Inference overhead comparison across four backbones. We report the mean latency (ms) and peak GPU memory (MB).

validate our design goals: enhancing VFM-driven optimization, stabilizing training under incomplete supervision conditions, and distilling this stability into an ISTD CNN student.

Impact of VFM on CNN Representations. To visually demonstrate how VFM priors reshape ISTD detectors, we visualize the intermediate representations of the CNN encoder before and after SCAM modulation. Specifically, for three encoder layers, we calculate the channel mean intensity maps of the original features $\mathbf{f}_\ell$ and modulated features $\tilde{\mathbf{f}}_\ell$. Additionally, we also visualize the VFM patch grid score map generated by token-aware aggregation, which directly reflects the attention distribution of DINO patch tokens. Fig. 5 shows SCAM boosts target responses and suppresses clutter across layers, producing clearer spatial contrast. The VFM score maps often focus on subtle target cues that are easily overlooked by purely local CNN priors (e.g., faint or irregular targets). Therefore, VFM can provide complementary global guidance. This effect ultimately translates into more accurate and compact predicted shapes in difficult samples and improves overall detection performance.

Computational Overhead. We benchmark our framework on four backbones using a single NVIDIA A40 GPU with AMP enabled. We perform 20 warm-up iterations and report averaged latency over 100 measured iterations, repeated 3 times. The student is unchanged for inference, while the teacher is equipped with a frozen DINOv3 ViT-S+/16 backbone with about 29M parameters. As summarized in Table 2, the teacher introduces a moderate additional cost with a latency increase ranging from $1.04\times$ to $1.63\times$ and a peak memory increase from $1.10\times$ to $1.33\times$, depending on the backbone. For SCTransNet, overhead adds $1.19\times$ latency and $1.33\times$ memory. On lightweight backbones (e.g., MLCLNet / ALCLNet), latency overhead becomes more visible but memory remains efficient. Importantly, since all teacher components are removed for deployment, our improvements come from the proposed training strategy without increasing inference-time complexity of the student model.

Efficiency of Proposed Mechanisms. Despite adding a frozen VFM branch and hierarchical training, the gains come from a few lightweight designs. Table 3 shows that removing VFM and SCAM decreases IoU to 63.06, suggesting pure CNNs lack semantic discrimination with evolving pseudo-masks. The two knowledge distillation terms are complementary: removing either lowers IoU and raises F_a. The inner term $L_{\mathrm{kd}}(p_t, p_s)$ stabilizes teacher adaptation, while the outer term $L_{\mathrm{kd}}(p_s, p_t)$ distills validation semantics from $\mathcal{D}_{\mathrm{val}}$ into the student. Removing the validation set causes one

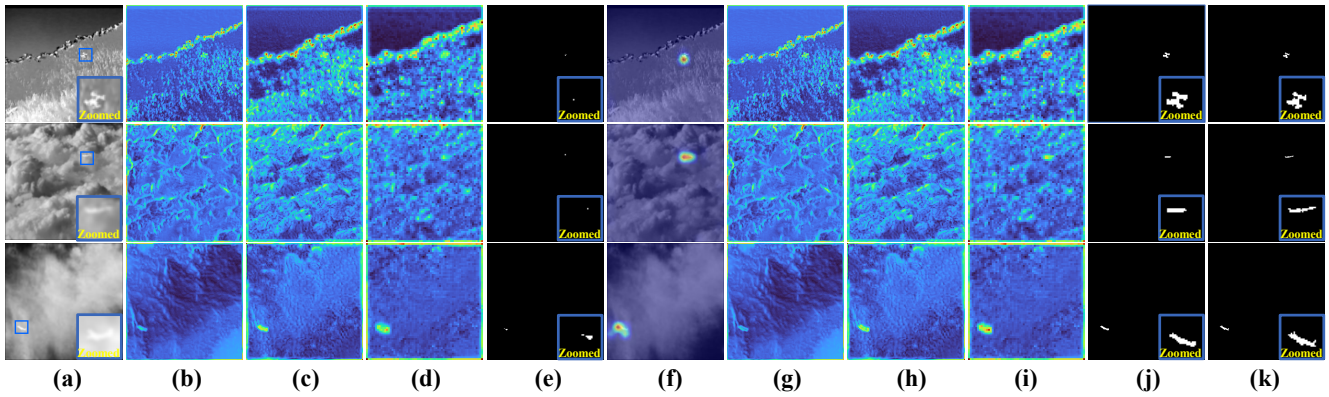

Figure 5: Impact of VFM semantics on CNN representations. (a) Input infrared image. (b–d) Encoder features from three layers before SCAM modulation. (e) Prediction of the unmodulated network. (f) DINO attention (score) map indicating semantic regions of interest. (g–i) Corresponding layer features after SCAM modulation. (j) Prediction of the modulated network. (k) Ground-truth label.

Method	VFM	SCAM	Outer-KD	Inner-KD	Val Set	Reweight	IoU↑	nIoU↑	P_d↑	F_a↓
w/o VFM	✗	✗	✗	✗	✗	✗	63.06	64.77	93.49	29.27
w/o $L_{\mathrm{kd}}(p_s, p_t)$	✓	✓	✗	✓	✓	✓	64.35	68.75	93.09	32.80
w/o $L_{\mathrm{kd}}(p_t, p_s)$	✓	✓	✓	✗	✓	✓	64.45	67.38	93.42	26.99
w/o $\mathcal{D}_{\mathrm{val}}$	✓	✓	✓	✓	✗	✓	63.33	66.07	**95.22**	22.43
$w_\alpha(x) \equiv 1$	✓	✓	✓	✓	✓	✗	65.22	67.31	91.30	23.72
Proposed	✓	✓	✓	✓	✓	✓	**66.91**	**69.69**	94.09	**20.90**

Table 3: Ablation study of the proposed core components.

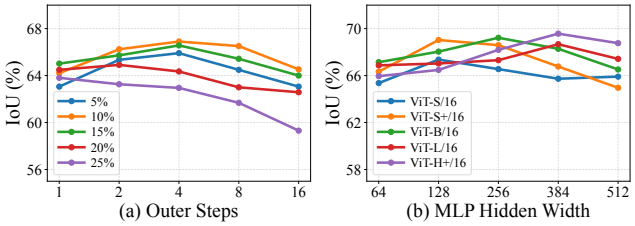

Figure 6: (a) IoU versus the number of outer optimization steps under different validation ratios (Student). (b) IoU versus the hidden width of the SCAM MLP for different DINOv3 variants (Teacher).

Method	Teacher				Student			
	IoU↑	nIoU↑	P_d↑	F_a↓	IoU↑	nIoU↑	P_d↑	F_a↓
Naive Concat Fusion	64.99	69.58	94.95	25.32	64.09	69.17	93.36	28.00
w/o TAP (GAP instead)	63.31	62.43	94.15	30.43	63.09	63.60	94.15	33.95
Single-depth VFM Feature	64.40	68.37	93.89	32.41	63.27	66.42	94.42	27.72
Single-layer Modulation	66.43	69.77	**95.55**	27.72	64.76	69.49	**94.55**	23.03
w/o $\mathcal{R}_{\mathrm{gate}}$	68.36	69.90	94.62	**18.54**	66.18	69.23	93.69	**17.85**
Proposed	**69.08**	**70.07**	95.48	19.96	**66.91**	**69.69**	94.09	20.90

Table 4: Ablation study of SCAM.

Subset	H_{norm} ↓		EffN↓		$p_{\max}$ ↑	
	Mean	Std	Mean	Std	Mean	Std
Salient	81.73	9.91	58.17	34.55	9.96	8.65
Filamentary	79.83	11.60	48.85	31.97	11.80	11.34
Faint	81.67	9.74	59.20	33.75	10.43	9.15
Camouflaged	**69.36**	12.42	**32.05**	24.29	**17.59**	15.60

Table 5: Characteristic-wise attention analysis of DINO modulator. H_{norm} (%) denotes the normalized attention entropy, EffN denotes the effective number of attended patches, and $p_{\max}$ (%) denotes the maximum attention weight.

of the largest performance drops, showing validation feedback counters cross-scene bias; while disabling cluster-level reweighting reduces P_d and increases F_a, indicating it mitigates imbalance. Fig. 6 (a) shows performance saturates with 10% $\mathcal{D}_{\mathrm{val}}$ and 4 outer steps; Fig.6 (b) indicates that the hidden layer width of the MLP in SCAM depends on the DINOv3 variant due to different semantic dimensions, so the overhead is controllable without sensitive tuning.

SCAM Design. In Table 4, we compared six VFM-CNN fusion strategies to better integrate VFM priors with ISTD CNN detectors. Replacing TAP with Global Average Pooling (GAP) or using a single VFM depth degrades performance; GAP increases F_a to 33.95. This suggests that selective token aggregation and multi-depth fusion are crucial in cluttered infrared backgrounds. For semantic injection, naive concatenation or single-layer modulation is inferior, while multi-layer affine modulation brings stronger and more stable gains, better aligning global semantics with fine-grained localization. Removing the layer-wise gate reduces IoU to 66.18 and P_d to 93.69, implying that gated semantic control is important under evolving pseudo-masks $\tilde{y}$. Table 5 shows that camouflaged targets have lower H_{norm}, smaller EffN, and larger $p_{\max}$, i.e., attention becomes more selective under heavy clutter and irregular shapes. Overall, SCAM balances performance and robustness via token-aware multi-depth aggregation, multi-layer affine modulation, and gated injection, enabling layer-wise guidance for stable bilevel transfer.

4 Conclusion

This paper proposes a bilevel knowledge distillation framework guided by a VFM for point-supervised ISTD. During training, the VFM is frozen and its knowledge is distilled into a lightweight detector. SCAM and dynamic collaborative learning reduce the impact of noisy pseudo-masks and stabilize training. Experiments on SIRST3 show consistent improvements across different backbones and data splits.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Nos. 62306059, 624B2033), the Fundamental Research Funds for the Central Universities.

References

[Bai and Zhou, 2010] Xiangzhi Bai and Fugen Zhou. Analysis of new top-hat transformation and the application for infrared dim small target detection. *Pattern Recognition*, 43(6):2145–2156, 2010.

[Chen *et al.*, 2014] CL Philip Chen, Hong Li, Yantao Wei, Tian Xia, and Yuan Yan Tang. A local contrast method for small infrared target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1):574–581, 2014.

[Dai *et al.*, 2021] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. Asymmetric contextual modulation for infrared small target detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 950–959, 2021.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[Li *et al.*, 2023] Boyang Li, Chao Xiao, Longguang Wang, Yingqian Wang, Zaiping Lin, Miao Li, Wei An, and Yulan Guo. Dense nested attention network for infrared small target detection. *IEEE Transactions on Image Processing*, 32:1745–1758, 2023.

[Liu *et al.*, 2023a] Zhu Liu, Zihang Chen, Jinyuan Liu, Long Ma, Xin Fan, and Risheng Liu. Enhancing infrared small target detection robustness with bi-level adversarial framework. *arXiv preprint arXiv:2309.01099*, 2023.

[Liu *et al.*, 2023b] Zhu Liu, Jinyuan Liu, Guanyao Wu, Long Ma, Xin Fan, and Risheng Liu. Bi-level dynamic learning for jointly multi-modality image fusion and beyond. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 1240–1248, 2023.

[Liu *et al.*, 2024a] Qiankun Liu, Rui Liu, Bolun Zheng, Hongkui Wang, and Ying Fu. Infrared small target detection with scale and location sensitivity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17490–17499, 2024.

[Liu *et al.*, 2024b] Risheng Liu, Zhu Liu, Jinyuan Liu, Xin Fan, and Zhongxuan Luo. A task-guided, implicitly-searched and meta-initialized deep model for image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6594–6609, 2024.

[Liu *et al.*, 2024c] Risheng Liu, Zhu Liu, Wei Yao, Shangzhi Zeng, and Jin Zhang. Moreau envelope for nonconvex bi-level optimization: a single-loop and hessian-free solution strategy. In *Proceedings of the 41st International Conference on Machine Learning*, pages 31566–31596, 2024.

[Liu *et al.*, 2025] Zhu Liu, Zijun Wang, Jinyuan Liu, Fanqi Meng, Long Ma, and Risheng Liu. DEAL: Data-efficient adversarial learning for high-quality infrared imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28198–28207, 2025.

[Liu *et al.*, 2026] Risheng Liu, Zhu Liu, Weihao Mao, Wei Yao, and Jin Zhang. Bilevel optimization for adversarial learning problems: Sharpness, generation, and beyond. *Advances in Neural Information Processing Systems*, 38:29102–29130, 2026.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[Rao *et al.*, 2021] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient vision transformers with dynamic token sparsification. *Advances in Neural Information Processing Systems*, 34:13937–13949, 2021.

[Siméoni *et al.*, 2025] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

[Wang *et al.*, 2023] Weimin Wang, Ting Yang, Yu Du, and Yu Liu. Snow removal for LiDAR point clouds with spatio-temporal conditional random fields. *IEEE Robotics and Automation Letters*, 8(10):6739–6746, 2023.

[Wang *et al.*, 2024] Weimin Wang, Yingxu Deng, Zezeng Li, Yu Liu, and Na Lei. MergeNet: Explicit mesh reconstruction from sparse point clouds via edge prediction. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[Wang *et al.*, 2025] Weimin Wang, Xin Tan, Liang Li, Yu Liu, and Qiong Chang. 3D-NLM: Voxel-based non-local means for 3D point cloud noise detection and smoothing. *Computers & Graphics*, page 104348, 2025.

[Wang *et al.*, 2026] Weimin Wang, Ruifeng Nie, Yingchi Liu, Long Ma, Chengpei Xu, Qi Jia, Yu Liu, and Na Lei. SWG-Fusion: Soft weather-guided multimodal fusion with VLM-assistance for BEV object detection under harsh weather. *Pattern Recognition*, page 113398, 2026.

[Wu *et al.*, 2023a] Shuanglin Wu, Chao Xiao, Longguang Wang, Yingqian Wang, Jungang Yang, and Wei An. RepISD-Net: Learning efficient infrared small-target detection network via structural re-parameterization. *IEEE*

Transactions on Geoscience and Remote Sensing, 61:1–12, 2023.

[Wu *et al.*, 2023b] Xin Wu, Danfeng Hong, and Jocelyn Chanussot. UIU-Net: U-Net in U-Net for infrared small object detection. *IEEE Transactions on Image Processing*, 32:364–376, 2023.

[Ying *et al.*, 2023] Xinyi Ying, Li Liu, Yingqian Wang, Ruojing Li, Nuo Chen, Zaiping Lin, Weidong Sheng, and Shilin Zhou. 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*, pages 15528–15538, 2023.

[Yu *et al.*, 2022a] Chuang Yu, Yunpeng Liu, Shuhang Wu, Zhuhua Hu, Xin Xia, Deyan Lan, and Xin Liu. Infrared small target detection based on multiscale local contrast learning networks. *Infrared Physics & Technology*, 123:104107, 2022.

[Yu *et al.*, 2022b] Chuang Yu, Yunpeng Liu, Shuhang Wu, Xin Xia, Zhuhua Hu, Deyan Lan, and Xin Liu. Pay attention to local contrast learning networks for infrared small target detection. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[Yu *et al.*, 2025] Chuang Yu, Jinmiao Zhao, Yunpeng Liu, Sicheng Zhao, Yimian Dai, and Xiangyu Yue. From easy to hard: Progressive active learning framework for infrared small target detection with single point supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2588–2598, 2025.

[Yuan *et al.*, 2024] Shuai Yuan, Hanlin Qin, Xiang Yan, Naveed Akhtar, and Ajmal Mian. SCTransNet: Spatial-channel cross transformer network for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.

[Zhang *et al.*, 2022] Mingjin Zhang, Rui Zhang, Yuxiang Yang, Haichen Bai, Jing Zhang, and Jie Guo. ISNet: Shape matters for infrared small target detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 877–886, 2022.

[Zhang *et al.*, 2024] Mingjin Zhang, Yuchun Wang, Jie Guo, Yunsong Li, Xinbo Gao, and Jing Zhang. IRSAM: Advancing Segment Anything Model for infrared small target detection. In *Proceedings of the European Conference on Computer Vision*, pages 233–249. Springer, 2024.

[Zhang *et al.*, 2025] Mingjin Zhang, Xiaolong Li, Fei Gao, Jie Guo, Xinbo Gao, and Jing Zhang. 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*, pages 9549–9558, 2025.

[Zhao *et al.*, 2023] Jinmiao Zhao, Chuang Yu, Zelin Shi, Yunpeng Liu, and Yingdi Zhang. Gradient-guided learning network for infrared small target detection. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023.