Rethinking Infrared Small Target Detection: A Foundation-Driven Efficient Paradigm

Chuang Yu[1,2,6] **Jinmiao Zhao**[1,2] **Yunpeng Liu**[1†] **Yaokun Li**[3] **Xiujun Shu**[2]
Yuanhao Feng[4] **Bo Wang**[5] **Xiangyu Yue**[6†]

[1]Shenyang Institute of Automation, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences [3]Sun Yat-sen University [4]Peking University
[5]Northwestern Polytechnical University [6]MMLab, The Chinese University of Hong Kong

Abstract

While large-scale visual foundation models (VFMs) exhibit strong generalization across diverse visual domains, their potential for single-frame infrared small target (SIRST) detection remains largely unexplored. To fill this gap, we first systematically introduce frozen VFM representations into the SIRST task and propose a **F**oundation-**D**riven **E**fficient **P**aradigm (**FDEP**), a general and efficient framework compatible with diverse VFMs and SIRST networks, which improves detection accuracy without additional VFM-related inference overhead. Specifically, a Semantic Alignment and Modulated Fusion (SAMF) module is designed to achieve dynamic alignment and deep fusion of the global semantic priors from VFMs with task-specific features. Meanwhile, to avoid the inference-time overhead introduced by VFMs, we propose a Collaborative Optimization-based Implicit Self-Distillation (CO-ISD) strategy, which enables implicit semantic transfer between the main and lightweight branches through parameter sharing and synchronized backpropagation. In addition, to unify the fragmented evaluation system, we construct a Holistic SIRST Evaluation (HSE) metric that performs multi-threshold integral evaluation at both pixel-level confidence and target-level robustness, providing a stable and comprehensive basis for fair model comparison. Extensive experiments demonstrate that the SIRST detection networks equipped with our FDEP framework achieve state-of-the-art (SOTA) performance on multiple public datasets. Our code is available at **https://github.com/YuChuang1205/FDEP-Framework**.

1 Introduction

Single-frame infrared small target (SIRST) detection is a key technology for infrared search and tracking systems [43, 9, 51, 44] and has been widely applied to target tracking [38, 29, 56], maritime assistance [48, 54, 10], and early warning [19, 35, 23]. However, current SIRST detection remains challenging due to complex backgrounds, scarce target features, and severe class imbalance.

Early SIRST studies focused on model-driven methods [27, 1, 28, 17, 3, 26, 2, 31, 7, 8, 12, 47, 4], which explicitly model infrared imaging and background priors but are often limited by handcrafted assumptions. With the rise of deep learning, data-driven methods have shown superior adaptability and robustness by automatically learning high-dimensional mappings from data [42, 11, 30]. Recently, deep learning-based methods [5, 53, 14, 34, 21, 33, 36, 45, 49, 55, 20, 18, 50, 52, 13, 16, 58, 39, 40] further enhance weak-target perception by incorporating task-specific cues, such as local contrast [39, 40], text [52, 13], and shape priors [49, 55]. Nevertheless, three limitations remain: *1. Insufficient visual global context and semantic modeling.* Existing methods focus on architectural modifications,

† Corresponding author: ypliu@sia.cn and xyyue@ie.cuhk.edu.hk.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

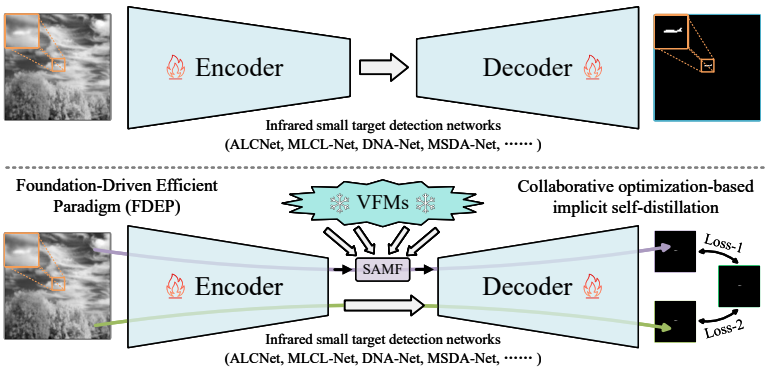

Figure 1: Comparison of different SIRST detection paradigms. **Top:** Traditional encoder-decoder-based architecture performs feature encoding and decoding in a sequential manner. **Bottom:** Our FDEP framework consists of a main branch (purple arrow) and a lightweight branch (green arrow). The main branch integrates frozen semantic representations from VFMs. The lightweight branch maintains the original structure. During inference, only the lightweight branch is used for efficiency.

local clue amplification, and text incorporation, but their ability to characterize visual global structural context and high-level semantic relations is relatively limited, resulting in unstable performance in complex backgrounds. *2. Limited performance-efficiency trade-off.* In pursuit of higher accuracy, some methods introduce substantial parameters and computational overhead, increasing latency and limiting their applicability to real-time or resource-constrained scenarios. *3. Fragmented and insufficient evaluation system.* Existing single metrics are insufficient to comprehensively assess detection capabilities, and the current evaluation system is highly sensitive to threshold settings.

To address these challenges, we rethink the SIRST detection paradigm. As shown in Fig. 1, this work first systematically introduces frozen representations from visual foundation models (VFMs) into SIRST detection, and proposes a Foundation-Driven Efficient Paradigm (FDEP). It provides a general and efficient framework compatible with diverse VFMs and SIRST networks, bridging the modality gap between natural and infrared imagery, and transferring VFM-derived priors to lightweight detectors without additional inference overhead. Specifically, considering the semantic and scale discrepancies between the high-level representations from VFMs and the task-specific features of SIRST, we design a Semantic Alignment and Modulated Fusion (SAMF) module, which achieves deep fusion between global semantic priors and task-specific representations through semantic alignment and dual-path modulation. Meanwhile, to avoid the inference time overhead introduced by VFMs and to mitigate the optimization instability and insufficient knowledge transfer of explicit distillation under extreme class imbalance, we further propose a Collaborative Optimization-based Implicit Self-Distillation (CO-ISD) strategy. By sharing encoder-decoder parameters and performing synchronous backpropagation within a unified optimization space, CO-ISD enables implicit semantic transfer between the main and lightweight branches during training, achieving stable performance gains while preserving the original detector architecture and computational cost at inference. In addition, to overcome the fragmentation and insufficiency of existing evaluation systems, we construct a Holistic SIRST Evaluation (HSE) metric, which performs multi-threshold integral assessment at both pixel-level confidence prediction and target-level robust detection perspectives, providing a more stable and comprehensive metric for fair model comparison and analysis. The main contributions of this work are summarized as follows:

- We systematically introduce frozen VFM representations into SIRST detection and propose FDEP, a general and efficient framework compatible with diverse VFMs and SIRST networks, which improves detection accuracy without additional VFM-related inference overhead.
- To bridge the semantic gap, we propose a SAMF module that achieves deep fusion of global semantic priors and task-specific representations through semantic alignment and dual-path modulation.
- A CO-ISD strategy is proposed to enable implicit semantic transfer through shared encoder-decoder and synchronous backpropagation, improving performance while enabling lightweight deployment.
- To overcome the limitations of existing evaluation systems, we construct the HSE metric that performs multi-threshold integral assessment from both pixel-level and target-level perspectives.

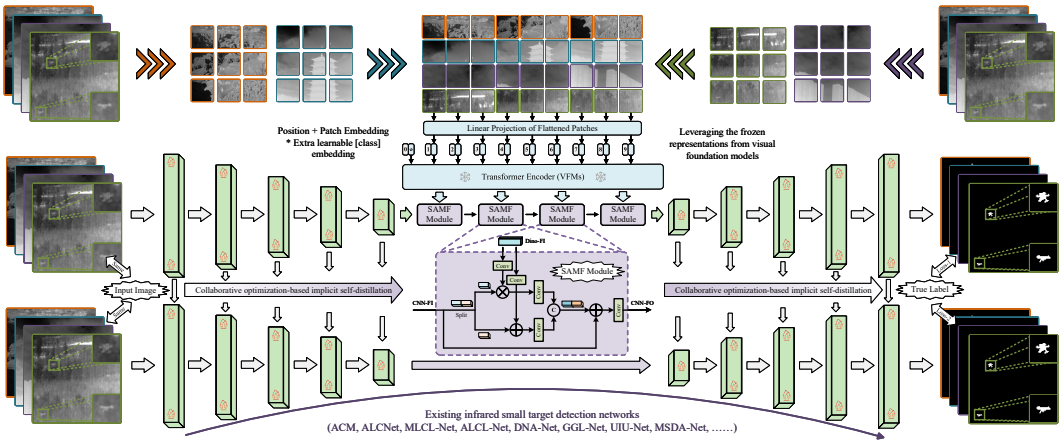

Figure 2: Overview of the FDEP Framework.

2 Related Work

Non-deep learning-based SIRST detection. These methods rely on handcrafted features and priors, including background suppression, human visual system, and image structure-based methods. Background suppression-based methods [27, 1, 28, 17, 3] enhance target responses by reducing background energy, but often degrade in complex scenes. Human visual system-based methods [26, 2, 31, 7] exploit visual attention and contrast mechanisms, but often produce false alarms in noisy backgrounds. Image structure-based methods [8, 12, 47, 4] formulate detection as low-rank sparse decomposition, but are computationally expensive and sensitive to complex textures.

Deep learning-based SIRST detection. Compared with non-deep learning methods, deep learning-based methods show stronger robustness and generalization [57, 15]. Existing methods can be broadly grouped into network structure optimization, contrast modeling, edge/shape enhancement, and text incorporation. Network structure optimization methods mainly build on U-Net variants to improve feature extraction and semantic modeling, including CNN-based [5, 53, 14, 34, 21] and Transformer-hybrid networks [33, 36, 45]. Contrast modeling methods [6, 39, 40] inherit the idea of local contrast to enhance target salience and foreground-background separability. To better capture fine-grained edges, recent works [49, 55, 20, 18, 16, 58] explore edge/shape enhancement, effectively improving boundary integrity through gradient, shape, and directional cues. Furthermore, some works [52, 13] introduce textual information to enhance detection performance. Despite their progress, existing methods focus on individual network optimization, with limited exploration of unified detection paradigms. More importantly, the potential of large-scale VFMs for SIRST detection has not yet been fully explored. Therefore, we systematically introduce VFMs into SIRST detection and propose a general and efficient framework that can be adapted to diverse VFMs and SIRST networks.

3 Methods

3.1 FDEP Framework

To further improve the detection performance of existing SIRST networks, we propose the FDEP framework. As shown in Fig. 2, FDEP consists of a frozen VFM representation branch, a main branch, and a lightweight branch. The frozen VFM branch extracts global semantic priors from the input infrared image, activating the universal visual knowledge learned from large-scale natural images. The main branch uses the encoder of an existing SIRST network to extract task-specific high-level semantic features, and employs the proposed SAMF module to deeply align and fuse them with VFM-derived global semantic priors. The main branch features after the SAMF module not only have stronger discriminative abilities, but also significantly improve robust detection capabilities in complex backgrounds. For the lightweight branch, we retain the structure of the original SIRST detection network to ensure efficient inference. During training optimization, the proposed CO-ISD strategy is used. By sharing the encoder-decoder structure and performing synchronous backpropagation, the main branch dynamically guides the lightweight branch to learn its high-level semantic representation, thereby achieving implicit knowledge transfer and efficient alignment.

3.2 SAMF Module

Although VFMs (such as DINOv2 [24]) can learn powerful global representations from large-scale natural images, their feature distributions remain misaligned with SIRST-specific semantics. To bridge the semantic gap, we design a SAMF module that dynamically aligns and fuses VFM-derived global semantic priors with task-specific discriminative features at the feature level. From Fig. 2, SAMF consists of three steps: semantic alignment, dual-path modulation, and residual refinement. Taking MSDA-Net equipped with the FDEP framework as an example, the process is as follows:

For semantic alignment, given multi-level features $F_{\mathrm{DINO}}^{(i)} \in \mathbb{R}^{C_d \times H_i \times W_i}, i \in \{6, 12, 18, 24\}$ from the frozen DINOv2 (ViT-L/14), we first project them to the spatial scale of the backbone encoder's top-level feature $F^{(t)} \in \mathbb{R}^{64 \times H \times W}$ using a 1×1 convolution followed by bilinear upsampling:

$$\tilde{P}^{(i)} = \mathrm{Up}(\mathrm{Conv}_{1\times1}(F_{\mathrm{DINO}}^{(i)})), \tag{1}$$

where $\tilde{P}^{(i)} \in \mathbb{R}^{32 \times H \times W}$ denotes the aligned semantic prior. Subsequently, two modulation maps are generated through two independent 1×1 convolutional branches:

$$M_{\mathrm{mul}}^{(i)} = \mathrm{Sigmoid}(\mathrm{Conv}_{1\times1}(\tilde{P}^{(i)})), \tag{2}$$

$$M_{\mathrm{add}}^{(i)} = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}_{1\times1}(\tilde{P}^{(i)}))), \tag{3}$$

where M_{mul} is a multiplicative gating map. M_{add} is an additive semantic map.

For dual-path modulation, we divide the backbone features $F^{(t)}$ into two equal parts $[F_A^{(t)}, F_B^{(t)}]$ along the channel dimension. The $F_A^{(t)}$ performs multiplicative modulation to suppress noise responses, while the $F_B^{(t)}$ performs additive modulation to inject semantic information:

$$B_{\mathrm{mul}}^{(i)} = \mathrm{Conv}_{3\times3}(F_A^{(t)} \odot M_{\mathrm{mul}}^{(i)}), \tag{4}$$

$$B_{\mathrm{add}}^{(i)} = \mathrm{Conv}_{1\times1}(F_B^{(t)} + M_{\mathrm{add}}^{(i)}), \tag{5}$$

where $\odot$ denotes element-wise multiplication.

For residual refinement, the dual-path modulated features are channel-wise concatenated, then residually superimposed with the original features, and finally integrated using a 3×3 convolution:

$$F^{(t+1)} = \mathrm{Conv}_{3\times3}(([B_{\mathrm{mul}}^{(i)}, B_{\mathrm{add}}^{(i)}]) + F^{(t)}), \tag{6}$$

where $F^{(t+1)}$ denotes the fused feature after the $(t+1)_{th}$ SAMF module. The top-level features of the backbone encoder are progressively fused with the features of the four DINO deep layers $(i \in \{6, 12, 18, 24\})$ to obtain the final semantically aligned fused features $F^4 \in \mathbb{R}^{64 \times H \times W}$.

Overall, guided by the global semantic priors from frozen VFMs, SAMF aligns and fuses task-specific and global semantic features through a three-stage mechanism. In addition, by progressively injecting multi-level semantic priors and performing decoupled feature fusion, stacked SAMF modules transfer the global prior knowledge to the task-specific detector, yielding semantically consistent, noise-resistant, and target-sensitive representations that significantly improve detection performance.

3.3 CO-ISD Strategy

To avoid the inference time overhead introduced by VFMs, we propose the CO-ISD strategy. From Fig. 2, the CO-ISD strategy consists of a main branch and a lightweight branch that are jointly optimized during training. The main branch integrates frozen VFM representations to learn globally contextual and highly discriminative features, while the lightweight branch retains the original SIRST network structure to enable efficient inference during deployment. Notably, the two branches share the same encoder-decoder parameters and compute their respective detection losses simultaneously during training. This strategy achieves implicit modeling of semantic consistency and knowledge transfer through dynamic collaborative optimization.

Specifically, the optimization objectives for the main and lightweight branches are formulated as:

$$\mathcal{L}_{\mathrm{main}} = \mathcal{L}_{EEDM}(Y_{\mathrm{main}}, Y_{\mathrm{gt}}), \tag{7}$$

$$\mathcal{L}_{light} = \mathcal{L}_{EEDM}(Y_{light}, Y_{\mathrm{gt}}), \tag{8}$$

Algorithm 1 Computation of Holistic SIRST Evaluation (*HSE*)

Input: Predicted probability maps $\{P_i\}_{i=1}^{N}$; Ground truth masks $\{G_i\}_{i=1}^{N}$; Threshold set $\mathcal{T} = \{t_j\}_{j=1}^{M}$; Centroid matching tolerance τ.
Output: Pixel-level *HSE-P*, Target-level *HSE-T*, Holistic *HSE*.

Initialize: Give initial value to M, τ, and Precision–Recall lists $\mathcal{P} \leftarrow \emptyset$, $\mathcal{R} \leftarrow \emptyset$
Step HSE-P-1: Flatten all $\{P_i\}_{i=1}^{N}$ and corresponding $\{G_i\}_{i=1}^{N}$:

$$\{(p_k, g_k)\}_{k=1}^{K} \leftarrow \text{Flatten}(\{(P_i, G_i)\})$$

Step HSE-P-2: Compute the Precision-Recall curve and integrate it to obtain the *HSE-P* according to Eq.(11).
Step HSE-T-1: Iterate over thresholds to compute successive $(\mathcal{P}, \mathcal{R})$ pairs:
for each threshold $t \in \mathcal{T}$ **do**
    $\widehat{P}_i(t) \leftarrow \mathbb{I}(P_i > t)$ ▷ Perform binarization to generate the binary masks
    $(\widehat{\mathcal{C}}_{\text{p}}(t), \mathcal{C}_{\text{g}}) \leftarrow \text{ConnA}(\widehat{P}_i(t), G_i)$ ▷ Obtain connected areas from the masks
    $\mathcal{M}(t) \leftarrow \text{Match}(\widehat{\mathcal{C}}_{\text{p}}(t), \mathcal{C}_{\text{g}}, \tau)$ ▷ Conduct centroid matching
    $(N_{\text{match}}(t), N_{\text{p}}(t), N_{\text{g}}) \leftarrow (|\mathcal{M}(t)|, |\widehat{\mathcal{C}}_{\text{p}}(t)|, |\mathcal{C}_{\text{g}}|)$ ▷ Statistical count
    if $N_{\text{p}}(t) = 0$ **then**
        $\text{Precision}(t) \leftarrow 0$ ▷ Set Precision to 0 to avoid division-by-zero
    else
        $\text{Precision}(t) \leftarrow N_{\text{match}}(t)/N_{\text{p}}(t)$ ▷ Compute the Precision at threshold t
    end if
    $\text{Recall}(t) \leftarrow N_{\text{match}}(t)/N_{\text{g}}$ ▷ Compute the Recall at threshold t
    $(\mathcal{P}, \mathcal{R}) \leftarrow (\mathcal{P}, \mathcal{R}) \cup \{(\text{Precision}(t), \text{Recall}(t))\}$ ▷ Append to the list
end for
Step HSE-T-2: Integrate the Precision-Recall pairs $(\mathcal{P}, \mathcal{R})$ across all thresholds to obtain the *HSE-T* according to Eq.(12).
Step HSE-FINAL: Fuse pixel-level *HSE-P* and target-level *HSE-T* to obtain the final *HSE* according to Eq.(10).
Return: *HSE-P*, *HSE-T*, and *HSE*.

where Y_{main} and Y_{light} denote the outputs of the main and lightweight branches, Y_{gt} denotes the ground truth, and $\mathcal{L}_{EEDM}$ is the edge-enhanced difficulty-mining loss [43, 57]. During training, the two branches share parameters for synchronized backpropagation, and the overall objective is:

$$\mathcal{L}_{CO\text{-}ISD} = \mathcal{L}_{main} + \alpha \cdot \mathcal{L}_{light} \tag{9}$$

where $\mathcal{L}_{CO\text{-}ISD}$ denotes the total loss. α is set to 1.

Unlike explicit distillation, CO-ISD introduces no extra feature- or logit-matching loss. Instead, VFM-enhanced gradients from the main branch implicitly regularize the shared encoder-decoder parameters through synchronous optimization, enabling the lightweight branch to inherit semantic priors. After training, the VFM branch and SAMF modules are removed, and only the lightweight branch is retained for standalone inference without additional VFM-related overhead.

3.4 HSE Metric

In the SIRST task, existing evaluation protocols suffer from fragmentation and inherent limitations. Specifically, existing studies rely on a set of discrete metrics (e.g. IoU, $nIoU$, P_d, F_a), which fail to comprehensively reflect the quality of model-generated probability maps. At the same time, the $ROC\text{-}AUC$ aggregates performance across thresholds but is insensitive to false positives under extreme class imbalance (please see Sec. 4.5). To address these limitations, we propose a unified and comprehensive HSE metric as shown in Algorithm 1. Notably, *HSE* differs fundamentally from *hIoU* [25]. The *hIoU* adopts hierarchical discrete decisions and emphasizes target-level results by considering pixel-level IoU only for correctly detected targets. In contrast, our *HSE* unifies pixel- and target-level evaluation via continuous multi-threshold integration, jointly capturing pixel-level confidence stability and target-level robustness while avoiding single-threshold bias. Tab. 14 in Appendix D provides a detailed quantitative comparison.

Table 1: Comparison of different methods on the SIRST3 dataset. *"+ FDEP w/o CO-ISD"* denotes the FDEP without using CO-ISD. Best and second-best results are highlighted in red and blue.

Methods	Description	Conventional Accuracy Metrics				HSE Series Accuracy Metrics			Computational Cost	
		IoU ↑	nIoU ↑	P_d ↑	F_a ↓	HSE-P ↑	HSE-T ↑	HSE ↑	Param. (M)	Time (s)
ACM [5]	Original	62.76	61.25	94.62	18.78	82.41	82.58	68.05	**0.40**	**0.022**
	+ FDEP w/o CO-ISD	**70.43**	**69.73**	**96.48**	19.70	**90.02 (⇑ 7.61)**	**86.66 (⇑ 4.08)**	**78.01 (⇑ 9.96)**	0.85	0.082
	+ FDEP Framework	68.56	68.01	95.42	**13.92**	89.17 (⇑ 6.76)	85.02 (⇑ 2.44)	75.81 (⇑ 7.76)	0.85	**0.022**
ALCNet [6]	Original	68.30	69.55	95.28	34.07	87.56	84.74	74.20	**0.43**	**0.021**
	+ FDEP w/o CO-ISD	**73.17**	**73.23**	**97.14**	24.95	**91.73 (⇑ 4.17)**	**88.67 (⇑ 3.93)**	**81.34 (⇑ 7.14)**	0.88	0.082
	+ FDEP Framework	70.57	72.88	95.61	**22.05**	89.75 (⇑ 2.19)	87.25 (⇑ 2.51)	78.31 (⇑ 4.11)	0.88	**0.021**
MLCL-Net [39]	Original	79.83	82.88	96.35	19.02	94.37	91.98	86.80	**0.66**	**0.025**
	+ FDEP w/o CO-ISD	**84.16**	**85.08**	**97.54**	**8.59**	**96.63 (⇑ 2.26)**	**93.57 (⇑ 1.59)**	**90.43 (⇑ 3.63)**	1.11	0.087
	+ FDEP Framework	80.24	83.46	95.42	12.25	94.27 (⇓ 0.10)	92.50 (⇑ 0.52)	87.20 (⇑ 0.40)	1.11	**0.025**
ALCL-Net [40]	Original	80.94	83.39	96.74	11.22	95.09	91.23	86.75	**5.67**	**0.024**
	+ FDEP w/o CO-ISD	**83.06**	**84.16**	**97.21**	**8.49**	**96.31 (⇑ 1.22)**	**92.30 (⇑ 1.07)**	**88.90 (⇑ 2.15)**	9.74	0.086
	+ FDEP Framework	81.83	83.95	96.41	16.09	95.74 (⇑ 0.65)	91.29 (⇑ 0.06)	87.40 (⇑ 0.65)	9.74	**0.024**
DNA-Net [14]	Original	82.59	86.08	96.68	9.29	95.38	93.41	89.10	**4.70**	**0.095**
	+ FDEP w/o CO-ISD	**83.06**	**86.09**	**98.14**	**8.90**	**96.33 (⇑ 0.95)**	**94.99 (⇑ 1.58)**	**91.50 (⇑ 2.40)**	8.77	0.160
	+ FDEP Framework	82.54	86.06	97.54	11.93	95.46 (⇑ 0.08)	93.89 (⇑ 0.48)	89.62 (⇑ 0.52)	8.77	**0.095**
GGL-Net [55]	Original	81.69	84.74	97.21	**7.78**	95.24	92.39	87.99	**8.99**	**0.036**
	+ FDEP w/o CO-ISD	**84.35**	85.62	**97.54**	13.90	**96.78 (⇑ 1.54)**	**94.58 (⇑ 2.19)**	**91.53 (⇑ 3.54)**	13.06	0.100
	+ FDEP Framework	83.59	**86.29**	97.34	11.46	96.48 (⇑ 1.24)	94.00 (⇑ 1.61)	90.69 (⇑ 2.70)	13.06	**0.036**
UIU-Net [34]	Original	81.99	83.73	96.54	12.15	96.00	91.69	88.02	**50.54**	**0.055**
	+ FDEP w/o CO-ISD	**84.34**	**85.58**	**97.41**	**5.35**	**96.93 (⇑ 0.93)**	**93.30 (⇑ 1.61)**	**90.44 (⇑ 2.42)**	64.71	0.120
	+ FDEP Framework	83.12	85.05	97.08	13.31	96.57 (⇑ 0.57)	93.22 (⇑ 1.53)	90.01 (⇑ 1.99)	64.71	**0.055**
MSDA-Net [58]	Original	82.86	85.46	97.01	12.53	96.43	93.04	89.72	**4.79**	**0.035**
	+ FDEP w/o CO-ISD	**84.71**	**86.57**	**98.14**	**11.95**	**97.25 (⇑ 0.82)**	**94.32 (⇑ 1.28)**	**91.73 (⇑ 2.01)**	5.24	0.100
	+ FDEP Framework	84.05	86.33	97.54	14.24	96.92 (⇑ 0.49)	94.21 (⇑ 1.17)	91.31 (⇑ 1.59)	5.24	**0.035**

The *HSE* consists of two complementary sub-metrics: *HSE-P* and *HSE-T*. The final score:

$$HSE = HSE\text{-}P \times HSE\text{-}T \tag{10}$$

where $HSE\text{-}P$ and $HSE\text{-}T$ evaluate pixel-level confidence quality and target-level detection robustness, respectively, and their product fusion jointly captures local precision and overall detectability.

For $HSE\text{-}P$, predicted probability maps $\{p_i\}_{i=1}^{N}$ and ground truths $\{G_i\}_{i=1}^{N}$ are flattened and concatenated to compute the $Precision$ - $Recall$ (PR) curve, whose area is used as the score:

$$HSE\text{-}P = \int_0^1 P(R)\,\mathrm{d}R\,, \tag{11}$$

where $P(R)$ is the $Precision$ as a function of $Recall$. Compared to $ROC\text{-}AUC$, $HSE\text{-}P$ is more sensitive to false positive pixels, offering a more reliable measure.

For $HSE\text{-}T$, we first binarize the $\{p_i\}_{i=1}^{N}$ under multiple threshold sets $T = \{t_1, t_2, \ldots, t_M\}$, and extract the connected component sets $\hat{C}_p(t_j)$ and $\hat{C}_g$ of the predicted target and the true target. For each true target centroid $c_g \in C_g$, iterate through all unmatched predicted target centroids $\hat{c}_p \in \hat{C}_p(t_j)$. If the Euclidean distance meets the set threshold, it is regarded as a correct detection, and the predicted target is marked as "matched" to prevent duplication. Then, based on the number of successful matches, predicted targets, and true targets, $Precision$ and $Recall$ are calculated at each threshold. Finally, $HSE\text{-}T$ is obtained by discretely integrating the *Precision-Recall* pairs:

$$HSE\text{-}T = \sum_{j=1}^{M} P(t_j)\,[R(t_j) - R(t_{j+1})], \tag{12}$$

where $P(t_j)$ and $R(t_j)$ denote the $Precision$ and $Recall$ at threshold t_j. Unlike traditional metrics such as P_d and F_a, $HSE\text{-}T$ more stably measures the model's detection robustness and generalization ability under different decision boundaries through discrete integration with multiple thresholds.

4 Experiments

4.1 Datasets

We conduct experiments on four public SIRST datasets: SIRST3 [43, 37], NUAA-SIRST [5], NUDT-SIRST [14] and IRSTD-1K [49]. They contain 2755, 427, 1327, and 1001 samples respectively. These datasets cover a wide range of scenarios, from real-world complex scenes to synthetic multi-target environments, providing a comprehensive basis for verifying the robustness and generalization.

Table 2: Comparison of different methods on the NUAA-SIRST, NUDT-SIRST, and IRSTD-1k datasets. *"+ FDEP w/o CO-ISD"* denotes the FDEP framework without CO-ISD strategy.

Methods	Description	NUAA-SIRST (213:214)					NUDT-SIRST (663:664)					IRSTD-1K (800: 201)				
		IoU	P_d	HSE-P	HSE-T	HSE	IoU	P_d	HSE-P	HSE-T	HSE	IoU	P_d	HSE-P	HSE-T	HSE
ACM	Original	67.07	91.63	87.25	89.23	77.86	67.31	97.88	86.31	85.22	73.55	63.26	**90.24**	83.86	**86.70**	72.71
	+ FDEP w/o CO-ISD	**70.94**	**94.68**	**90.95**	**90.16**	**82.01**	**72.16**	**98.41**	**90.85**	**85.78**	**77.93**	**67.71**	89.23	**88.57**	83.76	**74.19**
	+ FDEP Framework	70.34	**94.68**	89.60	89.39	80.09	69.82	97.67	89.56	85.24	76.34	65.96	89.90	87.44	84.34	73.75
ALCNet	Original	70.64	92.02	90.56	87.09	78.87	72.73	96.19	91.20	88.86	81.04	67.15	90.91	88.82	86.97	77.24
	+ FDEP w/o CO-ISD	**74.15**	**95.82**	**93.15**	**91.66**	**85.38**	**77.01**	**98.20**	**93.95**	**89.42**	**84.01**	**68.73**	**95.29**	89.52	**88.45**	**79.18**
	+ FDEP Framework	71.68	93.54	91.51	88.57	81.04	74.15	97.14	92.04	88.97	81.89	68.02	91.25	**89.76**	88.15	79.12
MLCL-Net	Original	76.37	96.96	93.80	88.79	83.28	94.47	98.73	99.35	96.73	96.11	66.07	90.24	84.62	85.72	72.53
	+ FDEP w/o CO-ISD	**80.35**	**97.72**	**95.88**	**93.52**	**89.66**	**95.82**	**99.47**	**99.48**	**98.21**	**97.70**	**71.68**	**91.58**	**91.05**	**87.36**	**79.54**
	+ FDEP Framework	77.59	95.82	92.98	91.84	85.40	94.93	99.05	99.37	97.30	96.69	66.41	90.91	85.82	85.07	73.01
ALCL-Net	Original	76.52	95.82	93.44	90.02	84.11	94.25	**99.15**	**99.38**	94.25	93.66	67.35	92.26	86.49	**86.11**	74.48
	+ FDEP w/o CO-ISD	**79.07**	**97.34**	**95.11**	**92.14**	**87.63**	94.39	98.10	99.01	96.41	95.46	**72.24**	**92.93**	**91.37**	85.74	**78.34**
	+ FDEP Framework	77.69	95.82	93.10	91.86	85.53	**94.44**	98.94	99.31	**96.54**	**95.88**	69.82	92.26	88.95	83.55	75.15
DNA-Net	Original	76.29	96.58	93.56	91.19	85.31	95.17	99.26	99.45	96.88	96.34	68.07	94.28	86.27	86.83	74.91
	+ FDEP w/o CO-ISD	**80.36**	**97.72**	**95.70**	**93.07**	**89.07**	95.20	99.26	**99.47**	**98.21**	**97.69**	**71.53**	**94.61**	90.78	**87.49**	**79.42**
	+ FDEP Framework	79.26	96.58	95.09	92.14	87.62	**95.45**	**99.37**	99.45	98.00	97.47	71.05	93.27	**90.80**	84.10	76.37
GGL-Net	Original	79.46	97.34	94.64	93.44	88.43	94.68	98.94	99.25	97.26	96.53	67.68	89.56	87.06	86.84	75.60
	+ FDEP w/o CO-ISD	79.59	**98.48**	**95.70**	**94.48**	**90.41**	**95.49**	**99.26**	**99.40**	**98.00**	**97.41**	**68.77**	**94.28**	**88.61**	89.16	**79.00**
	+ FDEP Framework	**80.12**	98.10	94.84	94.04	89.19	95.23	99.05	99.36	97.46	96.84	67.42	92.93	86.06	**90.24**	77.66
UIU-Net	Original	78.23	96.58	94.37	91.03	85.91	94.88	99.05	99.44	97.59	97.05	67.64	93.27	86.27	**88.73**	76.55
	+ FDEP w/o CO-ISD	79.34	**98.10**	**95.87**	90.11	**86.39**	95.52	99.05	99.40	97.89	97.30	**72.44**	**94.28**	**91.95**	85.84	78.93
	+ FDEP Framework	**79.65**	97.34	94.97	**90.51**	85.96	**95.55**	**99.37**	**99.47**	**97.90**	**97.38**	68.46	91.58	90.14	87.67	**79.03**
MSDA-Net	Original	75.94	96.96	93.25	89.64	83.58	96.03	99.15	99.54	97.10	96.66	71.47	92.26	91.21	85.89	78.34
	+ FDEP w/o CO-ISD	**79.48**	**98.10**	**95.29**	**92.35**	**88.00**	96.01	99.05	99.43	**98.08**	**97.53**	**71.96**	93.27	**91.86**	86.07	79.06
	+ FDEP Framework	75.98	95.06	92.84	90.44	83.96	**96.50**	**99.37**	**99.50**	97.91	97.43	71.52	**93.60**	91.74	**86.51**	**79.37**

4.2 Implementation Details

Experiment setting. We use the AdamW [22] optimizer with an initial learning rate of 1.0×10^{-3}. The batch size is 16, the number of epochs is 400, and the GPU is an RTX 4090 24 GB. During training, all input images are normalized and randomly cropped into 448 × 448 pixels image patches. Unless otherwise specified, the default VFMs used are *DINOv2 (ViT-L/14)*.

Evaluation metrics. To facilitate a smooth transition, in addition to HSE series, we also adopt four commonly used metrics: pixel-level metrics (IoU and $nIoU$) and target-level metrics (P_d and F_a). Consistent with previous research [57, 43, 41, 15], the model is invalid when $F_a > 1e^{-4}$.

4.3 Comparison with SOTA Methods

We embed multiple excellent SIRST detection networks [14, 39, 40, 55, 6, 58, 5, 34] into our framework. For fair comparison, we retrain all models under the same settings. More experimental results are provided in Tab. 12 and Tab. 13 of Appendix B.

Evaluation on the SIRST3 dataset. From Tab. 1, first, when the frozen VFM is introduced (i.e., *"+FDEP w/o CO-ISD"*), all networks exhibit consistent improvements, with IoU, $nIoU$, P_d, and HSE increasing by 0.47 - 7.67, 0.01 - 8.48, 0.33 - 1.86, and 2.01 - 9.96, respectively. Second, compared with *"+FDEP w/o CO-ISD"*, the introduction of the CO-ISD strategy further reduces inference time by 40.6% (from 0.160 to 0.095) - 74.4% (from 0.082 to 0.021), while maintaining comparable detection performance. For most networks (e.g., MSDA-Net), the difference in HSE is less than 1.0, indicating effective representation alignment between the main and lightweight branches. Finally, the FDEP framework consistently enhances performance across diverse models without increasing inference cost, with HSE improving by 0.40 - 7.76, highlighting its strong transferability and generalization. From Fig. 4 of the Appendix C, both *"+FDEP w/o CO-ISD"* and *"+FDEP Framework"* deliver comparable results, which are better than those of the original networks.

Evaluation on three individual datasets. As shown in Tab. 2, both *"+ FDEP w/o CO-ISD"* and *"+ FDEP Framework"* consistently improve performance over the *"Original"* across all three datasets. Specifically, *"+ FDEP w/o CO-ISD"* increases the HSE by 0.48 - 6.51, 0.25 - 4.38, 0.72 - 7.01 on the NUAA-SIRST, NUDT-SIRST, and IRSTD-1K datasets, respectively. Similarly, *"+ FDEP Framework"* increases the HSE by 0.05 - 2.31, 0.31 - 2.79, and 0.48 - 2.48, respectively. Overall, these results fully validate that our FDEP delivers stable gains under diverse challenging conditions with few-shot samples. From Figs. 5–7 of the Appendix C, both *"+FDEP w/o CO-ISD"* and *"+FDEP Framework"* deliver comparable results, which are better than those of the original networks.

Table 3: Break-down ablation on the SIRST3 dataset.

Scheme	Variants			MSDA-Net + FDEP Framework				
	VFMs	SAMF	CO-ISD	HSE-P ↑	HSE-T ↑	HSE ↑	Param. (M) ↓	Time (s) ↓
Original	✗	✗	✗	96.43	93.04	89.72	**4.79**	**0.037**
Scheme1	✓	✗	✗	96.93 (+0.50)	94.29 (+1.25)	91.40 (+1.68)	5.05	0.100
Scheme2	✓	✓	✗	**97.25** (+0.82)	**94.32** (+1.28)	**91.73** (+2.01)	5.24	0.100
Scheme3	✓	✗	✓	96.76 (+0.33)	93.73 (+0.69)	90.92 (+1.20)	5.05	**0.037**
Scheme4	✓	✓	✓	**96.92** (+0.49)	**94.21** (+1.17)	**91.31** (+1.59)	5.24	**0.037**

Table 4: Evaluation of the VFMs directly applied to the SIRST detection on the SIRST3 dataset.

Scheme	VFM Frozen	Conventional Metrics				HSE Series Metrics		
		IoU ↑	nIoU ↑	P_d ↑	F_a ↓	HSE-P ↑	HSE-T ↑	HSE ↑
VFMs + Decoder	✓	34.15	26.98	55.48	91.11	52.62 (+44.63)	41.38 (+52.94)	21.77 (+69.96)
VFMs + Decoder	✗	33.87	26.87	56.81	80.67	52.03 (+45.22)	44.27 (+50.05)	23.03 (+68.70)
VFMs + MSDA-Net (Decoder)	✓	69.08	69.40	96.40	18.68	90.00 (+7.25)	90.96 (+3.36)	81.86 (+9.87)
VFMs + MSDA-Net (Decoder)	✗	69.98	70.91	96.81	17.30	90.31 (+6.94)	91.50 (+2.82)	82.64 (+9.09)
FDEP w/o CO-ISD	✓	**84.71**	**86.57**	**98.14**	**11.95**	**97.25**	**94.32**	**91.73**

Table 5: Performance comparison of FDEP framework with different VFMs on the SIRST3 dataset. "Baseline (N/A)" denotes the original MSDA-Net without VFM integration.

VFMs (Scales)	MSDA-Net + FDEP			VFMs (Scales)	MSDA-Net + FDEP		
	HSE-P ↑	HSE-T ↑	**HSE ↑**		HSE-P ↑	HSE-T ↑	**HSE ↑**
Baseline (N/A)	96.43	93.04	**89.72**	CLIP (ViT-H-14)	96.86	93.46	**90.52** (+0.80)
MAE (ViT-B)	96.89	93.72	**90.81** (+1.09)	DINOv2 (ViT-S/14)	96.66	93.63	**90.51** (+0.79)
MAE (ViT-L)	96.89	94.44	**91.51** (+1.79)	DINOv2 (ViT-B/14)	96.91	93.87	**90.98** (+1.26)
SAM1 (ViT-L)	96.70	93.36	**90.29** (+0.57)	DINOv2 (ViT-L/14)	96.92	94.21	**91.31** (+1.59)
SAM1 (ViT-H)	97.15	93.39	**90.73** (+1.01)	DINOv3 (ViT-S+/16)	96.68	93.88	**90.76** (+1.04)
SAM2.1 (Base+)	96.85	93.98	**91.02** (+1.30)	DINOv3 (ViT-B/16)	97.10	94.00	**91.27** (+1.55)
SAM2.1 (Large)	96.24	95.27	**91.69** (+1.97)	DINOv3 (ViT-L/16)	96.63	94.85	**91.66** (+1.94)

4.4 Ablation Experiment

To fully validate the proposed FDEP framework, we perform detailed break-down ablation experiments. From Tab. 3, compared with the *"Original"*, introducing the frozen VFM features improves the HSE metric by 1.68, indicating that the global semantic priors provided by VFMs can effectively compensate for the sparsity and limited semantic representation of target features. At the same time, from *"Scheme1-Scheme2"* and *"Scheme3-Scheme4"*, the proposed SAMF module further enhances the performance gains brought by VFM feature integration, yielding an additional average 0.36 HSE gain with negligible inference cost. In addition, from *"Scheme1-Scheme3"* and *"Scheme2-Scheme4"*, incorporating the proposed CO-ISD strategy significantly reduces inference time by 63.0% (from 0.100 to 0.037) while maintaining only a negligible decrease in HSE, markedly enhancing inference efficiency. Furthermore, by comparing the *"Original"* and *"Scheme4"*, SIRST networks equipped with the FDEP framework achieve an HSE improvement of 1.59 under the same inference time. In summary, the frozen VFM features provide strong global semantic priors, the SAMF module enables efficient multi-level semantic fusion, and the CO-ISD strategy ensures an optimal balance between performance and deployability. Together, these three components constitute the core driving forces behind the performance improvements achieved by the proposed FDEP framework. More detailed ablation experiments (VFM feature-level integration, VFM scale, SAMF module placement, etc.) and analysis can be found in the Appendix A.

4.5 Discussion

1) The limitations of directly applying VFMs to SIRST detection. As shown in Tab. 4, naively transferring VFMs to the SIRST task performs poorly under the tested decoders. Specifically, directly connecting a VFM to a decoder or replacing the MSDA-Net encoder remains far below the FDEP-based method. For instance, the *"VFMs + Decoder"* achieves an HSE of only 21.77, and unfreezing the VFM improves it marginally to 23.03. This shows that although VFMs exhibit powerful representational capability for natural images, their global semantic features are not directly compatible with SIRST feature distribution, leading to a notable domain gap. Moreover, compared with the *"DINO +*

Table 6: Cross-dataset generalization evaluated by HSE. The first and second header rows indicate the training and testing datasets, respectively. All variants use the MSDA-Net as the detector.

Schemes	NUAA-SIRST		NUDT-SIRST		IRSTD-1K	
	NUDT-SIRST	IRSTD-1K	NUAA-SIRST	IRSTD-1K	NUAA-SIRST	NUDT-SIRST
Original	48.70	45.32	63.12	31.10	75.39	38.97
+FDEP w/o CO-ISD	61.35 (+12.65)	62.61 (+17.29)	68.21 (+5.09)	38.17 (+7.07)	80.46 (+5.07)	53.32 (+14.35)
+FDEP Framework	52.26 (+3.56)	49.50 (+4.18)	67.23 (+4.11)	31.41 (+0.31)	83.34 (+7.95)	41.98 (+3.01)

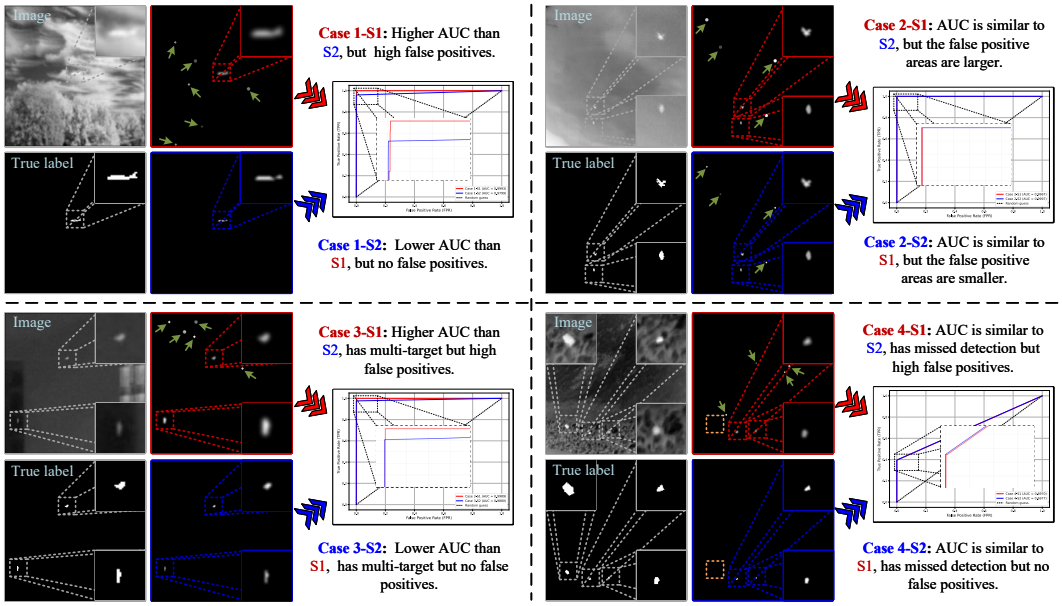

Figure 3: ROC comparisons across multiple different and representative cases. The red and blue correspond to two prediction samples (S1 and S2), respectively. The orange boxes indicate missed detections, and the green arrows denote false detections.

Decoder", using *"DINO + MSDA-Net (Decoder)"* yields a substantial improvement, indicating that task-specific network structures are better suited to capturing the subtle characteristics of infrared small targets. However, even under this configuration, it still fails to fully leverage the potential of VFMs. In contrast, the proposed FDEP achieves adaptive transfer from foundation-level semantics to task-specific semantics through explicit semantic fusion and implicit knowledge transfer.

2) The limitations of using ROC curves for SIRST detection evaluation. Conventional ROC curves fail to accurately reflect model performance in SIRST detection due to the extreme class imbalance. Specifically, target pixels occupy only a very small percentage of the image, while background pixels overwhelmingly dominate. Under such conditions, even if a model generates a large number of false positives, as long as the overall negative sample prediction is correct, the ROC curve will still present the illusion of a high AUC. The Cases in Fig. 3 can effectively verify this. For example, Case I-S1 still attains a high AUC of 0.9993 despite many false positives, whereas Case I-S2 shows much better overall detection yet a lower AUC (0.9799). Tab. 14 in the Appendix D provides more results.

3) The FDEP can be extended to various VFMs. In addition to seamlessly adapting to existing encoder-decoder-based SIRST networks, our FDEP can also be easily adapted to various VFMs, demonstrating good versatility and scalability. From Tab. 5, introducing semantic priors from different VFMs into FDEP consistently yields stable gains. Specifically, compared to the *"Baseline (N/A)"* without introducing VFMs, FDEP with different VFMs improves the *HSE* metric by 0.57-1.97, verifying that the framework does not depend on a specific VFM, but can widely benefit from the high-level semantic representation of VFMs. In addition, Tab. 5 also shows that higher *HSE* often corresponds to larger VFM scales, indicating that richer semantic and structural representations can provide stronger transferable priors for FDEP.

4) The role of VFM-derived semantic priors in SIRST detection. Although infrared small targets exhibit limited semantic appearance, the VFM-derived priors in FDEP provide complementary

information about background structure, scene context, and mid-level visual patterns. Specifically, background structural priors capture regional continuity and structural regularities, providing a global reference for identifying local anomalies. Scene-level contextual priors encode spatial layouts and dependencies between regions, helping distinguish potential targets from bright background structures such as clouds and building edges. Mid-level visual representations capture generic edges, contours, and textures that complement local contrast and target boundary cues. Thus, the semantic priors primarily describe contextual and structural information rather than target category semantics. Through SAMF, these priors are aligned and fused with task-specific features to support small target discrimination and background suppression.

5) Evaluation of Cross-Dataset Generalization. To evaluate generalization under dataset shifts, we conduct cross-dataset experiments on NUAA-SIRST, NUDT-SIRST, and IRSTD-1K, training on one dataset and testing on each of the other two. As shown in Tab. 6, compared with the original MSDA-Net, the complete FDEP framework improves HSE by 0.31-7.95 across all six transfer settings without additional VFM-related inference overhead. The accuracy-oriented variant, "+ FDEP w/o CO-ISD", achieves gains of 5.07-17.29, generally providing stronger cross-dataset performance by retaining the VFM and SAMF during inference. These results suggest that frozen VFM representations provide transferable contextual and structural priors, while CO-ISD enables the lightweight branch to retain part of their benefits after removing the VFM and SAMF. Nevertheless, the gains vary across transfer settings, with a modest improvement of 0.31 for the complete FDEP when transferring from NUDT-SIRST to IRSTD-1K, indicating that substantial domain shifts remain challenging. The two variants therefore offer complementary choices for accuracy-oriented applications and efficient deployment.

5 Conclusion

This work rethinks the SIRST detection paradigm and proposes the FDEP framework, which can seamlessly adapt to diverse VFMs and SIRST networks. Our FDEP introduces frozen VFM representations as global semantic priors and employs a SAMF module to achieve feature alignment and semantic injection, thereby enhancing weak-target responses and suppressing complex-background interference. Meanwhile, a CO-ISD strategy is proposed, which achieves implicit transfer of semantic priors through parameter sharing and synchronous backpropagation. In addition, we construct the HSE metric to address the redundancy and incompleteness of existing evaluation systems. Extensive experiments demonstrate that the SIRST networks equipped with the FDEP framework significantly improve detection accuracy while maintaining VFM-free inference efficiency. We hope this study draws attention to research on deep collaboration between VFMs and task-specific networks.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (No. 62306261), HK RGC-Early Career Scheme (No. 24211525), ITSP Platform Project (No. ITS/600/24FP), the Centre for Perceptual and Interactive Intelligence, a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government, Hong Kong RGC Strategic Topics Grant (No. STG1/E-403/24-N and No. STG1/M-404/26-N), CUHK-CUHK(SZ)-GDST Joint Collaboration Fund (No. YSP26-4760949), the LiaoNing Revitalization Program under Grant no. XLYC2201001, and the Institute Basic Research Program (No. E429040701).

References

[1] Gonzalor Arce and Michaelp Mcloughlin. Theoretical analysis of the max/median filter. *IEEE transactions on acoustics, speech, and signal processing*, 35(1):60–69, 1987.

[2] 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, 2013.

[3] Zhong Chen, Song Luo, Ting Xie, Jianguo Liu, Guoyou Wang, and Gao Lei. A novel infrared small target detection method based on bemd and local inverse entropy. *Infrared Physics & Technology*, 66:114–124, 2014.

[4] Yimian Dai, Yiquan Wu, Yu Song, and Jun Guo. Non-negative infrared patch-image model: Robust target-background separation via partial sum minimization of singular values. *Infrared Physics & Technology*, 81:182–194, 2017.

[5] 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 (WACV)*, pages 950–959, 2021.

[6] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. Attentional local contrast networks for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 59(11): 9813–9824, 2021.

[7] He Deng, Xianping Sun, Maili Liu, Chaohui Ye, and Xin Zhou. Small infrared target detection based on weighted local difference measure. *IEEE Transactions on Geoscience and Remote Sensing*, 54(7):4204–4214, 2016.

[8] Chenqiang Gao, Deyu Meng, Yi Yang, Yongtao Wang, Xiaofang Zhou, and Alexander G Hauptmann. Infrared patch-image model for small target detection in a single image. *IEEE transactions on image processing*, 22(12):4996–5009, 2013.

[9] Huaiyu Gao, Xiaoyong Jiang, Xinyu Ma, Minrui Ye, Jie Yang, Junyao Zhang, Yangchen Gao, Tangxin Li, Hailu Wang, Jian Mei, et al. Bio-inspired mid-infrared neuromorphic transistors for dynamic trajectory perception using pdse2/pentacene heterostructure. *Nature Communications*, 16(1):5241, 2025.

[10] Yaqi Han, Jingwen Liao, Tianshu Lu, Tian Pu, and Zhenming Peng. Kcpnet: Knowledge-driven context perception networks for ship detection in infrared imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–19, 2022.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[12] YuJie He, Min Li, JinLi Zhang, and Qi An. Small infrared target detection based on low-rank and sparse representation. *Infrared Physics & Technology*, 68:98–109, 2015.

[13] Feng Huang, Shuyuan Zheng, Zhaobing Qiu, Huanxian Liu, Huanxin Bai, and Liqiong Chen. Text-irstd: Leveraging semantic text to promote infrared small target detection in complex scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10635–10644, 2025.

[14] 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, 2022.

[15] Boyang Li, Xinyi Ying, Ruojing Li, Yongxian Liu, Yangsi Shi, Miao Li, Xin Zhang, Mingyuan Hu, Chenyang Wu, Yukai Zhang, et al. Icpr 2024 competition on resource-limited infrared small target detection challenge: Methods and results. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 62–77. Springer, 2024.

[16] Haoqing Li, Jinfu Yang, Runshi Wang, and Yifei Xu. Ilnet: Low-level matters for salient infrared small target detection. *IEEE Transactions on Aerospace and Electronic Systems*, 2025.

[17] Jianfu Li, Weiguo Gong, Weihong Li, and Xiaoying Liu. Robust pedestrian detection in thermal infrared imagery using the wavelet transform. *Infrared Physics & Technology*, 53(4):267–273, 2010.

[18] Qiang Li, Wei Zhang, Wanxuan Lu, and Qi Wang. Multi-branch mutual-guiding learning for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[19] Yansheng Li and Yongjun Zhang. Robust infrared small target detection using local steering kernel reconstruction. *Pattern Recognition*, 77:113–125, 2018.

[20] Fanzhao Lin, Shiming Ge, Kexin Bao, Chenggang Yan, and Dan Zeng. Learning shape-biased representations for infrared small target detection. *IEEE Transactions on Multimedia*, 26: 4681–4692, 2023.

[21] 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 (CVPR)*, pages 17490–17499, 2024.

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[23] Husain Najafi, Pallav Kumar Shrestha, Oldrich Rakovec, Heiko Apel, Sergiy Vorogushyn, Rohini Kumar, Stephan Thober, Bruno Merz, and Luis Samaniego. High-resolution impact-based early warning system for riverine flooding. *Nature communications*, 15(1):3726, 2024.

[24] 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.

[25] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, Huchuan Lu, Georges El Fakhri, Xiaofeng Liu, and Shijian Lu. Rethinking evaluation of infrared small target detection. *arXiv preprint arXiv:2509.16888*, 2025.

[26] Baojun Qi, Tao Wu, and Hangen He. Robust detection of small infrared objects in maritime scenarios using local minimum patterns and spatio-temporal context. *Optical Engineering*, 51 (2):027205–027205, 2012.

[27] Yao Qin, Lorenzo Bruzzone, Chengqiang Gao, and Biao Li. Infrared small target detection based on facet kernel and random walker. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):7104–7118, 2019.

[28] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 839–846. IEEE, 1998.

[29] Zhengzheng Tu, Yan Ma, Zhun Li, Chenglong Li, Jieming Xu, and Yongtao Liu. Rgbt salient object detection: A large-scale dataset and benchmark. *IEEE Transactions on Multimedia*, 25: 4163–4176, 2022.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems (NeurIPS)*, 30, 2017.

[31] Yantao Wei, Xinge You, and Hong Li. Multiscale patch-based contrast measure for small infrared target detection. *Pattern recognition*, 58:216–226, 2016.

[32] Fengyi Wu, Tianfang Zhang, Lei Li, Yian Huang, and Zhenming Peng. Rpcanet: Deep unfolding rpca based infrared small target detection. In *WACV*, pages 4809–4818, 2024.

[33] Tianhao Wu, Boyang Li, Yihang Luo, Yingqian Wang, Chao Xiao, Ting Liu, Jungang Yang, Wei An, and Yulan Guo. Mtu-net: Multilevel transunet for space-based infrared tiny ship detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[34] 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, 2022.

[35] Hai Xu, Sheng Zhong, Tianxu Zhang, and Xu Zou. Multiscale multilevel residual feature fusion for real-time infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023.

[36] Huoren Yang, Tingkui Mu, Ziyue Dong, Zicheng Zhang, Bin Wang, Wei Ke, Qiujie Yang, and Zhiping He. Pbt: Progressive background-aware transformer for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024.

[37] 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 (CVPR)*, pages 15528–15538, 2023.

[38] Xinyi Ying, Li Liu, Zaipin Lin, Yangsi Shi, Yingqian Wang, Ruojing Li, Xu Cao, Boyang Li, Shilin Zhou, and Wei An. Infrared small target detection in satellite videos: a new dataset and a novel recurrent feature refinement framework. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[39] 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.

[40] 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.

[41] Chuang Yu, Yunpeng Liu, Jinmiao Zhao, and Zelin Shi. Lr-net: a lightweight and robust network for infrared small target detection. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 19–33. Springer, 2024.

[42] Chuang Yu, Yunpeng Liu, Jinmiao Zhao, Dou Quan, Zelin Shi, and Xiangyu Yue. Relational representation learning network for cross-spectral image patch matching. *Information Fusion*, page 103749, 2025.

[43] 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 (ICCV)*, pages 2588–2598, 2025.

[44] Di Yuan, Haiping Zhang, Xiu Shu, Qiao Liu, Xiaojun Chang, Zhenyu He, and Guangming Shi. Thermal infrared target tracking: A comprehensive review. *IEEE Transactions on Instrumentation and Measurement*, 73:1–19, 2023.

[45] 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.

[46] Taoran Yue, Xiaojin Lu, Jiaxi Cai, Yuanping Chen, and Shibing Chu. Sds-net: Shallow-deep synergism-detection network for infrared small target detection. *arXiv preprint arXiv:2506.06042*, 2025.

[47] Landan Zhang, Lingbing Peng, Tianfang Zhang, Siying Cao, and Zhenming Peng. Infrared small target detection via non-convex rank approximation minimization joint l 2, 1 norm. *Remote Sensing*, 10(11):1821, 2018.

[48] Mingjin Zhang, Ke Yue, Jing Zhang, Yunsong Li, and Xinbo Gao. Exploring feature compensation and cross-level correlation for infrared small target detection. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 1857–1865, 2022.

[49] 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 (CVPR)*, pages 877–886, 2022.

[50] Mingjin Zhang, Yuchun Wang, Jie Guo, Yunsong Li, Xinbo Gao, and Jing Zhang. Irsam: Advancing segment anything model for infrared small target detection. In *European Conference on Computer Vision*, pages 233–249. Springer, 2024.

[51] Mingjin Zhang, Ke Yue, Boyang Li, Jie Guo, Yunsong Li, and Xinbo Gao. Single-frame infrared small target detection via gaussian curvature inspired network. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024.

[52] 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 (CVPR)*, pages 9549–9558, 2025.

[53] Tianfang Zhang, Lei Li, Siying Cao, Tian Pu, and Zhenming Peng. Attention-guided pyramid context networks for detecting infrared small target under complex background. *IEEE Transactions on Aerospace and Electronic Systems*, 59(4):4250–4261, 2023.

[54] Enzhong Zhao, Lili Dong, Chenxi Li, and Yingjie Ji. Infrared maritime target detection based on temporal weight and total variation regularization under strong wave interferences. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[55] 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.

[56] Jinmiao Zhao, Zelin Shi, Chuang Yu, and Yunpeng Liu. Refined infrared small target detection scheme with single-point supervision. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 3–18. Springer, 2024.

[57] Jinmiao Zhao, Zelin Shi, Chuang Yu, Yunpeng Liu, and Yimain Dai. Towards robust infrared small target detection: A feature-enhanced and sensitivity-tunable framework. *Knowledge-Based Systems*, page 114519, 2025.

[58] Jinmiao Zhao, Zelin Shi, Chuang Yu, Yunpeng Liu, Xinyi Ying, and Yimian Dai. Multi-scale direction-aware network for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

Table 7: Investigation of introducing different VFMs levels on the SIRST3 dataset.

Levels	MSDA-Net + FDEP			Levels	MSDA-Net + FDEP		
	HSE-P ↑	HSE-T ↑	HSE ↑		HSE-P ↑	HSE-T ↑	HSE ↑
1	96.17	93.69	90.10	12/24	96.70	94.31	91.21
8	96.91	93.95	91.04	8/16/24	96.91	94.19	91.28
16	96.68	94.20	91.07	6/12/18/24	**96.92**	94.21	**91.31**
24	96.92	93.88	90.98	4/9/14/19/24	96.41	**94.59**	91.19

Table 8: Investigation of VFM scale and SAMF strategy in FDEP on the SIRST3 dataset.

Scales	SAMF	MSDA-Net + FDEP			SAMF	MSDA-Net + FDEP		
		HSE-P ↑	HSE-T ↑	HSE ↑		HSE-P ↑	HSE-T ↑	HSE ↑
ViT-S/14	✗	96.52	93.45	90.20	✓	**96.66** (+0.14)	**93.63** (+0.18)	**90.51** (+0.31)
ViT-B/14	✗	96.49	93.70	90.41	✓	**96.91** (+0.42)	**93.87** (+0.17)	**90.98** (+0.57)
ViT-L/14	✗	96.76	93.73	90.92	✓	**96.92** (+0.16)	**94.21** (+0.48)	**91.31** (+0.39)

In this appendix, we provide additional details and results to complement the main paper. In Appendix A, we present more ablation experiments to thoroughly explore the performance of the proposed FDEP framework. In Appendix B, we offer quantitative comparisons with more existing methods. In Appendix C, we provide additional qualitative comparisons on three individual datasets. In Appendix D, we provide further discussion. In Appendix E, we provide a discussion of the potential societal impacts and practical applications of our work. In Appendix F, we provide the limitations of the current study and outline possible directions for future work.

A More Ablation Experiments

In this section, we present additional ablation studies to further investigate VFM feature-level integration, VFM scale and integration strategy, SAMF module placement, and the CO-ISD strategy.

1) Investigation of VFM feature-level integration. We evaluate DINO feature-level integration on the SIRST3 dataset. From Tab. 7, compared with shallow-layer features (e.g., the 1_{st} layer), deeper features yield more substantial and stable gains. Specifically, relative to the *"Original"* in Tab. 3, introducing only the 1_{st} layer improves the HSE by 0.38 (from 89.72 to 90.10), while the 8_{th}, 16_{th}, and 24_{th} layers achieve gains of 1.32 (from 89.72 to 91.04), 1.35 (from 89.72 to 91.07), and 1.26 (from 89.72 to 90.98), respectively. This indicates that high-level semantic features play a crucial role in enhancing robustness and suppressing complex backgrounds. In addition, multi-level feature fusion yields the best HSE at layers 6/12/18/24, indicating complementary semantics across layers for hierarchical representations. Notably, adding more layers ("4/9/14/19/24") may lead to feature-space redundancy and overlap, providing limited additional benefits.

2) Investigation of VFM scale and integration strategy. From Tab. 8, as the scale of the VFMs increases, the model exhibits a consistent improvement in the HSE. This indicates that larger-scale VFMs provide richer global priors for SIRST detection. Meanwhile, compared to not using the SAMF module, using this module can achieve further performance improvements across all settings. Specifically, the HSE metric improves by 0.31 (from 90.20 to 90.51) - 0.57 (from 90.41 to 90.98), demonstrating that the SAMF module effectively aligns the semantics from the VFMs with task-specific features, thereby achieving more efficient semantic fusion. To further investigate the impact of different scales of VFMs and their feature integration methods on the performance, in addition to the exploration on the *"+ FDEP Framework"*, we also conduct detailed experimental studies on the high-precision *"+ FDEP w/o CO-ISD"* framework. The experimental results are shown in Tab. 9. Consistent with the experiments conducted using *"+ FDEP Framework"*, the *HSE* continues to improve as the scale of the VFMs increases. This indicates that larger backbone models possess stronger semantic abstraction and feature representation capabilities, thereby providing richer global priors for SIRST detection. Specifically, compared with *"ViT-S/14"*, using *"ViT-B/14"* improves *HSE* by 0.45 (from 90.58 to 91.03) and 0.34 (from 91.00 to 91.34). Using *"ViT-L/14"* improves *HSE* by 0.82 (from 90.58 to 91.40) and 0.73 (from 91.00 to 91.73). In addition, compared with the variants without the SAMF module, equipping the model with the proposed SAMF module consistently yields

Table 9: Investigation of VFM scale and SAMF strategy in FDEP w/o CO-ISD on the SIRST3 dataset.

Scales	SAMF	MSDA-Net + FDEP w/o CO-ISD			SAMF	MSDA-Net + FDEP w/o CO-ISD		
		HSE-P ↑	HSE-T ↑	HSE ↑		HSE-P ↑	HSE-T ↑	HSE ↑
ViT-S/14	✗	97.08	93.30	90.58	✓	**97.21** (+0.13)	**93.61** (+0.31)	**91.00** (+0.42)
ViT-B/14	✗	96.76	94.08	91.03	✓	**97.34** (+0.58)	**93.83** (-0.25)	**91.34** (+0.31)
ViT-L/14	✗	96.93	94.29	91.40	✓	**97.25** (+0.32)	**94.32** (+0.03)	**91.73** (+0.33)

Table 10: Investigation of SAMF placement on the SIRST3 dataset. *"Baseline"* denotes the original MSDA-Net. *"Each stage"* indicates injecting VFM features via SAMF at each stage. *CL*: Cross-layer.

Scheme	Conventional Metrics				HSE Series Metrics		
	mIoU ↑	nIoU ↑	P_d ↑	F_a ↓	HSE-P ↑	HSE-T ↑	HSE ↑
Baseline (MSDA-Net)	82.86	85.46	97.01	12.53	96.43	93.04	89.72
Encoder (Each stage)	82.71	85.13	**97.81**	15.20	96.31 (-0.12)	93.47 (+0.43)	90.02 (+0.30)
Decoder (Each stage)	83.86	85.55	96.54	**11.49**	96.80 (+0.37)	93.27 (+0.23)	90.28 (+0.56)
CL Connection (Each stage)	83.23	85.81	97.28	21.71	96.58 (+0.15)	93.52 (+0.48)	90.33 (+0.61)
Final FDEP (Ours)	**84.05**	**86.33**	97.54	14.24	**96.92** (+0.49)	**94.21** (+1.17)	**91.31** (+1.59)

Table 11: Investigation of CO-ISD strategy on the SIRST3 dataset. *EDL*: Explicit Distillation Loss.

Share	EDL	MSDA-Net + FDEP			EDL	MSDA-Net + FDEP		
		HSE-P ↑	HSE-T ↑	HSE ↑		HSE-P ↑	HSE-T ↑	HSE ↑
Encoder (E)	✓	94.71 (+1.88)	90.36 (+3.30)	85.58 (+4.89)	✗	96.00 (+0.92)	92.50 (+1.71)	88.81 (+2.50)
Decoder (D)	✓	94.67 (+1.92)	89.13 (+4.53)	84.38 (+6.09)	✗	96.40 (+0.52)	91.96 (+2.25)	88.64 (+2.67)
E + D	✓	96.59	93.66	90.47	✗	**96.92**	**94.21**	**91.31**

performance gains across all backbone scales. Specifically, using SAMF improves *HSE* by 0.31 (from 91.03 to 91.34) - 0.42 (from 90.58 to 91.00). These results further validate that the SAMF module can effectively align the semantic features of the VFMs with task-specific representations, enabling more efficient semantic transfer and fusion.

3) Investigation of SAMF module placement. To systematically evaluate the impact of SAMF injection location on performance, we deployed it in each stage in the encoder, decoder, and cross-layer connections of MSDA-Net. The experimental results are shown in Tab. 10. Overall, compared to the baseline without VFM, all three locations (encoder, decoder, and cross-layer connections) provide stable gains, improving *HSE* by 0.30 (from 89.72 to 90.02), 0.56 (from 89.72 to 90.28), and 0.61 (from 89.72 to 90.33) respectively. The proposed FDEP further improves *HSE* by 1.59 (from 89.72 to 91.31), achieving the best performance. This shows that VFMs priors have universal validity, but the injection location determines the upper limit of gain. Specifically, although the encoder's stage-by-stage injection improves P_d and slightly increases *HSE-T*, it is accompanied by a slight decrease in mIoU/nIoU and an increase in F_a. This indicates that introducing too strong semantic priors in the early layers, which are dominated by local textures, can easily activate false background responses, thereby weakening the overall gains. In contrast, the stage-by-stage injection of the decoder is more conducive to reconstruction and denoising (F_a: 12.53→11.49, mIoU: 82.86→83.86), thus bringing a more significant improvement in *HSE* (+0.56). However, P_d slightly decreases, indicating that its main contribution is in refinement rather than recall enhancement. Cross-layer connections and staged injection can improve target-level stability (*HSE-T* is significantly improved, *HSE* +0.61), but F_a surges to 21.71, reflecting that direct superposition of semantic modulation in skip paths easily amplifies high-frequency background interference. Ultimately, FDEP simultaneously improved mIoU/nIoU and P_d while keeping F_a at 14.24, achieving a maximum *HSE* of 91.31 (+1.59), further validating that its advantage stems from a more reasonable semantic alignment and fusion paradigm, rather than simple position stacking.

4) Investigation of CO-ISD strategy. We evaluate different parameter-sharing schemes and the explicit distillation loss. From Tab. 11, when only the encoder is shared, the absence of gradient collaboration in decoding limits the HSE to 88.81. When only the decoder is shared, the absence of global constraints at the feature level limits the HSE to 88.64. In contrast, sharing both encoder and decoder leads to a notable performance improvement, indicating that end-to-end gradient collaboration within a unified parameter space is crucial for effective implicit semantic transfer. In addition, the

Table 12: Performance comparison of different methods on the SIRST3 dataset. *"+ FDEP w/o CO-ISD"* denotes the FDEP framework without using the CO-ISD strategy. Red denotes the best result, and blue denotes the second best result. The values in parentheses denote the difference between each method and the *"MSDA-Net + FDEP"*. The order of magnitude of F_a is 10^{-6}.

Methods	Description	Conventional Accuracy Metrics				HSE Series Accuracy Metrics			Computational Cost	
		IoU ↑	nIoU ↑	P_d ↑	F_a ↓	HSE-P ↑	HSE-T ↑	HSE ↑	Param (M)	Time (s)
FKRW [27]	Filtering	14.46	23.67	60.66	22.50	20.48 (⇑ 76.44)	40.73 (⇑ 53.48)	8.34 (⇑ 82.97)	-	0.189
MPCM [31]	Local Contrast	18.54	26.80	71.96	123.76	31.36 (⇑ 65.56)	13.50 (⇑ 80.71)	4.24 (⇑ 87.07)	-	0.104
NIPPS [4]	Low Rank	18.82	28.59	70.56	15.17	30.46 (⇑ 66.46)	34.05 (⇑ 60.16)	10.37 (⇑ 80.94)	-	31.047
ACM [5]	CNN-based	62.76	61.25	94.62	18.78	82.41 (⇑ 14.51)	82.58 (⇑ 11.63)	68.05 (⇑ 23.26)	**0.40**	0.022
ALCNet [6]	CNN-based	68.30	69.55	95.28	34.07	87.56 (⇑ 9.36)	84.74 (⇑ 9.47)	74.20 (⇑ 17.11)	0.43	0.021
MLCL-Net [39]	CNN-based	79.83	82.88	96.35	19.02	94.37 (⇑ 2.55)	91.98 (⇑ 2.23)	86.80 (⇑ 4.51)	0.66	0.025
ALCL-Net [40]	CNN-based	80.94	83.39	96.74	11.22	95.09 (⇑ 1.83)	91.23 (⇑ 2.98)	86.75 (⇑ 4.56)	5.67	0.024
DNA-Net [14]	CNN-based	82.59	86.08	96.68	9.29	95.38 (⇑ 1.54)	93.41 (⇑ 0.80)	89.10 (⇑ 2.21)	4.70	0.095
GGL-Net [55]	CNN-based	81.69	84.74	97.21	7.78	95.24 (⇑ 1.68)	92.39 (⇑ 1.82)	87.99 (⇑ 3.32)	8.99	0.036
UIU-Net [34]	CNN-based	81.99	83.73	96.54	12.15	96.00 (⇑ 0.92)	91.69 (⇑ 2.52)	88.02 (⇑ 3.29)	50.54	0.055
MTU-Net [33]	Transformer-based	67.92	66.83	92.36	36.98	84.38 (⇑ 12.54)	80.61 (⇑ 13.60)	68.02 (⇑ 23.29)	12.75	0.028
RPCANet [32]	CNN-based	73.68	74.82	92.06	32.24	85.40 (⇑ 11.52)	82.58 (⇑ 11.63)	70.53 (⇑ 20.78)	0.68	1.853
SCTransNet [45]	Transformer-based	82.58	86.24	96.61	9.96	94.93 (⇑ 1.99)	92.94 (⇑ 1.27)	88.23 (⇑ 3.08)	11.32	0.048
ILNet [16]	CNN-based	82.89	81.02	96.41	**6.20**	95.96 (⇑ 0.96)	90.23 (⇑ 3.98)	86.59 (⇑ 4.72)	4.04	**0.020**
SDSNet [46]	Transformer-based	81.22	83.60	96.35	13.91	95.81 (⇑ 1.11)	92.84 (⇑ 1.37)	88.96 (⇑ 2.35)	2.55	0.029
MSDA-Net [58]	CNN-based	82.86	85.46	97.01	12.53	96.43 (⇑ 0.49)	93.04 (⇑ 1.17)	89.72 (⇑ 1.59)	4.79	0.035
	+ FDEP w/o CO-ISD	**84.71**	**86.57**	**98.14**	11.95	**97.25 (⇓ 0.33)**	**94.32 (⇓ 0.11)**	**91.73 (⇓ 0.42)**	5.24	0.100
	+ FDEP Framework	84.05	86.33	97.54	14.24	96.92	94.21	91.31	5.24	0.035

Table 13: Performance comparison of different methods on the NUAA-SIRST, NUDT-SIRST, and IRSTD-1k datasets. *"+ FDEP w/o CO-ISD"* denotes the FDEP without CO-ISD strategy. Red denotes the best result, and blue denotes the second best result.

Methods	Description	NUAA-SIRST (213:214)					NUDT-SIRST (663:664)					IRSTD-1K (800: 201)				
		IoU	P_d	HSE-P	HSE-T	HSE	IoU	P_d	HSE-P	HSE-T	HSE	IoU	P_d	HSE-P	HSE-T	HSE
FKRW	Filtering	22.84	77.19	35.72	62.73	22.41	14.70	58.10	21.74	37.70	8.20	9.25	54.21	12.49	33.40	4.17
MPCM	Local Contrast	42.92	88.59	61.20	62.18	38.05	12.55	64.34	22.42	10.57	2.37	21.36	81.48	35.35	11.36	4.02
NIPPS	Low Rank	26.82	80.23	61.40	54.71	33.59	18.66	68.89	54.29	42.13	22.87	14.93	67.34	27.18	46.72	12.70
ACM	CNN-based	67.07	91.63	87.25	89.23	77.86	67.31	97.88	86.31	85.22	73.55	63.26	90.24	83.86	86.70	72.71
ALCNet	CNN-based	70.64	92.02	90.56	87.09	78.87	72.73	96.19	91.20	88.86	81.04	67.15	90.91	88.82	86.97	77.24
MLCL-Net	CNN-based	76.37	96.96	93.80	88.79	83.28	94.47	98.73	99.35	96.73	96.11	66.07	90.24	84.62	85.72	72.53
ALCL-Net	CNN-based	76.52	95.82	93.44	90.02	84.11	94.25	99.15	99.38	94.25	93.66	67.35	92.26	86.49	86.11	74.48
DNA-Net	CNN-based	76.29	96.58	93.56	91.19	85.31	95.17	99.26	99.45	96.88	96.34	68.07	**94.28**	86.27	86.83	74.91
GGL-Net	CNN-based	79.46	97.34	94.64	**93.44**	**88.43**	94.68	98.94	99.25	97.26	96.53	67.68	89.56	87.06	86.84	75.60
UIU-Net	CNN-based	78.23	96.58	94.37	91.03	85.91	94.88	99.05	99.44	97.59	97.05	67.64	93.27	86.27	**88.73**	76.55
MTU-Net	Transformer-based	76.14	95.06	89.59	86.07	77.11	84.65	97.66	95.02	93.86	89.18	67.42	87.88	83.89	76.66	64.31
RPCANet	CNN-based	57.49	88.97	73.50	68.99	50.71	87.54	95.56	94.11	90.07	84.77	63.54	87.63	78.52	75.55	59.33
SCTransNet	Transformer-based	73.66	96.20	92.12	88.73	81.74	93.74	98.84	98.71	96.66	95.41	67.03	92.26	86.23	83.82	72.28
ILNet	CNN-based	77.32	95.44	94.86	87.66	83.15	91.74	98.62	98.79	94.76	93.61	68.25	93.60	86.78	85.90	74.54
SDSNet	Transformer-based	73.86	95.06	90.96	88.42	80.44	92.97	99.26	99.07	96.18	95.29	70.92	88.89	90.77	84.62	76.82
MSDA-Net	CNN-based	75.94	96.96	93.25	89.64	83.58	96.03	99.15	**99.54**	97.10	96.66	71.47	92.26	91.21	85.89	78.34
	+ FDEP w/o CO-ISD	**79.48**	**98.10**	**95.29**	92.35	88.00	96.01	99.05	99.43	**98.08**	**97.53**	**71.96**	93.27	**91.86**	86.07	79.06
	+ FDEP Framework	75.98	95.06	92.84	90.44	83.96	**96.50**	**99.37**	99.50	97.91	97.43	71.52	93.60	91.74	86.51	**79.37**

results using distillation loss show that adding an explicit distillation loss provides no benefit and slightly degrades performance. This indicates that explicit distillation tends to destabilize optimization in the extremely imbalanced SIRST task, whereas our CO-ISD strategy enables stable and efficient knowledge transfer. Meanwhile, these results suggest that CO-ISD is not a conventional teacher-student distillation process, but a gradient-level collaborative optimization strategy where VFM-induced semantics are transferred through shared deployable parameters. To further examine the interaction between the two branches, we retrain MSDA-Net equipped with FDEP on NUAA-SIRST and analyze their gradients with respect to the shared encoder-decoder parameters. The cosine similarity between the branch gradients increases from 0.550 in early training to 0.842 in late training, while their gradient norms remain of the same order of magnitude. These results suggest that, under the shared ground-truth detection objective, both branches contribute to shared parameter updates and gradually develop more consistent optimization directions. Together with the parameter-sharing

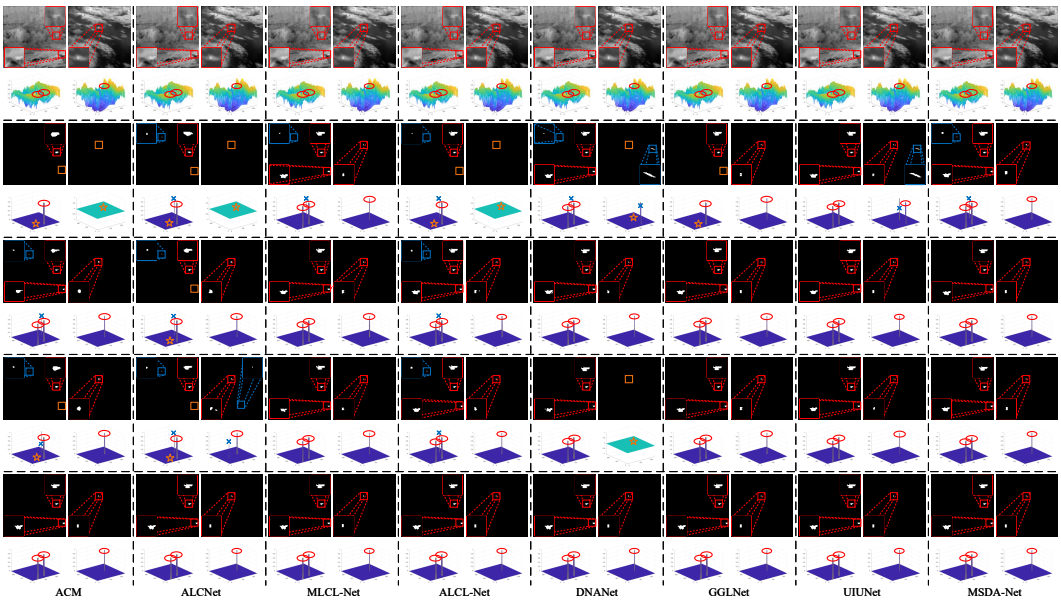

Figure 4: Visualization of several excellent methods on the SIRST3 dataset. Red: correct detections. Blue: false positives. Orange: missed detections. Every two rows from top to bottom: *Image*, *"Original"*, *"+ FDEP w/o CO-ISD"*, *"+ FDEP Framework"*, *True label*.

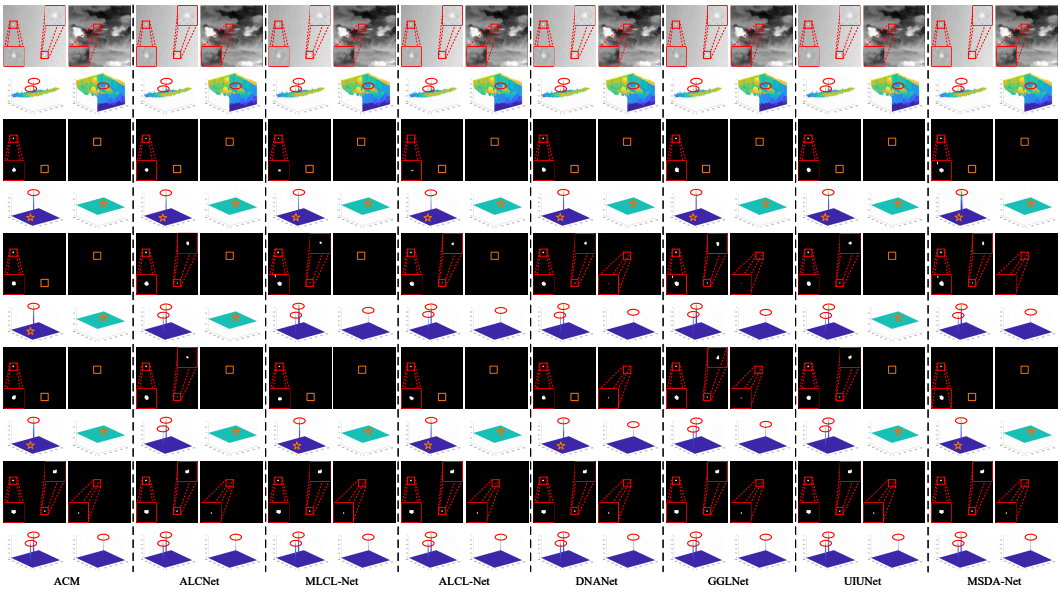

Figure 5: Visualization of some excellent methods on the NUAA-SIRST dataset. Red: correct detections; blue: false positives; orange: missed detections. Every two rows from top to bottom: *Image*, *"Original"*, *"+ FDEP w/o CO-ISD"*, *"+ FDEP Framework"*, *True label*.

ablations, they provide complementary optimization-level evidence supporting the collaborative training mechanism of CO-ISD.

B Compared with more SOTA Methods

To fully validate the performance of MSDA-Net equipped with the FDEP framework, we conduct detailed quantitative comparisons against a series of recent excellent methods on the SIRST3 [43, 37], NUAA-SIRST [5], NUDT-SIRST [14], and IRSTD-1K [49] datasets. In addition, we include three representative non-deep learning-based methods [27, 31, 4] to ensure the completeness and rigor of the evaluation.

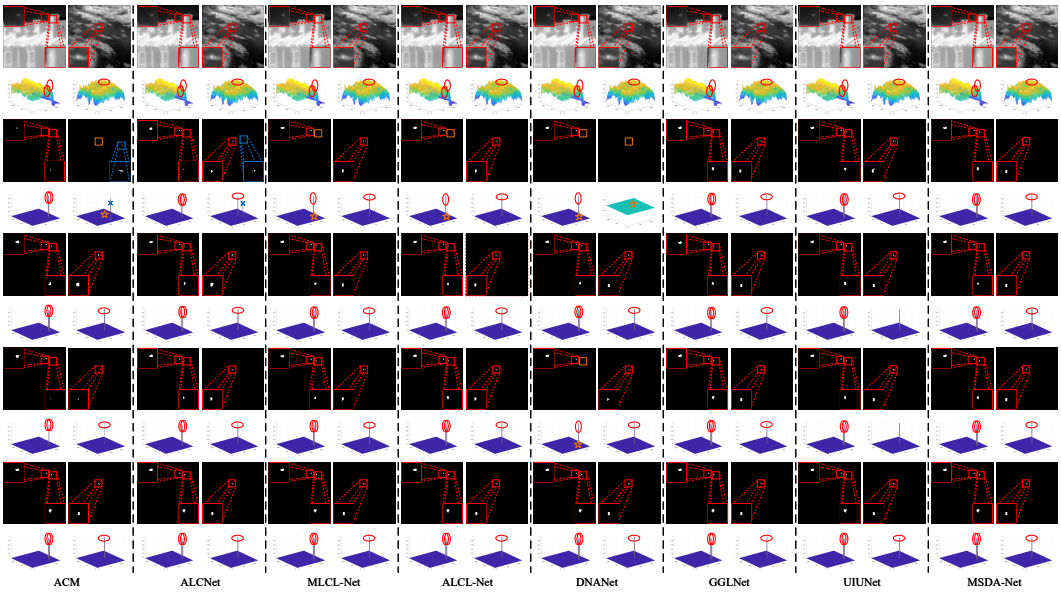

Figure 6: Visualization of some excellent methods on the NUDT-SIRST dataset. Red: correct detections; blue: false positives; orange: missed detections. Every two rows from top to bottom: *Image*, *"Original"*, *"+ FDEP w/o CO-ISD"*, *"+ FDEP Framework"*, *True label*.

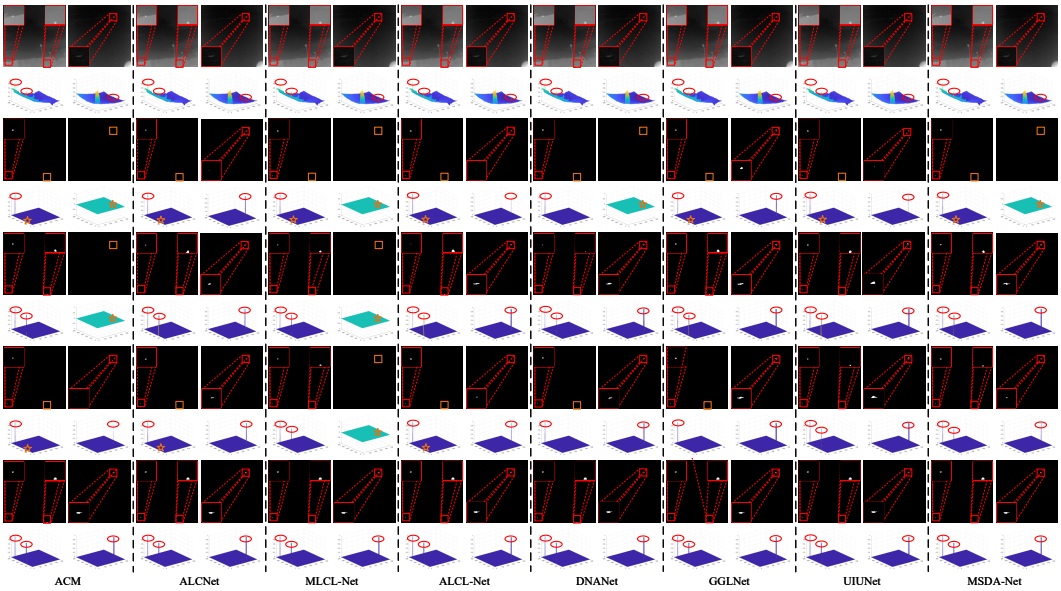

Figure 7: Visualization of some excellent methods on the IRSTD-1K dataset. Red: correct detections; blue: false positives; orange: missed detections. Every two rows from top to bottom: *Image*, *"Original"*, *"+ FDEP w/o CO-ISD"*, *"+ FDEP Framework"*, *True label*.

1) Evaluation on the SIRST3 dataset. The experimental results are presented in Tab. 12. First, compared with traditional non-deep learning-based methods, deep learning-based SIRST detection networks achieve significantly better performance. Second, compared to other non-MSDA-Net methods, MSDA-Net equipped with the FDEP framework achieves state-of-the-art (SOTA) performance. Specifically, it improves *HSE-P* by 0.92 (from 96.00 to 96.92) - 76.44 (from 20.48 to 96.92), improves *HSE-T* by 0.80 (from 93.41 to 94.21) - 80.71 (from 13.50 to 94.21), and improves the overall *HSE* by 2.21 (from 89.10 to 91.31) - 87.07 (from 4.24 to 91.31). Third, compared with MSDA-Net using *"+ FDEP Framework"*, the variant with *"+ FDEP w/o CO-ISD"* achieves higher accuracy, making it more suitable for scenarios requiring high precision. Although *"+ FDEP Framework"* yields slightly lower accuracy, it accelerates inference by 65% (from 0.100 s to 0.035 s) compared with *"+ FDEP w/o CO-ISD"*. Finally, compared with the original MSDA-Net, *"+ FDEP Framework"* improves

Table 14: Quantitative comparison of multiple cases in Fig. 3 of the Main Text.

Cases	Case 1		Case 2		Case 3		Case 4	
	S1	S2	S1	S2	S1	S2	S1	S2
AUC ↑	0.9993	0.9799	0.9987	0.9997	0.9989	0.9880	0.6970	0.6977
hIoU (%) ↑	16.67	19.00	27.43	27.43	18.87	38.10	15.93	38.52
HSE-P (%) ↑	45.37	89.75	14.24	36.25	31.95	91.27	17.04	64.46
HSE-T (%) ↑	33.33	100.00	36.67	36.67	45.00	100.00	21.67	66.67
HSE (%) ↑	**15.12**	**89.75**	**5.22**	**13.29**	**14.39**	**91.27**	**3.70**	**42.97**

HSE by 1.59 (from 89.72 to 91.31) without increasing inference cost, fully demonstrating the strong performance of the FDEP framework.

2) Evaluation on three individual datasets. From Tab. 13, compared to other non-MSDA-Net methods, MSDA-Net equipped with the FDEP framework achieves SOTA performance on both NUDT-SIRST and IRSTD-1K datasets. At the same time, although original MSDA-Net performs consistently well on NUDT-SIRST and IRSTD-1K, its performance on NUAA-SIRST is relatively weaker. However, after being equipped with *"+ FDEP w/o CO-ISD"*, MSDA-Net obtains a much more notable performance gain on NUAA-SIRST. This is because the original MSDA-Net lacks sufficient robustness with extremely limited training samples, whereas incorporating the semantic features from VFMs can effectively mitigate this issue. These results further demonstrate that systematically leveraging the semantic priors of VFMs is beneficial for improving SIRST detection performance, especially in challenging or extreme scenarios, highlighting the importance of integrating VFMs into the SIRST task. In addition, compared with the original MSDA-Net, both *"+ FDEP w/o CO-ISD"* and *"+ FDEP Framework"* yield significant improvements in the overall *HSE* metric, further validating the robustness of the proposed FDEP framework in few-shot scenarios.

C More Qualitative Analyses

To more intuitively demonstrate the effectiveness of the proposed FDEP framework under mixed or few-shot scenarios, we visualize detection results on a mixed dataset (SIRST3) and three individual datasets (NUAA-SIRST, NUDT-SIRST, and IRSTD-1K). These data cover diverse imaging conditions and target characteristics with few-shot samples. The results are shown in Figs. 4–7. Compared with the baseline model without FDEP, both *"+ FDEP w/o CO-ISD"* and *"+ FDEP Framework"* produce more complete target structures, clearer boundary contours, and significantly fewer missed and false detections across various scenes. This clearly indicates that the proposed FDEP framework can effectively enhance the model's sensitivity to small targets and improve its robustness under complex backgrounds and extreme class-imbalance conditions. At the same time, for the same SIRST network, the detection performance using *"+ FDEP w/o CO-ISD"* and *"+ FDEP Framework"* is similar, indicating that the CO-ISD strategy ensures that the lightweight branch can obtain semantic representations similar to the main branch while maintaining lightweight inference, making the detection results stable and reliable. Notably, compared to using *"+ FDEP w/o CO-ISD"*, using *"+ FDEP Framework"* reduces inference time overhead by 40.6% (from 0.160 to 0.095) to 74.4% (from 0.082 to 0.021) while maintaining competitive detection accuracy. In summary, the proposed FDEP framework not only significantly enhances the model's capability in small target discrimination and background suppression under few-shot conditions, but also delivers stable and consistent improvements in both visual quality and quantitative metrics, demonstrating strong potential for practical application.

D More Discussion

1) Conventional distillation methods fail in SIRST detection tasks. According to our experimental observations, when following a traditional model distillation process, that is, when pseudo-labels are generated by the model trained with the *"FDEP w/o CO-ISD"* version and then used to perform fully supervised training on the original SIRST network, the network will collapse directly in the early iteration stage. Even under conditions of joint supervision by real and pseudo labels, the training still collapses. The reason is that SIRST task has characteristics such as extreme class imbalance, small target size and low signal-to-noise ratio. The pseudo-labels generated in the traditional distillation often have semantic noise and confidence bias, which will introduce incorrect gradient signals during

Table 15: Multi-metric performance comparison on the SIRST3 dataset. Values in parentheses denote absolute *HSE* gains over the corresponding original detector. S@0.1 FPPI and S@1.0 FPPI denote FROC sensitivity at 0.1 and 1.0 false positives per image (FPPI), respectively. nFROC-AUC@1 FPPI denotes normalized FROC-AUC up to 1.0 FPPI.

Schemes	HSE ↑	hIoU ↑	ROC-AUC ↑	AP ↑	S@0.1 FPPI ↑	S@1.0 FPPI ↑	nFROC-AUC@1 FPPI ↑
ACM (Original)	68.05	55.05	98.69	81.86	36.11	96.74	87.42
+ FDEP Framework	75.81 (+7.76)	61.99	99.65	88.98	54.90	97.41	90.77
ALCNet (Original)	74.20	62.18	99.37	86.50	69.58	97.01	90.00
+ FDEP Framework	78.31 (+4.11)	67.06	98.83	88.92	94.68	96.35	91.56
MLCL-Net (Original)	86.80	75.55	98.74	93.84	96.15	97.87	93.37
+ FDEP Framework	87.20 (+0.40)	79.08	98.31	93.42	95.75	97.67	94.46
ALCL-Net (Original)	86.75	77.73	99.06	94.63	96.66	97.81	93.88
+ FDEP Framework	87.40 (+0.65)	79.27	99.19	95.28	96.74	97.54	92.16
DNA-Net (Original)	89.10	81.55	98.46	94.51	97.01	98.34	94.97
+ FDEP Framework	89.62 (+0.52)	81.38	98.49	94.53	97.80	98.41	95.36
GGL-Net (Original)	87.99	81.11	98.83	94.62	97.94	98.21	95.85
+ FDEP Framework	90.69 (+2.70)	82.05	99.19	95.95	97.61	98.07	95.55
UIU-Net (Original)	88.02	79.97	99.42	95.66	96.74	97.08	94.80
+ FDEP Framework	90.01 (+1.99)	80.61	99.41	96.19	97.48	97.87	95.50
MSDA-Net (Original)	89.72	80.55	99.50	96.13	97.27	98.07	94.86
+ FDEP Framework	91.31 (+1.59)	81.59	99.39	96.35	98.14	98.47	95.93

Table 16: Spearman rank correlations of *HSE* with other metrics across 16 configurations in Tab. 15 on the SIRST3 dataset. S@0.1/1.0 FPPI denotes FROC sensitivity at 0.1/1.0 false positives per image (FPPI). nFROC-AUC@1 FPPI denotes normalized FROC-AUC up to 1.0 FPPI.

Compared Metric	hIoU	ROC-AUC	AP	S@0.1 FPPI	S@1.0 FPPI	nFROC-AUC@1 FPPI
Spearman ρ	0.935	0.191	0.879	0.904	0.786	0.909

backpropagation, leading to instability or even collapse of the network optimization. Furthermore, conventional distillation assumes a stable semantic consistency between the teacher and student networks. However, in SIRST detection, the teacher model's features are dominated by global semantics derived from the VFMs, while the student network focuses on low-level textures and local contrast modeling. This semantic discrepancy between the two leads to misaligned feature spaces, making direct semantic matching infeasible. Such cross-domain semantic inconsistency causes the explicit distillation loss to act as an optimization disturbance rather than a benefit during training. The proposed CO-ISD strategy fundamentally avoids the instability caused by explicit distillation. Its implicit semantic transfer not only ensures the stability of the training process but also achieves a significant improvement in lightweight branch performance without additional inference costs.

2) Rationale and advantages of the HSE. Existing single metrics are insufficient to comprehensively characterize detection capability, and the current evaluation system is highly sensitive to threshold settings. To overcome the fragmentation and insufficiency of existing evaluation systems, we construct a Holistic SIRST Evaluation (*HSE*) metric, which performs multi-threshold integral assessment at both pixel-level confidence prediction and target-level robust detection perspectives, providing a more stable and more comprehensive criterion for fair model comparison and performance analysis. Notably, the proposed *HSE* fundamentally differs from *hIoU* [25]. The *hIoU* adopts hierarchical discrete decisions and emphasizes target-level results by considering pixel-level IoU only for correctly detected targets. In contrast, our *HSE* unifies pixel- and target-level evaluation via continuous threshold integration, jointly capturing pixel-level confidence stability and target-level robustness while avoiding single-threshold bias. As shown in Fig. 3 of the Main Text and Tab. 14 of the Appendix, *hIoU* has limited discriminative power both between and within sample pairs. For example, Case 2-S1 and Case 2-S2 differ in false positive area, but their *hIoU* scores are the same. Our *HSE* can more reasonably distinguish these differences. In addition, *HSE-P*, a component of *HSE*, can alleviate the limitations of the ROC curve. From the Fig. 3 of the Main Text and the Tab. 14 of the Appendix, the AUC of ROC curves in multiple cases does not faithfully reflect true pixel-level performance. *HSE-P* mitigates this issue. For instance, compared with Case 1-S2, Case 1-S1 contains severe false positives but still achieves a higher AUC. For *HSE-P*, Case 1-S1 has a lower value, which better aligns with the actual pixel-level quality.

3) Comparison with existing evaluation metrics. To systematically examine the relationship between *HSE* and existing evaluation metrics, we compare 16 configurations on the SIRST3 dataset, covering

holistic metrics (*HSE* and hIoU), pixel-level metrics (AP and ROC-AUC), and FROC-style target-level metrics. These configurations comprise eight backbone detectors and their corresponding FDEP-equipped variants. First, *HSE* shows high overall ranking consistency with existing metrics. From Tab. 16, except for ROC-AUC, the Spearman correlations between *HSE* and the other metrics range from 0.786 to 0.935, indicating broad agreement with existing holistic, pixel-level, and target-level metrics across the evaluated configurations. In contrast, the correlation between *HSE* and ROC-AUC is only 0.191. For example, ACM equipped with FDEP achieves the highest ROC-AUC (99.65) but substantially lower AP, FROC, and *HSE* scores than several other configurations, further supporting our discussion regarding the limitations of ROC-AUC in SIRST evaluation. Second, local ranking differences are interpretable. *HSE* jointly considers pixel-level confidence quality and target-level detection robustness, so its ranking need not be identical to that of any individual metric. From Tab. 15, "GGL-Net + FDEP Framework" achieves a slightly higher hIoU than "MSDA-Net + FDEP Framework" (82.05 vs. 81.59), whereas the latter obtains higher AP and FROC scores, while also achieving a higher *HSE* (91.31 vs. 90.69). This difference reflects their distinct evaluation criteria: hIoU relies on hierarchical discrete decisions, whereas *HSE* unifies pixel- and target-level evaluation through multi-threshold integration, jointly capturing pixel-level confidence quality and target-level robustness while reducing single-threshold bias. In addition, the overall effectiveness of FDEP is not solely supported by *HSE*. FDEP improves *HSE* on all eight backbones, hIoU and AP on seven, Sensitivity@0.1 FPPI and normalized FROC-AUC@1 FPPI on six, and Sensitivity@1.0 FPPI on four. These results indicate that the overall advantage of FDEP is also supported by multiple existing evaluation metrics, although improvements are not uniform across all metrics and backbones. Overall, *HSE* shows strong ranking agreement with hIoU, AP, and FROC-style metrics, while providing reasonable differences when pixel- and target-level performance are imbalanced.

E Broader Impacts

This paper focuses on SIRST detection, aiming to improve detection accuracy while maintaining inference efficiency by introducing frozen VFM representations. The proposed FDEP framework utilizes global semantic priors during the training phase and retains only lightweight branches during the inference phase, contributing to efficient and robust infrared perception. Related technologies can serve practical applications such as maritime monitoring, autonomous driving, environmental perception, and early warning, enhancing the detection capabilities of weak and distant targets in complex backgrounds. We hope this work will promote the application of VFMs in SIRST detection and provide technical support for the efficient deployment and reliable application of related methods.

F Limitations and Future Work

Although the proposed FDEP framework shows strong compatibility, can be seamlessly adapted to existing SIRST detection networks, and achieves stable performance improvements on multiple public datasets, it still has several limitations. On the one hand, although this work has been extensively validated on multiple public datasets, the available data are still limited and do not fully cover complex real-world deployment scenarios, such as extreme weather and low-quality imaging. Therefore, its generalization ability in broader practical environments requires further evaluation. On the other hand, FDEP relies on VFMs to provide global semantic priors during training, which introduces additional training cost. According to our statistics, on the largest SIRST3 dataset, the training times of the original MSDA-Net and MSDA-Net equipped with FDEP are 11.46 h and 19.52 h, respectively. Nevertheless, it is worth emphasizing that FDEP retains only the lightweight branch during inference, and thus does not introduce additional inference-time overhead. As a result, the two models maintain the same inference speed in practical applications. Future work will further explore more efficient VFMs adaptation and feature compression strategies, and construct more comprehensive real-world datasets to improve training efficiency, generalization ability, and deployment reliability.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions and scope of the paper are well summarized in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides sufficient reproducibility details, including datasets, backbone networks, evaluation metrics, training settings, default VFM, and ablation protocols. The key components, including SAMF and CO-ISD are also clearly described.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our code is available at **https://github.com/YuChuang1205/FDEP-Framework**. All datasets used in the paper are publicly available.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: In Sec. 4.2, we specify the main experimental settings, including data splits, evaluation metrics, optimizer, learning rate, batch size, training epochs, input crop size, GPU setting, and default VFM choice.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We did not report error bars or confidence intervals in the paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Sec. 4, we report the GPU type, parameter counts, and inference time and other relevant details. In Appendix F, we report the training time.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirmed that our research conforms to it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed the potential social impacts in Appendix E.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cited the original papers that produced the code packages and datasets used in this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our code is publicly available with documentation covering setup, data preparation, training, and evaluation. The repository is linked in the abstract. No new datasets are introduced.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method we developed is original and completely without LLM.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.