Rethinking Evaluation of Infrared Small Target Detection

Youwei Pang[1,3] **Xiaoqi Zhao**[2*] **Lihe Zhang**[1*] **Huchuan Lu**[1]
Georges El Fakhri[2] **Xiaofeng Liu**[2] **Shijian Lu**[3]
[1]Dalian University of Technology [2]Yale University [3]Nanyang Technological University
lartpang@gmail.com, xiaoqi.zhao@yale.edu, {zhanglihe,lhchuan}@dlut.edu.cn
{georges.elfakhri,xiaofeng.liu}@yale.edu, shijian.lu@ntu.edu.sg

Abstract

As an essential vision task, infrared small target detection (IRSTD)[1] has seen significant advancements through deep learning. However, critical limitations in current evaluation protocols impede further progress. First, existing methods rely on fragmented pixel- and target-level specific metrics, which fails to provide a comprehensive view of model capabilities. Second, an excessive emphasis on overall performance scores obscures crucial error analysis, which is vital for identifying failure modes and improving real-world system performance. Third, the field predominantly adopts dataset-specific training-testing paradigms, hindering the understanding of model robustness and generalization across diverse infrared scenarios. This paper addresses these issues by introducing a hybrid-level metric incorporating pixel- and target-level performance, proposing a systematic error analysis method, and emphasizing the importance of cross-dataset evaluation. These aim to offer a more thorough and rational hierarchical analysis framework, ultimately fostering the development of more effective and robust IRSTD models. An open-source toolkit has be released to facilitate standardized benchmarking. [2]

1 Introduction

Infrared small target detection (IRSTD) is critical in applications such as maritime resource management, navigation, and environmental monitoring [30, 3, 2, 36]. Infrared imaging leverages the contrast between target and background radiation, offering advantages such as working in all weather conditions and operating day and night. Detecting small and low-contrast targets in infrared image presents significant challenges due to complex backgrounds, long-distance transmission effects, and a low signal-to-noise ratio [4, 16]. Given the unique challenges, significant research has been devoted to developing algorithms that accurately capture and segment these targets.

Despite advances in detection methods, the evaluation protocols used to benchmark these algorithms remains a subject of concern. Current IRSTD practices rely on level-specific metrics, including both pixel-level IoU_{pix}, nIoU_{pix}, and F1_{pix}, and target-level Pd and Fa, alongside independent training and testing on individual datasets. These protocols often prioritize the incomplete evaluation in data-constrained scenarios, failing to deliver comprehensive and detailed model analysis. Such an inadequate evaluation leads to several issues. First, existing fragmented and coupled metrics result in a lack of holistic evaluation, making it difficult to fully show the model's true performance. Furthermore, the current research frequently overlooks systematic failure mode investigation in

*Corresponding Authors: Xiaoqi Zhao and Lihe Zhang

[1]In the existing literature, IRSTD can refer to both detection and segmentation tasks. However, **this paper primarily focuses on the more challenging segmentation task as in [20, 17, 37]**.

[2]Our evaluation toolkit: https://github.com/lartpang/PyIRSTDMetrics

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

its rush to report competitive performance benchmarks, which is a fundamental requirement for diagnosing model vulnerabilities and enhancing algorithmic robustness. Additionally, the evaluation protocol remains constrained by its dataset-specific training-testing paradigm. Such a widespread practice incentivizes narrow optimization for dataset biases rather than general detection capability. It not only increases overfitting risks but may also exaggerate perceived performance.

To address these challenges, we propose a comprehensive hierarchical analysis framework, which introduces three key components: 1) *a hybrid-level performance metric, hierarchical IoU (hIoU)*, 2) *a systematic error analysis method*, and 3) *a cross-dataset evaluation setting*. Specifically, the proposed hIoU combines the performance from target-level localization and pixel-level segmentation, offering a more holistic view of model efficacy. And the proposed error analysis method is closely tied and complementary to the proposed metric hIoU. It allows for a detailed exploration of model failure modes and identifying key cues for improving method effectiveness. Besides, our cross-dataset setting systematically evaluates model performance across different dataset scenarios, providing valuable measurements of robustness and generalization. By addressing these challenges, our work provide a thorough and rational evaluation framework, ultimately contributing to the advancement of more reliable and transferable IRSTD applications.

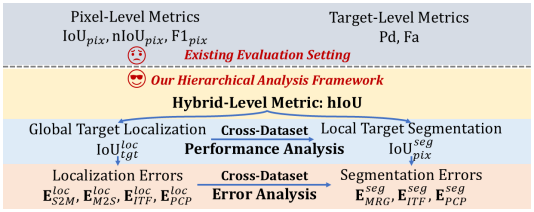


Figure 1: Our hierarchical analysis framework.

Our main contributions are summarized as follows:

- First to expose limitations in current IRSTD evaluation protocols and propose a hierarchical analysis framework.
- Introduce an hybrid-level metric capturing IRSTD performance across target and pixel levels.
- Reveal limited cross-dataset generalization of IRSTD algorithms through detailed evaluation.
- First to systematically analyze errors in IRSTD by quantifying model limitations under our metric.
- Develop a universal and comprehensive evaluation toolkit to advance IRSTD research.

2 Related Work

After the development of several decades, the IRSTD algorithm design has undergone a significant evolution from traditional methods to deep learning techniques [4, 16]. Traditional algorithms [1, 15, 31, 10, 13, 26, 14, 27, 12, 8, 7, 38, 43, 39] rely on filtering techniques and model-driven approaches, which perform well in simple backgrounds but lacked robustness in complex environments. Recent advances [9, 40, 17, 41, 42, 34, 29, 33, 22, 6, 20, 44, 37, 32, 18, 35, 19] based on deep learning have significantly propelled IRSTD.

Pioneering work [9] introduces asymmetric contextual modulation to enhance target-background discrimination while addressing data scarcity via a high-quality annotated dataset. Subsequent studies focus on specialized architectures. [17] designs dense nested interaction modules with dual attention to preserve targets across network depths, whereas [40] proposes feature compensation and cross-level correlation mechanisms to recover lost target details. For shape-aware detection, [41] integrates Taylor finite difference operators and orientation attention to capture geometric characteristics of targets. And [44] emphasizes lightweight multi-receptive field perception and feature fusion. Attention mechanisms became pivotal in later innovations. [42] developes pyramid context modules with global-local attention for complex backgrounds, while [34] embeds nested U-Nets with resolution-maintenance supervision to enhance multi-scale contrast. And [6] employs selective rank-aware attention to resolve hit-miss trade-offs. Transformer-based architectures emerge as powerful alternatives. [22] combines bilinear correlation and dilated convolutions for semantic refinement. [37] introduces spatial-channel cross-transformers to model full-level semantic differences. Besides, the loss function design proves critical. Scale- and location-sensitive losses [20] are developed to address target scale and location variations. Efforts to balance performance and interpretability are also explored by [32] which unfolds robust PCA into a deep network.

Despite these algorithmic advancements, the field has predominantly focused on architectural innovations while neglecting improvements in evaluation frameworks. Our work shifts the focus from

algorithm design to evaluation pipeline. And our framework provides actionable insights for real-world deployments, complementing rather than competing with existing algorithmic advancements.

3 Evaluation Metrics

3.1 Preliminaries

The existing deep learning-based IRSTD approaches utilize the hierarchical encoder-decoder[28, 21] architecture to process a dataset with K pairs of thermal infrared images $\{I_i\}_{i=1}^{K}$ and the corresponding ground truth (GT) binary masks $\{G_i\}_{i=1}^{K}$. In the pipeline, gray-scale predictions $\{P_i\}_{i=1}^{K}$ will be generated for these input images by the deep model in an end-to-end manner. The pixel values of P_i represent the probabilities of pixels belonging to infrared small targets within the i^{th} input image, and they range from 0 to 1. 0 indicates that the pixel is classified as a background pixel, whereas the value 1 signifies that the pixel is considered to be part of the foreground target. Gray-scale predictions $\{P_i\}_{i=1}^{K}$ are directly compared with their corresponding binary GT masks $\{G_i\}_{i=1}^{K}$ to calculate similarity. Thus, the average performance on the dataset can be calculated to evaluate the proposed algorithm. Currently, IRSTD metrics can be broadly classified into pixel-level and target-level categories according to their computational primitives. The following sections will provide detailed introductions.

3.2 Pixel-level Metrics

In this branch, existing metrics are all computed based on the statistical values the number of pixel-level true positives (TP_{pix}), false positives (FP_{pix}), true negatives (TN_{pix}), and false negatives (FN_{pix}) from the binary confusion matrix. Note that TP_{pix}, FP_{pix}, TN_{pix}, and FN_{pix} rely on the binary predictions and GTs with the same size. Therefore, the binarization strategy applied to the gray-scale predictions also influences the results of these metrics. However, the strategy design is beyond the scope of this work. Unless otherwise specified, a commonly-used threshold of 0.5 will be used by default for the prediction thresholding. The pixel-level metrics involved in existing works include Intersection over Union (IoU), and F1-score (F1) [3].

Intersection over Union (IoU) is a widely-used metric for measuring the overlap between the predicted foreground and the GT mask, normalizing their intersection cardinality with respect to the union. In existing works, there exist two different variants according to the difference of computational logic, including the conventional IoU (IoU_{pix}) and the normalized IoU (nIoU_{pix}) [9] as follows:

$$\text{IoU}_{pix} = \frac{\sum_{i=1}^{K} |\text{TP}_{pix}^{[i]}|}{\sum_{i=1}^{K} (|\text{TP}_{pix}^{[i]}| + |\text{FP}_{pix}^{[i]}| + |\text{FN}_{pix}^{[i]}|)} \quad \text{nIoU}_{pix} = \sum_{i=1}^{K} \frac{|\text{TP}_{pix}^{[i]}|/K}{|\text{TP}_{pix}^{[i]}| + |\text{FP}_{pix}^{[i]}| + |\text{FN}_{pix}^{[i]}|} \tag{1}$$

where $|\cdot|$ and $[i]$ represent the number of elements in the set and the sample index, respectively. IoU_{pix} aggregates global pixel statistics across all samples. nIoU_{pix} computes per-sample IoU before averaging, which ensures equal contribution from all samples, thereby fairly evaluating performance across diverse targets. When there is only one single sample (*i.e.*, $K = 1$), $\text{nIoU}_{pix} = \text{IoU}_{pix}$.

F1-score (F1_{pix}) is the harmonic mean of precision (Pre_{pix}) and recall (Rec_{pix}), designed to balance these metrics and provide a comprehensive evaluation. The calculation framework is defined as:

$$\text{Pre}_{pix} = \frac{\sum_{i=1}^{K} |\text{TP}_{pix}^{[i]}|}{\sum_{i=1}^{K} |\text{TP}_{pix}^{[i]}| + |\text{FP}_{pix}^{[i]}|} \quad \text{Rec}_{pix} = \frac{\sum_{i=1}^{K} |\text{TP}_{pix}^{[i]}|}{\sum_{i=1}^{K} |\text{TP}_{pix}^{[i]}| + |\text{FN}_{pix}^{[i]}|} \tag{2}$$

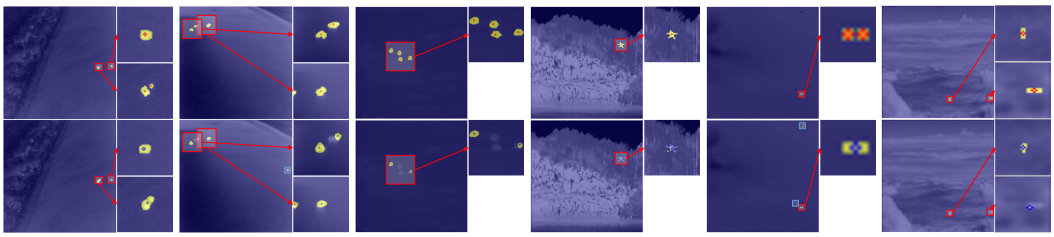
$$\text{F1}_{pix} = \frac{2\text{Pre}_{pix} \times \text{Rec}_{pix}}{\text{Pre}_{pix} + \text{Rec}_{pix}} \tag{3}$$

where Pre_{pix} and Rec_{pix} are primarily used to calculate the metric F1 and not used independently.

3.3 Target-level Metrics

Target-level metrics are designed to evaluate model performance at the level of individual targets and play an important role in IRSTD [17]. Unlike pixel-level metrics that aggregate performance across

[3] While IoU and F1 fundamentally measure region overlap, their standard implementation in IRSTD prioritizes global pixel-level evaluation over target-level alignment.



(a) (b) (c) (d) (e) (f)

Subfigure	IoU_{pix} ↑	$nIoU_{pix}$ ↑	$F1_{pix}$ ↑	Pd↑		Fa×10^6 ↓		hIoU↑	IoU^{seg}_{pix} ↑	E^{seg}_{ARG} ↓	E^{seg}_{TFF} ↓	E^{seg}_{PCP} ↓	IoU^{loc}_{tgt} ↑	E^{loc}_{S2M} ↓	E^{loc}_{M2S} ↓	E^{loc}_{TFF} ↓	E^{loc}_{PCP} ↓
				Ori. [17]	OPDC	Ori. [17]	OPDC	OPDC	OPDC	OPDC	OPDC	OPDC	OPDC	OPDC	OPDC	OPDC	OPDC
(a)	0.573	0.573	0.728	0.333	0.667	785.828	0.000	0.373	0.560	0.052	0.246	0.142	0.667	0.333	0.000	0.000	0.000
(b)	0.438	0.438	0.610	0.333	1.000	1457.214	198.364	0.344	0.459	0.000	0.225	0.316	0.750	0.000	0.000	0.250	0.000
(c)	0.310	0.310	0.473	0.500	0.500	0.000	0.000	0.300	0.600	0.000	0.000	0.400	0.500	0.000	0.000	0.000	0.500
(d)	0.477	0.477	0.646	1.000	1.000	579.834	625.610	0.005	0.036	0.000	0.000	0.964	0.143	0.000	0.857	0.000	0.000
(e)	0.514	0.514	0.679	0.500	0.500	244.141	244.141	0.118	0.474	0.474	0.053	0.000	0.250	0.250	0.000	0.500	0.000
(f)	0.326	0.326	0.492	1.000	1.000	0.000	0.000	0.313	0.313	0.000	0.085	0.602	1.000	0.000	0.000	0.000	0.000

Figure 2: Comparison of different metrics. Red and blue boxes to highlight the target regions. Red and blue points indicate the target centroids in ground truth (GT) masks and predictions, respectively. Zoom in on the digital color version for details. "Ori." [17] and "OPDC" refer to the original distance-based strategy and the proposed OPDC strategy for target matching.

the entire image, target-level metrics focus on the localization quality of predictions for each distinct small target. And target region sets $\{T_G^m\}$ and $\{T_P^n\}$ are extracted by the connected component analysis algorithm from the GT masks and binary predictions, respectively. Each individual predicted target T_P^n tries to match a nearby target T_G^m in the GT mask based on a specific criteria. Existing works [17, 22, 33] usually use *the centroid distance* to determine the matching relationship. If the distance between T_P^n and T_G^m is less than the predefined threshold (*e.g.*, 3 [17]), they will be viewed as the matched pair. Every predicted target matches to at most one GT target, and vice versa. Eventually, the predicted target regions matched to GT targets can be considered as belonging to target-level TP (TP_{tgt}), while the remaining T_P and T_G belong to FP_{tgt} and FN_{tgt}, respectively.

Probability of Detection (Pd). It quantifies the model's ability to detect true targets by computing the fraction of correctly matched target predictions (the count of TP_{tgt}) over the total number of GT targets (TP_{tgt} + FN_{tgt}) as follows:

$$\text{Pd} = \frac{\sum_{i=1}^{K} |\text{TP}_{tgt}^{[i]}|}{\sum_{i=1}^{K} (|\text{TP}_{tgt}^{[i]}| + |\text{FN}_{tgt}^{[i]}|)} \tag{4}$$

It effectively evaluates detection completeness for small targets regardless of their pixel area size.

False-Alarm Rate (Fa). It evaluates the spatial impact of unmatched predictions by normalizing the count of pixels in all FP_{tgt} targets against the entire image resolution:

$$\text{Fa} = \frac{\sum_{i=1}^{K} \texttt{TargetAreaOf}(\text{FP}_{tgt}^{[i]})}{\sum_{i=1}^{K} \texttt{ImageAreaOf}(G_i)} \tag{5}$$

where the normalization stabilizes comparisons across varying resolutions.

3.4 Current Limitations

Pixel-level Evaluation. Although IoU_{pix}, nIoU_{pix}, and F1_{pix} mitigate foreground imbalance by incorporating relative area proportions, they are inherently limited to global pixel-wise analysis and cannot assess target-level spatial localization or segmentation accuracy, both of which are critical to understanding the fine-grained IRSTD performance.

Target-level Evaluation. Pd measures detection completeness but ignores false positives, potentially overestimating performance in noisy environments. Fa quantifies the spatial impact of FP_{tgt} by normalizing their total area to the full image. But it maybe undercount smaller FP_{tgt} targets while overpenalizing large FP_{tgt} targets. Additionally, Fa disregards pixel-level errors (FP_{pix}) within TP_{tgt} targets and target-level errors FN_{tgt} targets. Although conventional Pd and Fa may compensate for each other's blind spots, their shared dependency on the distance threshold and data biases related to size and shape limit their ability to holistically reflect algorithm performance.

Fragmented Evaluation Paradigm. Current pipelines rely on several pixel-level (IoU_{pix}, nIoU_{pix}, and F1_{pix}) and target-level (Pd and Fa) metrics to approximate holistic performance. The former lacks spatial awareness, while the latter oversimplifies error patterns. While they individually address specific aspects, their simple combination creates a mismatch between evaluation pipeline and task requirements. Such fragmented paradigm obscures critical trade-offs, such as segmentation-localization dependency and target diversity tolerance, leaving incomplete or even contradictory performance insights. A new evaluation framework, considering hybrid-level modeling and data diversity, is essential to overcome these limitations.

4 New Evaluation Framework

4.1 Target Matching Strategy

Current target-level IRSTD metrics suffer from overly strict distance-only filtering [17], where centroid matching frequently misjudges offset, fragmented, or connected predictions as shown in Fig. 2a and Fig. 2b. By introducing overlap-priority constraint to enhance the matching mechanism, we propose the "Overlap Priority with Distance Compensation" (OPDC) strategy (Alg. 1), which can effectively alleviates these limitations.

Overlap priority constraint enforces shape coherence by computing pairwise overlap ratios between targets from GTs and predictions. For each target pair, if their IoU exceeds 0.5, it is marked as a valid candidate. The commonly-used assignment algorithm [5] is then applied to the full centroid Euclidean distance matrix to find the minimum-cost matching, followed by retaining only valid pairs satisfying the overlap constraint. This phase ensures morphological alignment by filtering mismatches caused by unreasonable shapes or over-prediction as the smallest target in Fig. 2a, resulting in initial matched pairs S_{TP} and unmatched sets S_{FN}/S_{FP}.

Distance-based compensation supplements residual unmatched pairs by evaluating centroid Euclidean distances. For targets in S_{FN} and S_{FP}, pairs with distances below the strict threshold (3 pixels, as in [17]) are re-matched via the assignment algorithm [5]. This phase specifically addresses small or low-overlap targets where shape metrics may underperform, leveraging spatial proximity as a secondary criterion to reassess their value.

By hierarchically integrating overlap and distance constraints, OPDC achieves a more intuitive matching. The former prioritizes overlap-based filtering to relax the original distance constraint, aligning with the real-world intuition where high overlap inherently indicates true morphological correspondence. The latter acts as a safety net exclusively for low-overlap residuals, ensuring spatial proximity without compromising the dominance of shape-aware matching. As shown in Fig. 2f, the right predicted target that do not satisfy the overlap constraint are re-matched with the GT target in the distance compensation.

4.2 Hierarchical Intersection over Union

Evaluation practices in current IRSTD studies typically focus on either isolated pixel-level or target-level similarity measurements between predictions and GTs. However, we propose a new hybrid-level metric, *i.e.*, hierarchical Intersection over Union (hIoU), which hierarchically combines both global target-level localization and local pixel-level segmentation performance as follows:

$$\text{IoU}_{tgt}^{loc} = \frac{\sum_{i=1}^{K} |\text{TP}_{tgt}^{[i]}|}{\sum_{i=1}^{K} |\text{TP}_{tgt}^{[i]}| + |\text{FP}_{tgt}^{[i]}| + |\text{FN}_{tgt}^{[i]}|} \quad \text{IoU}_{pix}^{seg} = \frac{\sum_{(T_G^m, T_P^n) \in \text{TP}_{tgt}} (T_G^m \cap T_P^n)/(T_G^m \cup T_P^n)}{\sum_{i=1}^{K} |\text{TP}_{tgt}^{[i]}|} \tag{6}$$

$$\text{hIoU} = \text{IoU}_{tgt}^{loc} \times \text{IoU}_{pix}^{seg} \tag{7}$$

where $(T_G^m \cap T_P^n)/(T_G^m \cup T_P^n)$ denotes the intersection-over-union ratio of the matched predicted target T_P^n and GT target T_G^m from TP_{tgt}. IoU_{tgt}^{loc} and IoU_{pix}^{seg} reflect the IoU-based localization and segmentation performance in infrared small target prediction, respectively. Specifically, we first adjust the IoU metric to measure the target-level localization performance IoU_{tgt}^{loc}. And then, for those predicted targets matched with GT targets, IoU_{pix}^{seg} is further applied to measure the local fine-grained segmentation similarity between them and the corresponding GT targets. The multiplicative

combination in hIoU inherently balances localization and segmentation: a method missing targets (low IoU_{tgt}^{loc}) or producing imprecise regions (low IoU_{pix}^{seg}) will be penalized proportionally. Unlike additive combinations that allow for linear error compensation (*e.g.*, high localization scores masking poor segmentation), this coupling measures the joint performance in the unit space $[0, 1]^2$, which makes improvements in one part necessary to maintain the other's performance. For example, achieving hIoU>0.64 requires both IoU_{tgt}^{loc} and IoU_{pix}^{seg} to exceed 0.64, whereas additive scoring would accept imbalanced solutions like $(0.9 + 0.38)/2 = 0.64$. This hybrid-level paradigm enables more comprehensive and nuanced performance evaluation, particularly crucial for the IRSTD task where *both complete localization (preventing target loss)* and *precise region delineation (ensuring target integrity)* are equally valuable for practical applications. Further discussions can be found in Sec. C.2.

4.3 Error Analysis Method

Existing protocols for evaluating IRSTD typically focus on overall average performance values. This analytical preference obscures critical failure modes and makes it difficult to diagnose and improve model deficiencies. For instance, a low IoU_{pix} and Pd could stem from background noise interference, adjacent target merging, or target perception limitations—each requiring distinct corrective strategies. To address this, building on our hierarchical evaluation paradigm as stated in Sec. 4.2, we categorize prediction errors into two associated levels: **target-level localization errors** ($\mathbf{E}_{S2M}^{loc}$, $\mathbf{E}_{M2S}^{loc}$, $\mathbf{E}_{ITF}^{loc}$, and $\mathbf{E}_{PCP}^{loc}$) and **pixel-level segmentation errors** ($\mathbf{E}_{MRG}^{seg}$, $\mathbf{E}_{ITF}^{seg}$, and $\mathbf{E}_{PCP}^{seg}$). They provide fine-grained decomposition of the performance losses reflected in IoU_{tgt}^{loc} (target localization IoU) and IoU_{pix}^{seg} (target segmentation IoU) metrics, respectively. Critically, the total error for each level is complementary to its corresponding IoU metric:

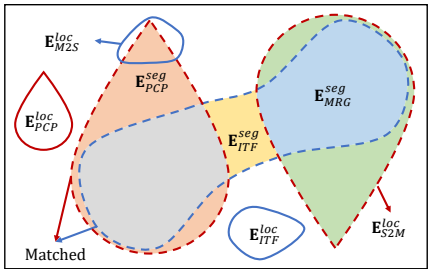


Figure 3: Error types (Sec. 4.3) for three predicted and three GT targets. Blue contours denote predictions and red contours denote GT. Under our OPDC strategy, only the middle prediction is matched to a GT and all others remain unmatched.

$$\text{Total Localization Error } \mathbf{E}^{loc} = 1 - \text{IoU}_{tgt}^{loc} \quad \text{Total Segmentation Error } \mathbf{E}^{seg} = 1 - \text{IoU}_{pix}^{seg} \tag{8}$$

Such a design establishes an explicit error-accuracy complementary relationship, where our error subtypes illustrated in Fig. 3 quantify distinct sources of deviation from the ideal performance of 1.

Target-level localization errors quantify mismatches and misidentifications in perceiving individual targets, including $\mathbf{E}_{S2M}^{loc}$, $\mathbf{E}_{M2S}^{loc}$, $\mathbf{E}_{ITF}^{loc}$, and $\mathbf{E}_{PCP}^{loc}$.

- $\mathbf{E}_{S2M}^{loc}$ (Single-to-Multi Mismatch): During the matching process, when a single predicted target satisfies overlap/distance constraints with multiple GT targets but ultimately gets assigned to only one GT target, the remaining unmatched GT targets contribute to this localization error. This is commonly observed in dense target clusters where a single prediction overlaps ambiguously with multiple GT targets, as illustrated in Fig. 2a and Fig. 2e.
- $\mathbf{E}_{M2S}^{loc}$ (Multi-to-Single Mismatch): When multiple predicted targets satisfy matching constraints with a single GT target, the one-to-one matching protocol forces retention of only one prediction. The excluded and unmatched competing predictions then become contributors to this localization error. This reflects model oversensitivity or fragmented predictions near valid targets, as demonstrated in Fig. 2d.
- $\mathbf{E}_{ITF}^{loc}$ (Interference Error): It corresponds to other false target predictions with no corresponding GT, which are primarily triggered by background clutter or noise artifacts, as marked by blue boxes in Fig. 2b and Fig. 2e.
- $\mathbf{E}_{PCP}^{loc}$ (Perception Error): It quantifies undetected GTs that fail to meet matching criteria, as exemplified in Fig. 2c. This indicates model insensitivity to low-contrast or morphologically variable targets.

Pixel-level segmentation errors capture neighboring interferences and shape distortions in matched targets, including $\mathbf{E}_{MRG}^{seg}$, $\mathbf{E}_{ITF}^{seg}$, and $\mathbf{E}_{PCP}^{seg}$.

Algorithm 1: OPDC Matching Strategy

Input: $\{T_G^m\}_m^M$: the mask set of M targets extracted from GT map G; $\{T_P^n\}_n^N$: the mask set of N targets extracted from prediction map P; MAX: a extremely large value used to avoid the algorithm choosing irrational matches;

Output: S_{TP}: the matched index pair set; S_{FN}: the unmatched GT target index set; S_{FP}: the unmatched predicted target index set;

// 1. Overlap Priority Constraint

Valid indicator $\mathbb{I} \in \mathbb{R}^{M\times N}, \mathbb{I}_{m,n} \in \{0,1\}, \mathbb{I}_{m,n} = 0, \forall m, n$;

Distance matrix $D \in \mathbb{R}^{M\times N}, D_{m,n} = \text{MAX}, \forall m, n$;

for $m = 1$ **to** M **do**

  for $n = 1$ **to** N **do**

    $D_{m,n} = \texttt{EuclideanDistance}(T_G^m, T_P^n)$;

    if $|T_G^m \cap T_P^n|/|T_G^m \cup T_P^n| \geq 0.5$ **then** $\mathbb{I}_{m,n} = 1$;

Initial match $A = \texttt{Assignment}(D)$; *// `scipy.optimize.linear_sum_assignment` [5]*

$S_{TP}, S_{FN}, S_{FP} \leftarrow A \cap \mathbb{I}$; *// Consider only pairwise relations that satisfy constraints.*

// 2. Distance-based Compensation

Valid indicator $\hat{\mathbb{I}} \in \mathbb{R}^{|S_{FN}|\times|S_{FP}|}, \hat{\mathbb{I}}_{m,n} \in \{0,1\}, \hat{\mathbb{I}}_{m,n} = 0, \forall m, n$;

Distance matrix $\hat{D} \in \mathbb{R}^{|S_{FN}|\times|S_{FP}|}, \hat{D}_{m,n} = \text{MAX}, \forall m, n$;

for $m = 1$ **to** $|S_{FN}|$ **do**

  for $n = 1$ **to** $|S_{FP}|$ **do**

    if $D_{S_{FN}^m, S_{FN}^n} < 3$ **then** *// Following the setting in [17].*

      $\hat{D}_{m,n} = D_{S_{FN}^m, S_{FN}^n}$;

      $\hat{\mathbb{I}}_{m,n} = 1$;

Compensation match $\hat{A} = \texttt{Assignment}(\hat{D})$; *// `scipy.optimize.linear_sum_assignment` [5]*

$\hat{S}_{TP}, \hat{S}_{FN}, \hat{S}_{FP} \leftarrow \hat{A} \cap \hat{\mathbb{I}}$;

$S_{TP} = S_{TP} \cup \hat{S}_{TP}$; *// Construct final matched index pair set.*

$S_{FN} = S_{FN} \setminus \hat{S}_{TP}$; *// Construct final unmatched GT target index set.*

$S_{FP} = S_{FP} \setminus \hat{S}_{TP}$; *// Construct final unmatched predicted target index set.*

- $\mathbf{E}_{MRG}^{seg}$ (Merging Error): This error quantifies false positive pixels within the regions of the GT targets that are not matched to the current predicted target. It specifically occurs when a target prediction extends into neighboring GT target regions, particularly when adjacent targets are erroneously merged, as shown in Fig. 2a and Fig. 2e.
- $\mathbf{E}_{ITF}^{seg}$ (Interference Error): It represents incorrectly predicted foreground pixels in real background region within target neighborhood, highlighting local noise interference or incomplete background suppression, as exemplified in Fig. 2a and Fig. 2b.
- $\mathbf{E}_{PCP}^{seg}$ (Perception Error): It accounts for the missing prediction within the region of the matched GT target, reflecting model uncertainty in delineating targets with faint edges or complex structures, as illustrated in Fig. 2d.

5 Experiment

5.1 Experimental Setup

Datasets. We evaluate IRSTD methods on three widely-used datasets: IRSTD1k [41], SIRST [9], and NUDT [17]. These datasets exhibit inherent diversity and complexity [6], enabling comprehensive evaluation of both within-dataset performance and cross-dataset generalization.

Metrics. We adopt several existing standard metrics, *i.e.*, IoU_{pix}, nIoU_{pix}, F1_{pix}, Pd, and Fa. Besides, we introduce the proposed hIoU to holistically evaluate localization-segmentation joint performance. For error analysis, we decompose hIoU into two components, *i.e.*, IoU_{tgt}^{loc} and IoU_{pix}^{seg}, and further break down their error statistics into fine-grained aspects based on a set of error subtypes.

Implementation Details. To ensure fair comparison across methods, we retrain all approaches using their official codebases under strictly controlled settings. All models are optimized via Adam

optimizer with an initial learning rate of 0.0005 and a multi-step decay schedule. They are trained for 400 epochs using a batch size of 16 on two NVIDIA GeForce RTX 3090 GPUs with a total of 48 GB memory. Following the existing practices [20], we introduce random horizontal flipping, cropping, and blurring to mitigate the overfitting risk. Besides, all images are bilinearly interpolated to 256 $\times$ 256 resolution during both training and testing phases.

5.2 Results and Discussion

We conduct a comprehensive benchmarking study of 14 recent deep learning-based[4] IRSTD methods [9, 40, 17, 41, 42, 34, 29, 33, 22, 6, 20, 44, 37, 32] through and dual evaluation protocols: conventional metrics and our hierarchical analysis framework. This analysis uniquely addresses two critical gaps in existing research: 1) over-optimistic single-dataset evaluation and 2) lack of detailed error analysis. By introducing cross-dataset setting and fine-grained error decomposition, we reveal previously obscured performance limitations.

Holistic Performance. Existing metrics show weak alignment in practical performance evaluation. As exemplified in Tab. 1, MSHNet [20], achieving top scores in conventional metrics (IoU_{pix}, F1_{pix} and Fa), fails to translate these advantages into superior holistic performance (hIoU). While DNANet [17] does not stand out in conventional metrics, its more balanced segmentation and localization performance propels it to hIoU leadership (0.557 *vs.* MSHNet's 0.549), as shown in Tab. 2. This inconsistency stems from their narrow focus on isolated aspects (pixel-level segmentation accuracy, target-level detection recall, *etc.*) without adequately modeling task hierarchy. Our hIoU demonstrates important value by reconciling these different aspects.

OPDC-Based Matching. To show OPDC's effects, we recalculated Pd and Fa metrics under this strategy ("+OPDC") as listed in Tab. 1. The observed recall improvement reveals that the original distance-based criterion overly constrained matching, discarding predictions with valuable positional reference for target localization. As shown in Fig. 2, when predictions span multiple adjacent GT targets (Fig. 2a and Fig. 2e) or partially intersect isolated targets (Fig. 2b), they fail centroid distance checks under the current protocol, but provide key location cues. OPDC alleviates these limitations through a dual-constraint mechanism that prioritizes overlap constraint as the primary matching trigger, with centroid distance constraint acting as supplementary validators as stated in Sec. 4.1.

Cross-Dataset Generalization. Cross-dataset analysis in Tab. 1 exposes critical generalization limitations. Most models exhibit severe performance degradation when they are tested on other datasets beyond their training one. Notably, IRSTD1k-trained models achieve better performance on SIRST_{TE} than on IRSTD1k_{TE}. And some IRSTD1k-trained models (*e.g.*, MRF3Net with 0.694 hIoU) even surpass counterparts trained on SIRST_{TR} (UIUNet: 0.679 hIoU; RPCANet: 0.675 hIoU) when evaluated on SIRST_{TE}. This phenomenon may be attributed to IRSTD1k primarily focuses on sky-background scenarios from SIRST, while introducing additional challenging ground scenarios (*e.g.*, woods and buildings). These complexities make IRSTD1k_{TE} inherently more difficult than SIRST_{TE}. Current single-dataset evaluation frameworks fail to expose such critical performance limitations. Establishing cross-dataset evaluation protocols would not only drive the community to prioritize model generalization across diverse scenarios, but also incentivize dataset creators to enhance scenario and target diversity. This paradigm shift addresses the current oversight in robustness validation and fosters progress in domain adaptation research.

Dense Multi-Object Scenarios. As shown in Fig. 4, these errors are predominantly dominated by interference- and perception-related terms. More specifically, interference item dominates the localization error, while perception item dominates the segmentation error. The former reflects the limited capability of current algorithms in suppressing background interference. The latter is closely tied to the intrinsic challenges of infrared small targets, such as blurred boundaries and complex structure patterns, which may lead to insufficient holistic perception during prediction. Furthermore, the analysis reveals that matching errors (*i.e.*, $\mathbf{E}^{loc}_{S2M}$ and $\mathbf{E}^{loc}_{M2S}$) and prediction merging errors (*i.e.*, $\mathbf{E}^{seg}_{MRG}$), often arising in dense multi-target scenarios or cases with complex target structures, exhibit relatively low proportions due to current data limitations. However, experiments in Sec. B using synthetic data reveal that they still make up a significant portion of total error. Their impact cannot be overlooked, as they represent critical failure modes in real-world applications. For instance, single-to-multi or multi-to-single errors may disrupt downstream tasks like counting, and merged

[4]Deep learning approaches are prioritized due to their demonstrated superiority in handling the IRSTD task.

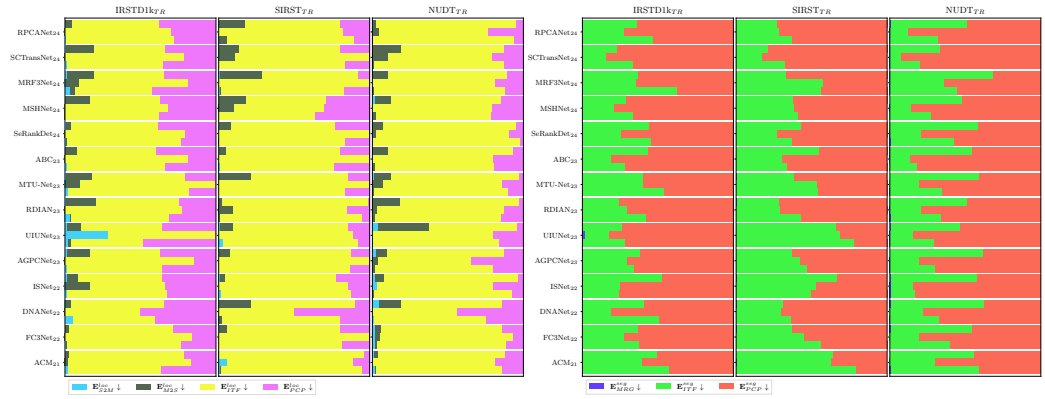
Table 1: Cross-dataset performance analysis. Colors **red**, **green** and **blue** represent the first, second and third ranked results.

		ACM$_{21}$ [9]	FC3Net$_{22}$ [40]	DNANet$_{22}$ [17]	ISNet$_{22}$ [41]	AGPCNet$_{23}$ [42]	UIUNet$_{23}$ [34]	RDIAN$_{23}$ [29]	MTU-Net$_{23}$ [33]	ABC$_{23}$ [22]	SeRankDet$_{24}$ [6]	MSHNet$_{24}$ [20]	MRF3Net$_{24}$ [44]	SCTransNet$_{24}$ [37]	RPCANet$_{24}$ [32]
Params.		0.5M	6.9M	4.7M	1.1M	12.4M	50.5M	0.1M	4.1M	73.5M	108.9M	4.1M	0.5M	11.2M	0.7M
FLOPs		2.0G	10.6G	56.1G	121.9G	327.5G	217.9G	14.8G	24.4G	332.6G	568.7G	24.4G	33.2G	67.4G	179.7G
															Trained on IRSTD1k$_{TR}$ [41].
IRSTD1k$_{TE}$ [41]	IoU$_{pix}$ ↑	0.439	0.358	0.637	0.578	0.605	0.570	0.603	0.610	0.624	**0.642**	**0.650**	0.636	**0.644**	0.608
	nIoU$_{pix}$ ↑	0.476	0.531	**0.625**	0.518	0.580	0.600	0.605	0.607	0.595	0.621	0.620	**0.630**	**0.622**	0.579
	F1$_{pix}$ ↑	0.610	0.527	0.778	0.733	0.754	0.726	0.753	0.757	0.768	**0.782**	**0.788**	0.777	**0.783**	0.756
	Pd↑	0.798	0.865	0.912	0.919	0.916	0.906	0.902	**0.929**	0.916	**0.926**	**0.933**	0.899	0.912	0.886
	+OPDC	0.835	0.875	0.916	0.926	0.919	0.912	0.912	**0.939**	0.919	**0.933**	**0.936**	0.909	0.923	0.896
	Fa×10^6 ↓	95.178	237.365	**13.854**	**13.266**	15.354	51.147	21.503	28.012	16.815	44.638	**11.539**	17.441	16.834	28.145
	+OPDC	61.187	233.266	**10.476**	**11.444**	11.862	44.277	17.062	23.647	13.475	41.108	**7.686**	12.146	**10.476**	21.844
	hIoU↑	0.356	0.383	**0.557**	0.443	0.496	0.530	0.511	0.493	0.508	0.520	**0.549**	**0.553**	0.537	0.470
SIRST$_{TE}$ [9]	IoU$_{pix}$ ↑	0.472	0.234	0.676	0.712	**0.763**	0.696	0.658	0.701	0.708	**0.734**	0.649	**0.752**	0.629	0.543
	nIoU$_{pix}$ ↑	0.567	0.595	0.733	0.719	0.735	0.688	0.735	**0.749**	0.737	**0.742**	0.699	**0.763**	0.686	0.676
	F1$_{pix}$ ↑	0.642	0.380	0.807	0.832	**0.866**	0.821	0.794	0.824	0.829	**0.847**	0.787	**0.858**	0.772	0.704
	Pd↑	0.908	0.872	0.963	**0.982**	**0.991**	0.963	0.963	**0.982**	**0.972**	**0.982**	0.954	**0.982**	0.963	0.927
	+OPDC	0.927	0.881	0.963	**0.982**	**0.991**	**0.982**	0.963	**1.000**	**0.982**	**0.982**	0.954	**0.982**	0.963	0.927
	Fa×10^6 ↓	127.650	932.797	**3.754**	5.632	**2.560**	86.351	17.407	42.152	22.526	17.236	11.946	**4.608**	20.820	124.066
	+OPDC	121.165	932.456	**3.754**	5.632	**2.560**	30.376	17.407	38.397	21.844	17.236	11.946	**4.608**	20.820	124.066
	hIoU↑	0.418	0.390	**0.687**	0.674	0.682	0.644	0.645	**0.693**	0.672	0.684	0.623	**0.694**	0.596	0.598
NUDT$_{TE}$ [17]	IoU$_{pix}$ ↑	0.331	0.288	**0.504**	0.443	0.468	0.450	0.441	0.416	0.469	**0.494**	0.463	**0.512**	0.377	0.291
	nIoU$_{pix}$ ↑	0.452	0.443	**0.627**	0.538	0.556	0.541	0.554	0.528	**0.582**	0.581	0.561	**0.609**	0.502	0.430
	F1$_{pix}$ ↑	0.498	0.448	**0.670**	0.614	0.638	0.620	0.612	0.588	0.638	**0.661**	0.633	**0.678**	0.548	0.451
	Pd↑	0.757	0.752	**0.848**	0.759	0.785	**0.836**	0.804	0.769	0.808	**0.820**	0.771	0.815	0.748	0.701
	+OPDC	0.792	0.787	**0.886**	0.806	0.850	**0.867**	**0.871**	0.857	0.841	0.832	0.843	0.862	0.818	0.722
	Fa×10^6 ↓	253.092	273.488	121.206	71.920	**40.690**	114.695	71.869	96.690	133.464	**63.883**	65.562	**43.844**	110.728	190.989
	+OPDC	236.308	266.469	114.899	57.068	**31.789**	106.049	62.002	79.753	124.563	59.916	**47.913**	**35.502**	92.723	187.581
	hIoU↑	0.274	0.333	**0.526**	0.434	0.456	0.457	0.408	0.366	**0.484**	**0.466**	0.445	**0.484**	0.393	0.344
															Trained on SIRST$_{TR}$ [9].
IRSTD1k$_{TE}$ [41]	IoU$_{pix}$ ↑	0.104	0.456	**0.564**	0.498	0.518	0.444	0.382	0.545	**0.574**	0.549	**0.581**	**0.581**	0.550	0.492
	nIoU$_{pix}$ ↑	0.306	0.469	**0.556**	0.495	0.481	0.470	0.470	**0.544**	**0.544**	0.543	0.542	**0.550**	0.519	0.490
	F1$_{pix}$ ↑	0.188	0.626	**0.721**	0.665	0.683	0.615	0.552	0.705	**0.730**	0.709	**0.735**	**0.735**	0.710	0.659
	Pd↑	0.818	0.865	**0.912**	**0.909**	**0.912**	**0.912**	0.879	**0.909**	**0.902**	**0.909**	0.892	0.899	**0.912**	0.879
	+OPDC	0.872	0.882	**0.919**	**0.916**	**0.919**	**0.929**	0.889	0.912	0.906	**0.916**	0.896	0.906	**0.916**	0.909
	Fa×10^6 ↓	1784.007	139.018	66.672	118.806	84.815	178.835	187.907	88.364	**60.959**	85.537	**44.524**	**56.822**	72.005	88.630
	+OPDC	1704.411	129.073	61.813	115.124	81.608	160.730	182.935	87.947	**60.409**	81.399	**44.144**	**52.666**	71.663	81.342
	hIoU↑	0.124	0.334	**0.435**	0.370	0.356	0.336	0.172	0.408	**0.443**	0.413	**0.459**	0.398	0.400	0.313
SIRST$_{TE}$ [9]	IoU$_{pix}$ ↑	0.607	0.651	0.708	0.746	**0.787**	0.738	0.670	**0.805**	0.775	**0.785**	0.715	0.777	0.764	0.690
	nIoU$_{pix}$ ↑	0.664	0.740	0.779	0.759	0.771	0.723	0.751	**0.796**	**0.793**	**0.800**	0.754	0.775	0.773	0.742
	F1$_{pix}$ ↑	0.756	0.788	0.829	0.854	**0.881**	0.849	0.803	**0.892**	0.873	**0.879**	0.834	0.875	0.866	0.817
	Pd↑	0.954	**1.000**	0.963	**0.982**	**1.000**	**0.991**	0.972	**1.000**	**1.000**	**1.000**	0.972	**1.000**	**1.000**	0.972
	+OPDC	**0.972**	**1.000**	**0.972**	**0.991**	**1.000**	**0.991**	**0.972**	**1.000**	**1.000**	**1.000**	**0.972**	**1.000**	**1.000**	**0.972**
	Fa×10^6 ↓	92.836	121.506	**4.778**	35.155	**8.362**	13.994	75.771	14.335	31.571	31.571	25.598	13.482	**7.679**	51.367
	+OPDC	51.879	121.506	**3.584**	22.526	**8.362**	13.994	75.771	14.335	31.571	31.571	25.598	13.482	**6.485**	51.367
	hIoU↑	0.584	0.655	**0.736**	0.720	0.723	0.679	0.655	0.732	**0.747**	**0.755**	0.699	0.715	0.712	0.675
NUDT$_{TE}$ [17]	IoU$_{pix}$ ↑	0.121	0.361	0.521	0.463	0.458	0.482	0.226	0.502	**0.599**	**0.586**	0.539	**0.543**	0.474	0.355
	nIoU$_{pix}$ ↑	0.373	0.461	0.635	0.552	0.563	0.549	0.527	0.586	**0.663**	**0.644**	0.616	**0.643**	0.572	0.467
	F1$_{pix}$ ↑	0.215	0.530	0.685	0.633	0.628	0.650	0.369	0.668	**0.749**	**0.739**	0.701	**0.704**	0.643	0.524
	Pd↑	0.855	0.757	0.841	0.832	0.846	0.846	0.825	0.846	**0.888**	**0.867**	0.864	**0.890**	0.827	0.750
	+OPDC	0.879	0.836	**0.923**	0.867	0.888	**0.916**	**0.923**	0.909	0.895	0.879	0.883	**0.946**	0.883	0.848
	Fa×10^6 ↓	2600.861	150.604	114.594	115.000	142.008	118.510	1013.184	100.352	**58.492**	**42.979**	**70.089**	97.198	91.553	139.211
	+OPDC	2571.971	139.567	100.352	111.084	137.126	109.049	996.348	82.855	**56.661**	**41.453**	**66.427**	90.078	84.686	124.003
	hIoU↑	0.125	0.317	0.379	0.419	0.340	0.409	0.070	0.404	**0.492**	**0.474**	**0.481**	0.401	0.406	0.328
															Trained on NUDT$_{TR}$ [17].
IRSTD1k$_{TE}$ [41]	IoU$_{pix}$ ↑	0.340	0.398	0.420	0.259	0.443	**0.464**	**0.448**	**0.450**	0.441	0.214	0.405	0.414	0.392	0.270
	nIoU$_{pix}$ ↑	0.420	0.425	0.506	0.342	0.479	**0.522**	**0.523**	0.506	0.492	0.341	**0.527**	0.511	0.484	0.409
	F1$_{pix}$ ↑	0.508	0.569	0.591	0.411	0.614	**0.633**	**0.619**	**0.621**	0.612	0.352	0.576	0.585	0.564	0.425
	Pd↑	0.862	0.859	0.852	0.865	0.875	**0.892**	0.862	**0.892**	**0.896**	**0.943**	**0.896**	0.889	**0.892**	0.855
	+OPDC	0.886	0.879	0.859	0.882	0.892	0.899	0.886	**0.906**	**0.906**	**0.943**	**0.902**	0.896	**0.902**	0.879
	Fa×10^6 ↓	251.409	134.596	**76.408**	268.699	79.710	107.741	**57.904**	112.847	**58.530**	678.048	104.724	81.247	89.465	265.985
	+OPDC	226.149	113.872	73.504	257.919	**70.695**	92.255	**49.325**	106.508	**54.810**	678.048	101.346	79.843	87.112	256.970
	hIoU↑	0.306	0.279	**0.383**	0.186	0.366	**0.382**	0.342	0.322	0.375	0.192	**0.408**	0.325	0.321	0.192
SIRST$_{TE}$ [9]	IoU$_{pix}$ ↑	0.605	0.557	**0.645**	0.572	0.587	**0.675**	**0.638**	0.603	0.611	0.615	0.587	0.625	0.569	0.484
	nIoU$_{pix}$ ↑	0.668	0.627	**0.716**	0.621	0.660	**0.720**	**0.720**	0.687	0.678	0.704	0.670	**0.709**	0.632	0.560
	F1$_{pix}$ ↑	0.754	0.716	**0.784**	0.728	0.740	**0.806**	**0.779**	0.752	0.759	0.762	0.740	0.770	0.725	0.652
	Pd↑	**0.963**	**0.954**	0.936	0.927	0.917	**0.963**	**0.963**	**0.954**	**0.945**	**0.954**	**0.954**	0.936	**0.945**	0.826
	+OPDC	**0.963**	**0.963**	0.936	0.936	0.917	**0.972**	**0.963**	**0.963**	0.945	**0.963**	**0.954**	0.936	0.945	0.826
	Fa×10^6 ↓	33.960	56.316	**8.191**	42.152	**15.188**	22.356	15.700	26.793	17.407	48.125	19.796	25.939	**11.605**	52.220
	+OPDC	33.960	55.463	**8.191**	20.820	**15.188**	19.967	15.700	25.939	17.407	47.613	19.796	25.939	**11.605**	53.586
	hIoU↑	**0.575**	0.441	**0.637**	0.491	0.557	**0.648**	0.566	0.557	0.555	0.526	0.559	0.554	0.511	0.372
NUDT$_{TE}$ [17]	IoU$_{pix}$ ↑	0.635	0.693	0.831	0.790	0.778	**0.840**	0.803	0.809	**0.854**	**0.840**	0.790	0.830	**0.856**	0.747
	nIoU$_{pix}$ ↑	0.668	0.723	0.850	0.813	0.796	0.844	0.826	0.821	**0.862**	**0.854**	0.817	0.847	**0.856**	0.772
	F1$_{pix}$ ↑	0.776	0.819	0.908	0.882	0.875	**0.913**	0.891	0.895	**0.921**	**0.913**	0.882	0.907	**0.923**	0.855
	Pd↑	0.974	0.953	**0.988**	**0.988**	0.986	**0.993**	**0.988**	**0.993**	**0.993**	**0.991**	0.974	**0.991**	**0.991**	0.970
	+OPDC	0.977	0.960	**0.991**	**0.993**	0.986	**0.995**	**0.995**	**0.993**	**0.995**	**0.991**	0.981	**0.993**	**0.995**	0.972
	Fa×10^6 ↓	46.844	46.539	11.698	17.700	13.987	**7.222**	22.125	20.549	12.309	31.586	30.518	**10.071**	**3.916**	57.576
	+OPDC	42.979	36.621	9.003	16.886	13.987	**5.442**	19.786	20.091	10.325	31.586	20.142	**7.782**	**2.391**	57.271
	hIoU↑	0.600	0.620	**0.817**	0.742	0.744	0.793	0.704	0.714	**0.814**	0.728	0.748	0.803	**0.824**	0.547

Table 2: Cross-dataset error analysis for IRSTD1k$_{TR}$ [41]-trained models. See Tab. 3 for details.

		ACM$_{21}$ [9]	FC3Net$_{22}$ [40]	DNANet$_{22}$ [17]	ISNet$_{22}$ [41]	AGPCNet$_{23}$ [42]	UIUNet$_{23}$ [34]	RDIAN$_{23}$ [29]	MTU-Net$_{23}$ [33]	ABC$_{23}$ [22]	SeRankDet$_{24}$ [6]	MSHNet$_{24}$ [20]	MRF3Net$_{24}$ [44]	SCTransNet$_{24}$ [37]	RPCANet$_{24}$ [32]
IRSTD1k$_{TE}$ [41]	IoU$^{neg}_{pix}$ ↑	5.387e-01	6.172e-01	6.694e-01	5.476e-01	6.139e-01	6.317e-01	6.529e-01	6.543e-01	6.361e-01	6.755e-01	6.497e-01	6.742e-01	6.610e-01	6.607e-01
	E$^{neg}_{MRG}$ ↓	2.132e-03	4.340e-04	5.610e-04	4.830e-04	4.340e-04	4.440e-04	5.410e-04	4.580e-04	4.680e-04	4.250e-04	0.000e+00	4.420e-04	4.770e-04	4.570e-04
	E$^{neg}_{FP}$ ↓	2.630e-01	1.431e-01	1.669e-01	1.089e-01	1.330e-01	1.048e-01	1.470e-01	1.855e-01	1.017e-01	1.463e-01	1.186e-01	2.039e-01	1.135e-01	1.595e-01
	E$^{neg}_{PCP}$ ↓	1.962e-01	2.393e-01	1.631e-01	3.430e-01	2.527e-01	2.630e-01	1.996e-01	1.597e-01	2.618e-01	1.778e-01	2.317e-01	1.215e-01	2.250e-01	1.794e-01
SIRST$_{TE}$ [9]	IoU$^{loc}_{tgt}$ ↑	6.613e-01	6.205e-01	8.318e-01	8.088e-01	8.077e-01	8.390e-01	7.832e-01	7.541e-01	7.982e-01	7.694e-01	8.450e-01	8.207e-01	8.131e-01	7.112e-01
	E$^{loc}_{2M}$ ↓	8.000e-03	2.387e-03	9.174e-03	2.941e-03	2.959e-03	3.096e-03	8.671e-03	5.405e-03	2.924e-03	2.778e-03	0.000e+00	6.079e-03	2.967e-03	2.674e-03
	E$^{loc}_{M2S}$ ↓	0.000e+00	2.387e-03	0.000e+00	0.000e+00	0.000e+00	3.096e-03	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	6.079e-03	0.000e+00	0.000e+00
	E$^{loc}_{FP}$ ↓	2.080e-01	2.888e-01	9.174e-02	1.265e-01	1.213e-01	7.740e-02	1.416e-01	1.973e-01	1.316e-01	1.750e-01	9.726e-02	9.119e-02	1.187e-01	2.059e-01
	E$^{loc}_{PCP}$ ↓	1.227e-01	8.592e-02	6.728e-02	6.177e-02	6.805e-02	7.740e-02	6.647e-02	4.324e-02	6.725e-02	5.278e-02	5.775e-02	7.599e-02	6.528e-02	8.021e-02
NUDT$_{TE}$ [17]	IoU$^{neg}_{pix}$ ↑	6.254e-01	7.317e-01	7.391e-01	7.113e-01	7.260e-01	6.857e-01	7.550e-01	7.564e-01	7.405e-01	7.480e-01	7.184e-01	7.658e-01	7.036e-01	7.575e-01
	E$^{neg}_{MRG}$ ↓	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	6.040e-03	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00
	E$^{neg}_{FP}$ ↓	1.488e-01	7.411e-02	5.026e-02	7.203e-02	8.022e-02	4.956e-02	7.283e-02	9.772e-02	4.874e-02	6.509e-02	5.877e-02	8.256e-02	4.715e-02	6.688e-02
	E$^{neg}_{PCP}$ ↓	2.257e-01	1.941e-01	2.106e-01	2.167e-01	1.937e-01	2.587e-01	1.721e-01	1.459e-01	2.107e-01	1.869e-01	2.229e-01	1.516e-01	2.493e-01	1.756e-01
SIRST$_{TE}$ [9]	IoU$^{loc}_{tgt}$ ↑	6.689e-01	5.333e-01	9.292e-01	9.469e-01	9.391e-01	9.386e-01	8.537e-01	9.160e-01	9.068e-01	9.145e-01	8.667e-01	9.068e-01	8.468e-01	7.891e-01
	E$^{loc}_{2M}$ ↓	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	1.754e-02	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00
	E$^{loc}_{M2S}$ ↓	6.623e-03	0.000e+00	0.000e+00	8.850e-03	0.000e+00	0.000e+00	0.000e+00	8.403e-03	0.000e+00	0.000e+00	0.000e+00	8.475e-03	0.000e+00	0.000e+00
	E$^{loc}_{FP}$ ↓	2.715e-01	3.944e-01	3.540e-02	2.655e-02	5.217e-02	4.386e-02	1.138e-01	7.563e-02	7.627e-02	6.838e-02	9.167e-02	6.780e-02	1.210e-01	1.484e-01
	E$^{loc}_{PCP}$ ↓	5.298e-02	7.222e-02	3.540e-02	1.770e-02	8.696e-03	0.000e+00	3.252e-02	0.000e+00	1.695e-02	1.709e-02	4.167e-02	1.695e-02	3.226e-02	6.250e-02
NUDT$_{TE}$ [17]	IoU$^{neg}_{pix}$ ↑	6.170e-01	6.490e-01	7.250e-01	7.177e-01	6.672e-01	6.491e-01	6.475e-01	6.569e-01	7.145e-01	7.075e-01	6.670e-01	7.032e-01	6.504e-01	6.966e-01
	E$^{neg}_{MRG}$ ↓	1.087e-03	1.249e-03	0.000e+00	1.242e-03	1.124e-03	1.110e-03	1.270e-03	1.291e-03	0.000e+00	1.248e-03	0.000e+00	1.284e-03	1.176e-03	1.456e-03
	E$^{neg}_{FP}$ ↓	1.875e-01	1.287e-01	1.127e-01	1.463e-01	1.261e-01	9.163e-02	8.371e-02	1.352e-01	1.242e-01	1.275e-01	9.738e-02	1.079e-01	7.780e-02	1.093e-01
	E$^{neg}_{PCP}$ ↓	1.944e-01	2.210e-01	1.622e-01	1.347e-01	2.056e-01	2.581e-01	2.676e-01	2.066e-01	1.613e-01	1.637e-01	2.357e-01	1.876e-01	2.706e-01	1.927e-01
	IoU$^{loc}_{tgt}$ ↑	4.443e-01	5.129e-01	7.261e-01	6.042e-01	6.829e-01	7.040e-01	6.301e-01	5.569e-01	6.767e-01	6.580e-01	6.673e-01	6.884e-01	6.045e-01	4.944e-01
	E$^{loc}_{2M}$ ↓	1.311e-03	1.522e-03	0.000e+00	1.751e-03	1.876e-03	1.898e-03	1.689e-03	1.517e-03	0.000e+00	1.848e-03	0.000e+00	1.866e-03	1.727e-03	1.600e-03
	E$^{loc}_{M2S}$ ↓	1.180e-02	1.218e-02	1.149e-02	2.627e-02	4.878e-02	2.656e-02	7.264e-02	7.436e-02	2.256e-02	1.294e-02	5.176e-02	5.597e-02	7.081e-02	2.080e-02
	E$^{loc}_{FP}$ ↓	4.273e-01	3.364e-01	1.686e-01	2.242e-01	1.482e-01	1.613e-01	2.044e-01	2.762e-01	1.729e-01	1.959e-01	1.571e-01	1.455e-01	1.900e-01	2.944e-01
	E$^{loc}_{PCP}$ ↓	1.153e-01	1.370e-01	9.387e-02	1.436e-01	1.182e-01	1.063e-01	9.122e-02	9.105e-02	1.278e-01	1.312e-01	1.238e-01	1.082e-01	1.330e-01	1.888e-01

segmentation masks could degrade the reliability of target region decisions. This highlights the necessity of future research to address such edge cases while balancing dataset biases.



(a) Localization errors. (b) Segmentation errors.

Figure 4: Cross-dataset error ratios corresponding to IRSTD1k$_{TE}$, SIRST$_{TE}$, and NUDT$_{TE}$ from bottom to top for the models trained on different datasets.

6 Limitations and Future Extensions

Broader Data and Application Scenarios. Given its status as a representative, mature, and challenging task with extensive research and diverse real-world applications, we selected IRSTD (Infrared Small Target Detection) as our primary focus. Our work addresses critical limitations in its existing evaluation framework through targeted enhancements. While this work provides focused analysis of 14 representative deep learning-based IRSTD methods [9, 40, 17, 41, 42, 34, 29, 33, 22, 6, 20, 44, 37, 32], its scope is fundamentally limited by relying exclusively on commonly-used infrared-only datasets (SIRST [9], IRSTD1k [41], and NUDT [17]). We also explore more challenging scenarios using synthetic data constructed based on data augmentation techniques in Sec. B. However, given the generalizability of our focal problem and the model- and data-agnostic nature of the proposed analysis framework, it can be readily extended to broader data and application scenarios, including the multi-frame IRSTD setting (Sec. C.5), medical image analysis (Sec. C.7), and other visual perception tasks involving small targets in industrial [46, 45, 50] and context-dependent [47, 24, 23, 49, 48, 25] scenarios. Based on the proposed framework, we also plan to conduct more targeted analysis and exploration on more diverse visual perception tasks to further advance the overall community.

Computational Efficiency. In our OPDC strategy described in Alg. 1, we integrate the commonly-used assignment algorithm (`scipy.optimize.linear_sum_assignment` [5]) to find the minimum-cost matching between predicted targets and GT targets. The algorithm ensures optimality but requires $O(n^3)$ time in worst-case scenarios. While our experiments show real-time feasibility in different datasets, scaling to larger-scale targets may necessitate trade-offs between optimality and speed (*e.g.*, via approximations) or hardware-specific optimizations. Additionally, as a performance analysis framework, our method is primarily designed for offline evaluation of IRSTD applications, where runtime is not a critical concern. However, to extend to application scenarios that may have real-time requirements, efficiency will also be one of the issues we will continue to focus on in the future.

7 Conclusion

Our work carefully reviews the limitations of existing evaluation protocols in IRSTD. And to this end, we introduce a novel hybrid-level performance metric and a systematic error analysis method, emphasizing the necessity of cross-dataset validation in evaluation. The proposed metric holistically models hybrid localization-segmentation information to comprehensively characterize the algorithm capabilities. Coupled with this, our error analysis provides fine-grained diagnostic insights into failure modes, thereby deepening the understanding of model performance. Through extensive within- and cross-dataset experiments on existing benchmarks, we demonstrate the effectiveness and interpretative power of our framework in evaluating the performance of IRSTD methods. By establishing an enhanced analytical paradigm, our work aims to advance the methodological foundation of IRSTD algorithm development and evaluation, with important implications for guiding more robust model design and validation in practical applications.

Acknowledgments and Disclosure of Funding

This work has greatly benefited from the open-source contributions of the research community. In particular, we would like to acknowledge the repository `BasicIRSTD` (https://github.com/XinyiYing/BasicIRSTD) developed by Xinyi Ying and collaborators, which offered important references and implementations that supported the development and validation of our study in the field of infrared small target detection (IRSTD).

This work was supported by the National Natural Science Foundation of China under Grant 62431004.

References

[1] Xiangzhi Bai and Fugen Zhou. Analysis of new top-hat transformation and the application for infrared dim small target detection. *Pattern Recognition*, 43(6):2145–2156, 2010. 2

[2] Michael Bruno, Alexander Sutin, Kil Woo Chung, Alexander Sedunov, Nikolay Sedunov, Hady Salloum, Hans Graber, and Paul Mallas. Satellite imaging and passive acoustics in layered approach for small boat detection and classification. *Marine Technology Society Journal*, 45(3):77–87, 2011. 1

[3] Zhaoyang Cao, Xuan Kong, Qiang Zhu, Siying Cao, and Zhenming Peng. Infrared dim target detection via mode-k1k2 extension tensor tubal rank under complex ocean environment. *ISPRS Journal of Photogrammetry and Remote Sensing*, 181:167–190, 2021. 1

[4] Yongbo Cheng, Xuefeng Lai, Yucheng Xia, and Jinmei Zhou. Infrared dim small target detection networks: A review. *Sensors*, 24(12):3885, June 2024. 1, 2

[5] David F. Crouse. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, August 2016. 5, 7, 10

[6] Yimian Dai, Peiwen Pan, Yulei Qian, Yuxuan Li, Xiang Li, Jian Yang, and Huan Wang. Pick of the bunch: Detecting infrared small targets beyond hit-miss trade-offs via selective rank-aware attention. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024. 2, 7, 8, 9, 10, 22, 25, 28

[7] Yimian Dai and Yiquan Wu. Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(8):3752–3767, 2017. 2

[8] 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. 2

[9] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. Asymmetric contextual modulation for infrared small target detection. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 949–958, Jan 2021. 2, 3, 7, 8, 9, 10, 15, 16, 18, 22, 23, 25, 28, 29, 30

[10] 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. 2

[11] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020. 29

[12] 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. 2

[13] Jinhui Han, Kun Liang, Bo Zhou, Xinying Zhu, Jie Zhao, and Linlin Zhao. Infrared small target detection utilizing the multiscale relative local contrast measure. *IEEE Geoscience and Remote Sensing Letters*, 15(4):612–616, 2018. 2

[14] Jinhui Han, Sibang Liu, Gang Qin, Qian Zhao, Honghui Zhang, and Nana Li. A local contrast method combined with adaptive background estimation for infrared small target detection. *IEEE Geoscience and Remote Sensing Letters*, 16(9):1442–1446, 2019. 2

[15] Jinhui Han, Yong Ma, Bo Zhou, Fan Fan, Kun Liang, and Yu Fang. A robust infrared small target detection algorithm based on human visual system. *IEEE Geoscience and Remote Sensing Letters*, 11(12):2168–2172, 2014. 2

[16] Renke Kou, Chunping Wang, Zhenming Peng, Zhihe Zhao, Yaohong Chen, Jinhui Han, Fuyu Huang, Ying Yu, and Qiang Fu. Infrared small target segmentation networks: A survey. *Pattern Recognition*, 143:109788, November 2023. 1, 2

[17] Boyang Li, Chao Xiao, Longguang Wang, Yingqian Wang, Zaiping Lin, Miao Li, Wei An, and Yulan Guo. Dense nested attention network for infrared small target detection. *IEEE Transactions on Image Processing*, 32:1745–1758, 2023. 1, 2, 3, 4, 5, 7, 8, 9, 10, 15, 16, 18, 21, 22, 25, 27, 28, 29, 30

[18] Ruojing Li, Wei An, Chao Xiao, Boyang Li, Yingqian Wang, Miao Li, and Yulan Guo. Direction-coded temporal u-shape module for multiframe infrared small target detection. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):555–568, 2025. 2, 27, 28

[19] Ruojing Li, Wei An, Xinyi Ying, Yingqian Wang, Yimian Dai, Longguang Wang, Miao Li, Yulan Guo, and Li Liu. Probing deep into temporal profile makes the infrared small target detector much better. *CoRR*, abs/2506.12766, 2025. 2, 27, 28

[20] 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 Conference on Computer Vision and Pattern Recognition*, pages 17490–17499, June 2024. 1, 2, 8, 9, 10, 16, 22, 25, 28

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 3

[22] Peiwen Pan, Huan Wang, Chenyi Wang, and Chang Nie. Abc: Attention with bilinear correlation for infrared small target detection. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 2381–2386, July 2023. 2, 4, 8, 9, 10, 22, 25, 28

[23] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 10

[24] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 10

[25] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 10

[26] 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. 2

[27] Zhaobing Qiu, Yong Ma, Fan Fan, Jun Huang, and Lang Wu. Global sparsity-weighted local contrast measure for infrared small target detection. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 2

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015. 3

[29] Heng Sun, Junxiang Bai, Fan Yang, and Xiangzhi Bai. Receptive-field and direction induced attention network for infrared dim small target detection with a large-scale dataset irdst. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. 2, 8, 9, 10, 22, 25, 28

[30] Fan Wang, Weixian Qian, Ye Qian, Chao Ma, He Zhang, Jiajie Wang, Minjie Wan, and Kan Ren. Maritime infrared small target detection based on the appearance stable isotropy measure in heavy sea clutter environments. *Sensors*, 23(24):9838, 2023. 1

[31] Yantao Wei, Xinge You, and Hong Li. Multiscale patch-based contrast measure for small infrared target detection. *Pattern Recognition*, 58:216–226, 2016. 2

[32] Fengyi Wu, Tianfang Zhang, Lei Li, Yian Huang, and Zhenming Peng. Rpcanet: Deep unfolding rpca based infrared small target detection. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 4797–4806, Jan 2024. 2, 8, 9, 10, 22, 25, 28, 29, 30

[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. 2, 4, 8, 9, 10, 22, 25, 28, 29, 30

[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, 2023. 2, 8, 9, 10, 22, 24, 25, 28

[35] Xinyi Ying, Li Liu, Zaiping 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*, 63:1–18, 2025. 2, 27, 28

[36] Xinyi Ying, Yingqian Wang, Longguang Wang, Weidong Sheng, Li Liu, Zaiping Lin, and Shilin Zhou. Mocopnet: Exploring local motion and contrast priors for infrared small target super-resolution. *CoRR*, abs/2201.01014, 2022. 1

[37] 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. 1, 2, 8, 9, 10, 22, 25, 27, 28

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

[39] Landan Zhang and Zhenming Peng. Infrared small target detection based on partial sum of the tensor nuclear norm. *Remote Sensing*, 11(4):382, 2019. 2

[40] 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*, 2022. 2, 8, 9, 10, 22, 23, 25, 28

[41] 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 Conference on Computer Vision and Pattern Recognition*, pages 867–876, June 2022. 2, 7, 8, 9, 10, 15, 16, 18, 19, 22, 24, 25, 28, 29, 30

[42] 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, Aug 2023. 2, 8, 9, 10, 22, 25, 28

[43] Tianfang Zhang, Hao Wu, Yuhan Liu, Lingbing Peng, Chunping Yang, and Zhenming Peng. Infrared small target detection based on non-convex optimization with l_p-norm constraint. *Remote Sensing*, 11(5):559, 2019. 2

[44] Xiaohan Zhang, Xue Zhang, Si-Yuan Cao, Beinan Yu, Chenghao Zhang, and Hui-Liang Shen. Mrf3net: An infrared small target detection network using multireceptive field perception and effective feature fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. 2, 8, 9, 10, 22, 25, 28

[45] Xiaoqi Zhao, Peiqian Cao, Lihe Zhang, Zonglei Feng, Hanqi Liu, Jiaming Zuo, Youwei Pang, Weisi Lin, Georges El Fakhri, Huchuan Lu, and Xiaofeng Liu. Power battery detection. *CoRR*, abs/2508.07797, 2025. 10

[46] Xiaoqi Zhao, Youwei Pang, Zhenyu Chen, Qian Yu, Lihe Zhang, Hanqi Liu, Jiaming Zuo, and Huchuan Lu. Towards automatic power battery detection: New challenge benchmark dataset and baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 10

[47] Xiaoqi Zhao, Youwei Pang, Wei Ji, Baicheng Sheng, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. Spider: A unified framework for context-dependent concept understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 10

[48] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *Proceedings of the European Conference on Computer Vision*, 2020. 10

[49] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Towards diverse binary segmentation via a simple yet general gated network. *International Journal of Computer Vision*, 2024. 10

[50] Yuan Zhao, Youwei Pang, Lihe Zhang, Hanqi Liu, Jiaming Zuo, Huchuan Lu, and Xiaoqi Zhao. Unimmad: Unified multi-modal and multi-class anomaly detection via moe-driven feature decompression, 2025. 10

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: We introduce the contributions and scope of this paper in Sec. Abstract and Sec. Introduction (Sec. 1).

   Guidelines:

   - The answer NA 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 NA 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 of our work and propose future potential extensions for our efforts in Sec. 6.

   Guidelines:

   - The answer NA 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: [NA]

Justification: Our work does not include theoretical results.

Guidelines:

- The answer NA 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: In the method section (Sec. 4), we provide a detailed description of our method and introduce the experimental details in the experimental section (Sec. 5.1). The datasets (SIRST [9], IRSTD1k [41], and NUDT [17]) used in this paper and the code for the relevant algorithms are publicly available. And these datasets involved have been widely used by the relevant communities.

Guidelines:

- The answer NA 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: As previously mentioned, the data (SIRST [9], IRSTD1k [41], and NUDT [17]) used in this paper and the code for the relevant algorithms are publicly available.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.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://nips.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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We detail the experimental details in Sec. 5.1. To ensure that the experiments are reasonable, **we follow the regular practice of existing algorithms** [20].

   Guidelines:

   - The answer NA 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: Each experiment requires a substantial amount of resources and time, so we do not conduct a statistical analysis.

   Guidelines:

   - The answer NA 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: We have specific instructions in the implementation details of the experiment section (Sec. 5.1).

   Guidelines:

   - The answer NA 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 carefully read the NeurIPS Code of Ethics and ensured compliance with it.

   Guidelines:

   - The answer NA 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 discuss the societal impacts of our work in Sec. D.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work poses no such risks.

    Guidelines:

    - The answer NA 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 ensure that the assets we use are licensed. The data (SIRST [9], IRSTD1k [41], and NUDT [17]) used in this paper and the code for the relevant algorithms are publicly available.

    Guidelines:

    - The answer NA 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: In our work, we augment the testset of the existing IRSTD1k [41] (MIT License) using a randomized copy-paste augmentation strategy and construct the synthetic dataset, *i.e.*, IRSTD1k$_{TE}^{AUG}$, and we include the relevant Croissant file with our paper submission.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer NA 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 rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

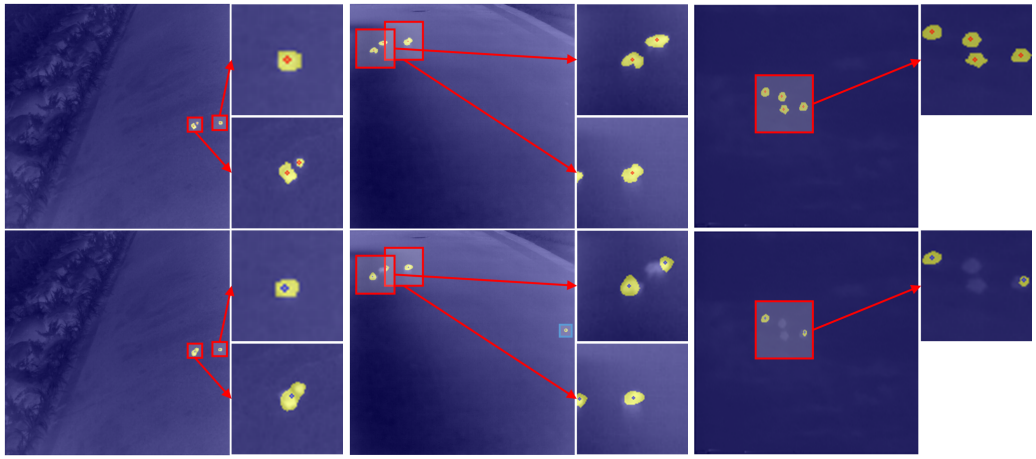
    Guidelines:

    - The answer NA 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 (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

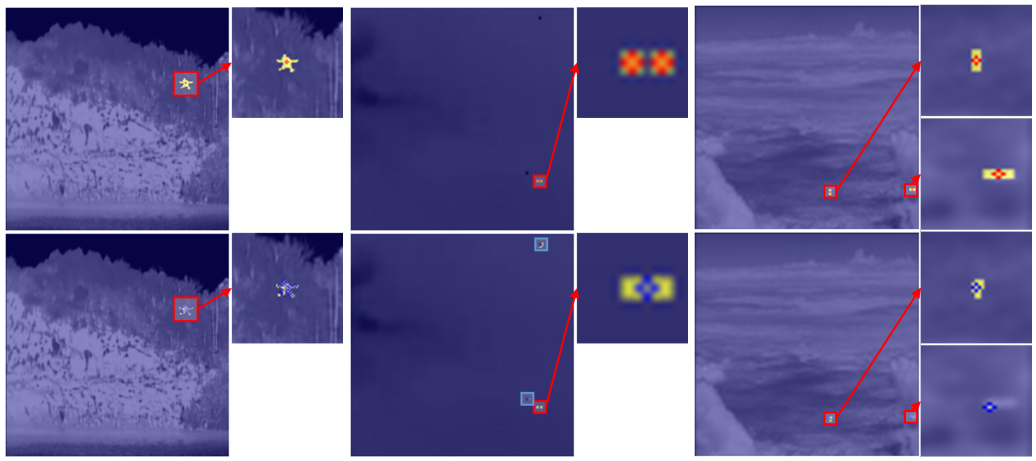
Technical Appendices and Supplementary Material

A Clearer Comparison

This section further adds the following:
- Fig. 5: Clearer figure visualization for improved readability.
- Tab. 3: Complete cross-dataset error analysis table (only the first section is presented in the main text due to space constraints).
- Fig. 6 and Fig. 7: Clearer statistical visualization of cross-dataset error ratios.



(a) (b) (c)



(d) (e) (f)

Subfigure	$IoU_{pix}\uparrow$	$nIoU_{pix}\uparrow$	$F1_{pix}\uparrow$	Pd↑		Fa×$10^6\downarrow$		hIoU↑	$IoU^{seg}_{pix}\uparrow$	$E^{seg}_{MRG}\downarrow$	$E^{seg}_{ITF}\downarrow$	$E^{seg}_{PCP}\downarrow$	$IoU^{loc}_{tgt}\uparrow$	$E^{loc}_{S2M}\downarrow$	$E^{loc}_{M2S}\downarrow$	$E^{loc}_{ITF}\downarrow$	$E^{loc}_{PCP}\downarrow$
				Ori. [17]	OPDC	Ori. [17]	OPDC	OPDC	OPDC	OPDC	OPDC	OPDC	OPDC	OPDC	OPDC	OPDC	OPDC
(a)	0.573	0.573	0.728	0.333	0.667	785.828	0.000	0.373	0.560	0.052	0.246	0.142	0.667	0.333	0.000	0.000	0.000
(b)	0.438	0.438	0.610	0.333	1.000	1457.214	198.364	0.344	0.459	0.000	0.225	0.316	0.750	0.000	0.000	0.250	0.000
(c)	0.310	0.310	0.473	0.500	0.500	0.000	0.000	0.300	0.600	0.000	0.000	0.400	0.500	0.000	0.000	0.000	0.500
(d)	0.477	0.477	0.646	1.000	1.000	579.834	625.610	0.005	0.036	0.000	0.000	0.964	0.143	0.000	0.857	0.000	0.000
(e)	0.514	0.514	0.679	0.500	0.500	244.141	244.141	0.118	0.474	0.474	0.053	0.000	0.250	0.250	0.000	0.500	0.000
(f)	0.326	0.326	0.492	1.000	1.000	0.000	0.000	0.313	0.313	0.000	0.085	0.602	1.000	0.000	0.000	0.000	0.000

Figure 5: Comparison of different metrics. Red and blue boxes to highlight the target regions. Red and blue points indicate the target centroids in ground truth (GT) masks and predictions, respectively. Zoom in on the digital color version for details. "Ori." [17] and "OPDC" refer to the original distance-based strategy and the proposed OPDC strategy for target matching.

Table 3: Cross-dataset error analysis for the models trained on different datasets.

		ACM$_{21}$ [9]	FC3Net$_{22}$ [40]	DNANet$_{22}$ [17]	ISNet$_{22}$ [41]	AGPCNet$_{23}$ [42]	UIUNet$_{23}$ [34]	RDIAN$_{23}$ [29]	MTU-Net$_{23}$ [33]	ABC$_{23}$ [22]	SeRankDet$_{24}$ [6]	MSHNet$_{24}$ [20]	MRF3Net$_{24}$ [44]	SCTransNet$_{24}$ [37]	RPCANet$_{24}$ [32]
															Trained on IRSTD1k$_{TR}$ [41].
IRSTD1k$_{TE}$ [41]	IoU$^{seg}_{pix}$ ↑	5.387e-01	6.172e-01	6.694e-01	5.476e-01	6.139e-01	6.317e-01	6.529e-01	6.543e-01	6.361e-01	6.755e-01	6.497e-01	6.742e-01	6.610e-01	6.607e-01
	E$^{seg}_{AUC}$ ↓	2.132e-03	4.340e-04	5.610e-04	4.830e-04	4.340e-04	4.440e-04	5.410e-04	4.580e-04	4.680e-04	4.250e-04	0.000e+00	4.420e-04	4.770e-04	4.570e-04
	E$^{seg}_{ITF}$ ↓	2.630e-01	1.431e-01	1.669e-01	1.089e-01	1.330e-01	1.048e-01	1.470e-01	1.855e-01	1.017e-01	1.463e-01	1.186e-01	2.039e-01	1.135e-01	1.595e-01
	E$^{seg}_{PCP}$ ↓	1.962e-01	2.393e-01	1.631e-01	3.430e-01	2.527e-01	2.630e-01	1.996e-01	1.597e-01	2.618e-01	1.778e-01	2.317e-01	1.215e-01	2.250e-01	1.794e-01
	IoU$^{loc}_{tgt}$ ↑	6.613e-01	6.205e-01	8.318e-01	8.088e-01	8.077e-01	8.390e-01	7.832e-01	7.541e-01	7.982e-01	7.694e-01	8.450e-01	8.207e-01	8.131e-01	7.112e-01
	E$^{loc}_{S2M}$ ↓	8.000e-03	2.387e-03	9.174e-03	2.941e-03	2.959e-03	3.096e-03	8.671e-03	5.405e-03	2.924e-03	2.778e-03	0.000e+00	6.079e-03	2.967e-03	2.674e-03
	E$^{loc}_{M2S}$ ↓	0.000e+00	2.387e-03	0.000e+00	0.000e+00	0.000e+00	3.096e-03	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	6.079e-03	0.000e+00	0.000e+00
	E$^{loc}_{ITF}$ ↓	2.080e-01	2.888e-01	9.174e-02	1.265e-01	1.213e-01	7.740e-02	1.416e-01	1.973e-01	1.316e-01	1.750e-01	9.726e-02	9.119e-02	1.187e-01	2.059e-01
	E$^{loc}_{PCP}$ ↓	1.227e-01	8.592e-02	6.728e-02	6.177e-02	6.805e-02	7.740e-02	6.647e-02	4.324e-02	6.725e-02	5.278e-02	5.775e-02	7.599e-02	6.528e-02	8.021e-02
SIRST$_{TE}$ [9]	IoU$^{seg}_{pix}$ ↑	6.254e-01	7.317e-01	7.391e-01	7.113e-01	7.260e-01	6.857e-01	7.550e-01	7.564e-01	7.405e-01	7.480e-01	7.184e-01	7.658e-01	7.036e-01	7.575e-01
	E$^{seg}_{AUC}$ ↓	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	6.040e-03	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00
	E$^{seg}_{ITF}$ ↓	1.488e-01	7.411e-02	5.026e-02	7.203e-02	8.022e-02	4.956e-02	7.283e-02	9.772e-02	4.874e-02	6.509e-02	5.877e-02	8.256e-02	4.715e-02	6.688e-02
	E$^{seg}_{PCP}$ ↓	2.257e-01	1.941e-01	2.106e-01	2.167e-01	1.937e-01	2.587e-01	1.721e-01	1.459e-01	2.107e-01	1.869e-01	2.229e-01	1.516e-01	2.493e-01	1.756e-01
	IoU$^{loc}_{tgt}$ ↑	6.689e-01	5.333e-01	9.292e-01	9.469e-01	9.391e-01	9.386e-01	8.537e-01	9.160e-01	9.068e-01	9.145e-01	8.667e-01	9.068e-01	8.468e-01	7.891e-01
	E$^{loc}_{S2M}$ ↓	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	1.754e-02	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00
	E$^{loc}_{M2S}$ ↓	6.623e-03	0.000e+00	0.000e+00	8.850e-03	0.000e+00	0.000e+00	0.000e+00	8.403e-03	0.000e+00	0.000e+00	0.000e+00	8.475e-03	0.000e+00	0.000e+00
	E$^{loc}_{ITF}$ ↓	2.715e-01	3.944e-01	3.540e-02	2.655e-02	5.217e-02	4.386e-02	1.138e-01	7.563e-02	7.627e-02	6.838e-02	9.167e-02	6.780e-02	1.210e-01	1.484e-01
	E$^{loc}_{PCP}$ ↓	5.298e-02	7.222e-02	3.540e-02	1.770e-02	8.696e-03	0.000e+00	3.252e-02	0.000e+00	1.695e-02	1.709e-02	4.167e-02	1.695e-02	3.226e-02	6.250e-02
NUDT$_{TE}$ [17]	IoU$^{seg}_{pix}$ ↑	6.170e-01	6.490e-01	7.250e-01	7.177e-01	6.672e-01	6.491e-01	6.475e-01	6.569e-01	7.145e-01	7.075e-01	6.670e-01	7.032e-01	6.504e-01	6.966e-01
	E$^{seg}_{AUC}$ ↓	1.087e-03	1.249e-03	0.000e+00	1.242e-03	1.124e-03	1.110e-03	1.270e-03	1.291e-03	0.000e+00	1.248e-03	0.000e+00	1.284e-03	1.176e-03	1.456e-03
	E$^{seg}_{ITF}$ ↓	1.875e-01	1.287e-01	1.127e-01	1.463e-01	1.261e-01	9.163e-02	8.371e-02	1.352e-01	1.242e-01	1.275e-01	9.738e-02	1.079e-01	7.780e-02	1.093e-01
	E$^{seg}_{PCP}$ ↓	1.944e-01	2.210e-01	1.622e-01	1.347e-01	2.056e-01	2.581e-01	2.676e-01	2.066e-01	1.613e-01	1.637e-01	2.357e-01	1.876e-01	2.706e-01	1.927e-01
	IoU$^{loc}_{tgt}$ ↑	4.443e-01	5.129e-01	7.261e-01	6.042e-01	6.829e-01	7.040e-01	6.301e-01	5.569e-01	6.767e-01	6.580e-01	6.673e-01	6.884e-01	6.045e-01	4.944e-01
	E$^{loc}_{S2M}$ ↓	1.311e-03	1.522e-03	0.000e+00	1.751e-03	1.876e-03	1.898e-03	1.689e-03	1.517e-03	0.000e+00	1.848e-03	0.000e+00	1.866e-03	1.727e-03	1.600e-03
	E$^{loc}_{M2S}$ ↓	1.180e-02	1.218e-02	1.149e-02	2.627e-02	4.878e-02	2.656e-02	7.264e-02	7.436e-02	2.256e-02	1.294e-02	5.176e-02	5.597e-02	7.081e-02	2.080e-02
	E$^{loc}_{ITF}$ ↓	4.273e-01	3.364e-01	1.688e-01	2.242e-01	1.482e-01	1.613e-01	2.044e-01	2.762e-01	1.729e-01	1.959e-01	1.571e-01	1.455e-01	1.900e-01	2.944e-01
	E$^{loc}_{PCP}$ ↓	1.153e-01	1.370e-01	9.387e-02	1.436e-01	1.182e-01	1.063e-01	9.122e-02	9.105e-02	1.278e-01	1.312e-01	1.238e-01	1.082e-01	1.330e-01	1.888e-01
															Trained on SIRST$_{TR}$ [9].
IRSTD1k$_{TE}$ [41]	IoU$^{seg}_{pix}$ ↑	4.501e-01	5.769e-01	6.311e-01	6.133e-01	5.876e-01	5.900e-01	6.009e-01	6.506e-01	6.301e-01	6.492e-01	6.070e-01	6.365e-01	6.268e-01	6.135e-01
	E$^{seg}_{AUC}$ ↓	2.131e-03	4.720e-04	0.000e+00	1.243e-03	4.990e-04	1.563e-03	4.870e-04	0.000e+00	4.820e-04	0.000e+00	0.000e+00	4.300e-04	0.000e+00	5.070e-04
	E$^{seg}_{ITF}$ ↓	4.390e-01	2.374e-01	1.354e-01	1.902e-01	1.946e-01	3.172e-01	1.721e-01	1.896e-01	1.239e-01	1.310e-01	1.628e-01	2.031e-01	1.226e-01	1.445e-01
	E$^{seg}_{PCP}$ ↓	1.088e-01	1.852e-01	2.335e-01	1.953e-01	2.172e-01	9.125e-02	2.265e-01	1.598e-01	2.456e-01	2.198e-01	2.302e-01	1.600e-01	2.506e-01	2.415e-01
	IoU$^{loc}_{tgt}$ ↑	2.761e-01	5.796e-01	6.894e-01	6.031e-01	6.053e-01	5.702e-01	2.870e-01	6.273e-01	7.024e-01	6.355e-01	7.557e-01	6.256e-01	6.385e-01	5.104e-01
	E$^{loc}_{S2M}$ ↓	5.330e-03	2.212e-03	0.000e+00	6.652e-03	2.217e-03	1.033e-02	1.087e-03	0.000e+00	2.611e-03	0.000e+00	0.000e+00	2.326e-03	0.000e+00	1.890e-03
	E$^{loc}_{M2S}$ ↓	1.066e-03	0.000e+00	7.576e-03	0.000e+00	0.000e+00	0.000e+00	6.522e-03	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	2.457e-02
	E$^{loc}_{ITF}$ ↓	6.823e-01	3.429e-01	2.424e-01	3.415e-01	3.415e-01	3.864e-01	6.707e-01	3.125e-01	2.245e-01	3.061e-01	1.562e-01	3.093e-01	3.028e-01	4.140e-01
	E$^{loc}_{PCP}$ ↓	3.518e-02	7.522e-02	6.061e-02	4.878e-02	5.100e-02	3.306e-02	3.478e-02	6.019e-02	7.050e-02	5.841e-02	8.807e-02	6.279e-02	5.869e-02	4.915e-02
SIRST$_{TE}$ [9]	IoU$^{seg}_{pix}$ ↑	6.835e-01	7.630e-01	7.776e-01	7.665e-01	7.763e-01	7.360e-01	7.792e-01	8.062e-01	7.948e-01	8.036e-01	7.645e-01	7.809e-01	7.710e-01	7.831e-01
	E$^{seg}_{AUC}$ ↓	1.741e-03	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00
	E$^{seg}_{ITF}$ ↓	1.976e-01	1.071e-01	6.655e-02	1.240e-01	9.534e-02	1.829e-01	6.401e-02	1.037e-01	6.307e-02	5.319e-02	9.104e-02	1.270e-01	3.959e-02	6.215e-02
	E$^{seg}_{PCP}$ ↓	1.171e-01	1.299e-01	1.559e-01	1.094e-01	1.284e-01	8.107e-02	1.568e-01	9.009e-02	1.422e-01	1.432e-01	1.444e-01	9.207e-02	1.895e-01	1.548e-01
	IoU$^{loc}_{tgt}$ ↑	8.548e-01	8.583e-01	9.464e-01	9.391e-01	9.316e-01	9.231e-01	8.413e-01	9.083e-01	9.397e-01	9.397e-01	9.138e-01	9.160e-01	9.237e-01	8.618e-01
	E$^{loc}_{S2M}$ ↓	8.065e-03	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00
	E$^{loc}_{M2S}$ ↓	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	1.587e-02	0.000e+00	0.000e+00	0.000e+00	8.621e-03	0.000e+00	8.475e-03	0.000e+00
	E$^{loc}_{ITF}$ ↓	1.210e-01	1.417e-01	2.679e-02	5.217e-02	6.838e-02	6.838e-02	1.190e-01	9.167e-02	6.035e-02	6.035e-02	5.172e-02	8.403e-02	6.780e-02	1.138e-01
	E$^{loc}_{PCP}$ ↓	1.613e-02	0.000e+00	2.679e-02	8.696e-03	0.000e+00	8.547e-03	2.381e-02	0.000e+00	0.000e+00	0.000e+00	2.586e-02	0.000e+00	0.000e+00	2.439e-02
NUDT$_{TE}$ [17]	IoU$^{seg}_{pix}$ ↑	5.597e-01	6.327e-01	6.988e-01	7.029e-01	6.763e-01	6.587e-01	6.541e-01	6.777e-01	7.832e-01	7.566e-01	6.999e-01	6.855e-01	6.793e-01	6.279e-01
	E$^{seg}_{AUC}$ ↓	6.380e-04	1.143e-03	1.139e-03	1.155e-03	1.003e-03	9.980e-04	1.085e-03	1.157e-03	1.119e-03	1.260e-03	1.058e-03	1.010e-03	1.253e-03	1.181e-03
	E$^{seg}_{ITF}$ ↓	2.834e-01	1.364e-01	9.303e-02	1.979e-01	1.202e-01	2.237e-01	9.840e-02	1.232e-01	1.178e-01	1.033e-01	1.122e-01	1.015e-01	6.661e-02	9.961e-02
	E$^{seg}_{PCP}$ ↓	1.563e-01	2.297e-01	2.070e-01	9.805e-02	2.025e-01	1.166e-01	2.464e-01	1.980e-01	9.795e-02	1.388e-01	1.868e-01	2.120e-01	2.529e-01	2.713e-01
	IoU$^{loc}_{tgt}$ ↑	2.242e-01	5.007e-01	5.418e-01	5.955e-01	5.026e-01	6.212e-01	1.063e-01	5.957e-01	6.279e-01	6.267e-01	6.873e-01	5.853e-01	5.981e-01	5.231e-01
	E$^{loc}_{S2M}$ ↓	5.960e-04	1.399e-03	1.372e-03	1.605e-03	1.323e-03	1.585e-03	2.690e-04	1.531e-03	1.639e-03	1.667e-03	1.818e-03	1.445e-03	1.582e-03	1.441e-03
	E$^{loc}_{M2S}$ ↓	1.789e-03	2.238e-02	9.877e-02	1.284e-02	3.571e-02	3.328e-02	1.856e-02	8.729e-02	1.639e-02	2.833e-02	5.455e-02	1.127e-01	5.380e-02	8.069e-02
	E$^{loc}_{ITF}$ ↓	7.430e-01	3.790e-01	3.141e-01	3.002e-01	3.981e-01	2.884e-01	8.663e-01	2.573e-01	2.820e-01	2.583e-01	1.673e-01	2.688e-01	2.690e-01	3.026e-01
	E$^{loc}_{PCP}$ ↓	3.041e-02	9.650e-02	4.390e-02	8.989e-02	6.217e-02	5.547e-02	8.609e-03	5.819e-02	7.213e-02	8.500e-02	8.909e-02	3.179e-02	7.753e-02	9.222e-02
															Trained on NUDT$_{TR}$ [17].
IRSTD1k$_{TE}$ [41]	IoU$^{seg}_{pix}$ ↑	5.488e-01	5.966e-01	6.225e-01	4.989e-01	5.792e-01	6.397e-01	6.289e-01	6.459e-01	5.752e-01	6.597e-01	6.141e-01	6.347e-01	5.949e-01	6.154e-01
	E$^{seg}_{AUC}$ ↓	1.617e-03	1.387e-03	5.760e-04	1.380e-03	5.170e-04	5.010e-04	5.120e-04	5.060e-04	0.000e+00	4.660e-04	0.000e+00	0.000e+00	0.000e+00	4.960e-04
	E$^{seg}_{ITF}$ ↓	2.653e-01	1.621e-01	1.242e-01	8.047e-02	1.250e-01	1.052e-01	1.241e-01	1.214e-01	7.609e-02	1.440e-01	1.045e-01	1.614e-01	8.208e-02	1.219e-01
	E$^{seg}_{PCP}$ ↓	1.843e-01	2.399e-01	2.527e-01	4.193e-01	2.953e-01	2.546e-01	2.465e-01	2.321e-01	3.487e-01	1.959e-01	2.815e-01	2.039e-01	3.230e-01	2.623e-01
	IoU$^{loc}_{tgt}$ ↑	5.572e-01	4.677e-01	6.145e-01	3.727e-01	6.325e-01	5.973e-01	5.445e-01	4.991e-01	6.513e-01	2.917e-01	6.650e-01	5.115e-01	5.392e-01	3.115e-01
	E$^{loc}_{S2M}$ ↓	8.475e-03	3.584e-03	4.819e-03	2.845e-03	2.387e-03	2.237e-03	2.070e-03	1.855e-03	0.000e+00	1.042e-03	0.000e+00	0.000e+00	0.000e+00	1.193e-03
	E$^{loc}_{M2S}$ ↓	2.119e-03	1.792e-03	0.000e+00	4.267e-03	4.773e-03	0.000e+00	6.211e-03	5.566e-03	7.264e-03	1.042e-03	2.481e-03	1.923e-03	6.036e-03	4.773e-03
	E$^{loc}_{ITF}$ ↓	3.686e-01	4.659e-01	2.843e-01	5.733e-01	2.864e-01	3.356e-01	3.789e-01	4.434e-01	2.736e-01	6.896e-01	2.605e-01	4.269e-01	3.964e-01	6.408e-01
	E$^{loc}_{PCP}$ ↓	6.356e-02	6.093e-02	9.639e-02	4.694e-02	7.399e-02	6.488e-02	6.832e-02	5.009e-02	6.780e-02	1.667e-02	7.196e-02	5.962e-02	5.835e-02	4.177e-02
SIRST$_{TE}$ [9]	IoU$^{seg}_{pix}$ ↑	6.896e-01	6.802e-01	7.372e-01	6.354e-01	7.015e-01	7.395e-01	7.330e-01	7.107e-01	7.165e-01	7.369e-01	6.876e-01	7.547e-01	6.549e-01	7.070e-01
	E$^{seg}_{AUC}$ ↓	0.000e+00	0.000e+00	0.000e+00	2.723e-03	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00
	E$^{seg}_{ITF}$ ↓	9.297e-02	6.004e-02	5.745e-02	5.292e-02	5.888e-02	4.036e-02	5.550e-02	5.562e-02	3.822e-02	5.499e-02	4.493e-02	8.826e-02	2.612e-02	3.665e-02
	E$^{seg}_{PCP}$ ↓	2.174e-01	2.597e-01	2.054e-01	3.090e-01	2.396e-01	2.201e-01	2.115e-01	2.337e-01	2.453e-01	2.081e-01	2.675e-01	1.570e-01	3.189e-01	2.563e-01
	IoU$^{loc}_{tgt}$ ↑	8.333e-01	6.481e-01	8.644e-01	7.727e-01	7.937e-01	8.760e-01	7.721e-01	7.836e-01	7.744e-01	7.143e-01	8.125e-01	7.338e-01	7.803e-01	5.263e-01
	E$^{loc}_{S2M}$ ↓	0.000e+00	0.000e+00	0.000e+00	7.576e-03	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00	0.000e+00
	E$^{loc}_{M2S}$ ↓	0.000e+00	6.173e-03	0.000e+00	0.000e+00	7.937e-03	0.000e+00	7.353e-03	1.493e-02	0.000e+00	6.803e-03	0.000e+00	0.000e+00	2.273e-02	2.339e-02
	E$^{loc}_{ITF}$ ↓	1.349e-01	3.210e-01	7.627e-02	1.742e-01	1.270e-01	9.917e-02	1.912e-01	1.716e-01	1.805e-01	2.517e-01	1.484e-01	2.158e-01	1.515e-01	3.392e-01
	E$^{loc}_{PCP}$ ↓	3.175e-02	2.469e-02	5.932e-02	4.546e-02	7.143e-02	2.479e-02	2.941e-02	2.985e-02	4.511e-02	2.721e-02	3.906e-02	5.036e-02	4.546e-02	1.111e-01
NUDT$_{TE}$ [17]	IoU$^{seg}_{pix}$ ↑	7.046e-01	7.813e-01	8.570e-01	8.347e-01	8.131e-01	8.380e-01	8.310e-01	8.321e-01	8.677e-01	8.745e-01	8.336e-01	8.557e-01	8.550e-01	8.304e-01
	E$^{seg}_{AUC}$ ↓	9.110e-04	1.810e-03	9.710e-04	1.046e-03	8.730e-04	9.670e-04	0.000e+00	1.046e-03	0.000e+00	0.000e+00	7.750e-04	0.000e+00	0.000e+00	9.900e-04
	E$^{seg}_{ITF}$ ↓	1.645e-01	1.171e-01	8.854e-02	9.513e-02	1.156e-01	7.499e-02	8.621e-02	9.827e-02	7.214e-02	7.323e-02	7.871e-02	9.872e-02	4.867e-02	8.605e-02
	E$^{seg}_{PCP}$ ↓	1.300e-01	9.976e-02	5.352e-02	6.914e-02	7.048e-02	8.606e-02	8.283e-02	6.854e-02	6.014e-02	5.225e-02	8.687e-02	4.556e-02	9.635e-02	8.258e-02
	IoU$^{loc}_{tgt}$ ↑	8.513e-01	7.934e-01	9.528e-01	8.891e-01	9.154e-01	9.467e-01	8.469e-01	8.586e-01	9.383e-01	8.330e-01	8.974e-01	9.382e-01	9.638e-01	6.582e-01
	E$^{loc}_{S2M}$ ↓	2.037e-03	3.861e-03	2.247e-03	2.092e-03	2.169e-03	2.222e-03	0.000e+00	2.020e-03	0.000e+00	0.000e+00	2.137e-03	0.000e+00	0.000e+00	1.582e-03
	E$^{loc}_{M2S}$ ↓	4.073e-03	7.722e-03	6.742e-03	6.276e-03	6.508e-03	1.778e-02	2.783e-02	1.616e-02	6.608e-03	1.179e-02	1.068e-02	6.623e-03	6.787e-03	7.911e-03
	E$^{loc}_{ITF}$ ↓	1.242e-01	1.660e-01	3.146e-02	9.833e-02	6.508e-02	3.111e-02	1.213e-01	1.192e-01	5.066e-02	1.473e-01	7.479e-02	4.856e-02	2.489e-02	3.149e-01
	E$^{loc}_{PCP}$ ↓	1.833e-02	2.896e-02	6.742e-03	4.184e-03	1.085e-02	2.222e-03	3.976e-03	4.040e-03	4.405e-03	7.859e-03	1.496e-02	6.623e-03	4.525e-03	1.741e-02

B Synthetic Data Experiment

To further investigate algorithm performance in complex scenarios with with densely distributed targets, we augment the testset of the existing IRSTD1k [41] using a randomized copy-paste augmentation strategy and construct the synthetic dataset, *i.e.*, IRSTD1k$^{AUG}_{TE}$. For copyright compliance, we exclusively performed data augmentation on the IRSTD1k dataset (MIT License), the only benchmark in our study with explicit redistribution permissions. All synthetic data will be publicly released under the same license terms (MIT License), providing legally compliant yet challenging test cases that faithfully extend the benchmark's inherent properties.

The strategy intelligently generates new target instances by:

1. Extract valid target regions from the ground-truth mask using connected component analysis.
2. Select targets for replication based on area-weighted sampling, prioritizing smaller targets.
3. Generate copies at $2\times$ the original target count, with a maximum of 7 new targets per image.
4. Augment each copy with random transformations including scaling (0.5-$1.1\times$), rotation ($0°$-$180°$), and positional perturbations (75% near original targets and 25% globally distributed).

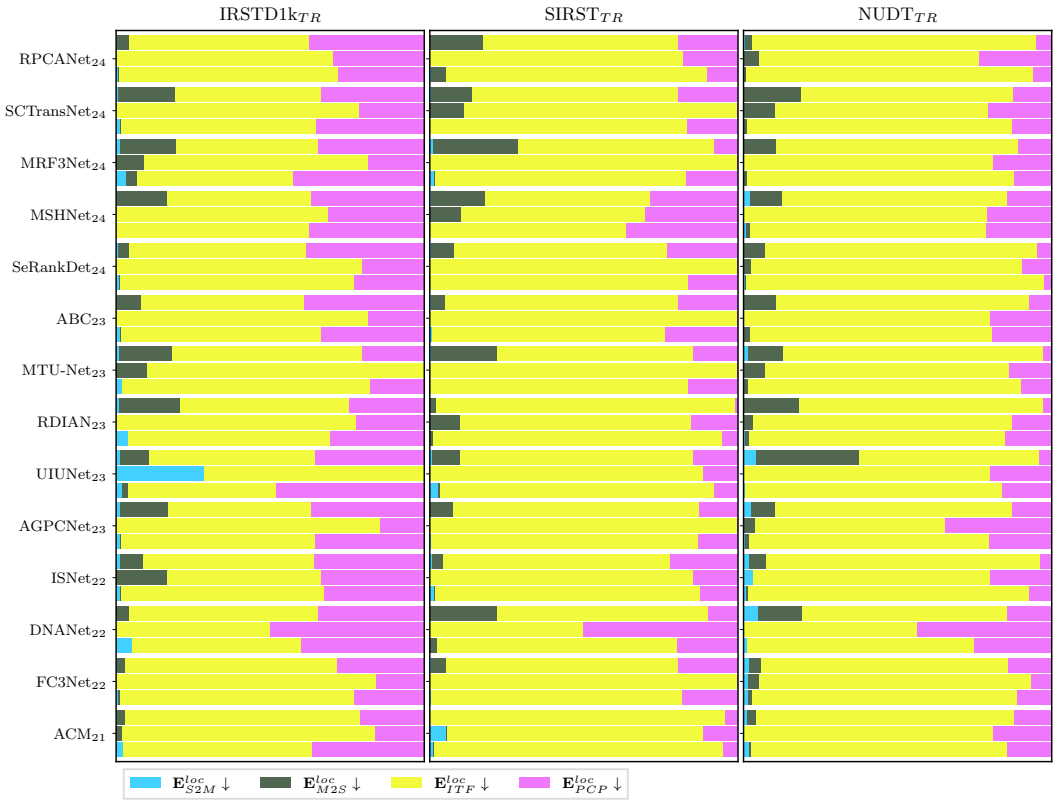


Figure 6: Cross-dataset localization error ratios corresponding to IRSTD1k$_{TE}$, SIRST$_{TE}$, and NUDT$_{TE}$ from bottom to top for the models trained on different datasets.

This approach realistically increases object density and morphological diversity, expanding the original dataset while maintaining scene coherence, thereby creating more challenging test conditions for evaluating algorithm robustness. And some samples are shown in Fig. 8.

As evidenced in Tab 4, existing models exhibit significant performance degradation on this new challenging dataset, with our error statistics showing marked increases across all error types. Fig. 9 provides a visual breakdown of how different error subtypes contribute to the overall performance decline. Notably, when comparing Fig. 9a with Fig. 6, we observe a substantial increase in the proportion of $\mathbf{E}^{loc}_{S2M}$, which directly validates our analysis about it in Sec. 4.3 of the main text. Similarly, the dramatic rise in $\mathbf{E}^{seg}_{MRG}$ shown in Fig. 9b versus Fig. 7 indicates these models struggle to effectively distinguish between individual target instances in high-density target distributions.

C Other Discussions

C.1 Further Error Analysis

Significant Value of Error Analysis. The goal of our error analysis is not merely to present total performance but to provide a more fine-grained understanding of model behavior, exposing the limitations and failure modes that are often hidden behind aggregated performance metrics. For instance, as demonstrated in Sec. B, Fig. 4 and Fig. 9 present the relative proportions of error components, which help identify which types of error dominate under different conditions. In dense-target scenarios, our analysis reveals that $\mathbf{E}^{seg}_{MRG}$ and $\mathbf{E}^{tgt}_{S2M}$ are the most significant contributors to failure, as discussed in Sec. 4.3. Furthermore, Tab. 3 provide quantitative error values and evaluate the severity of each error type in detail, enabling more targeted performance improvement strategies beyond overall score optimization.

Potential Impacts of Model Structure. Different structural choices in the model design clearly correlate with distinct error distributions. For instance, some models such as ACM [9] and FC3Net [40],

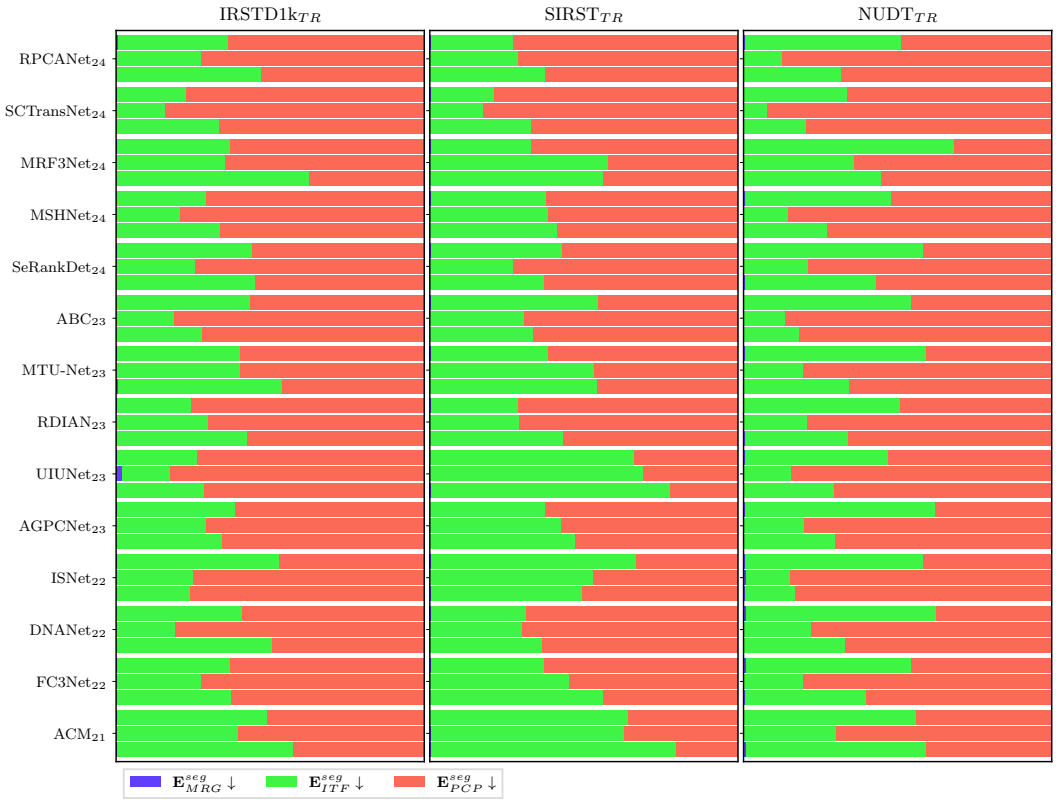
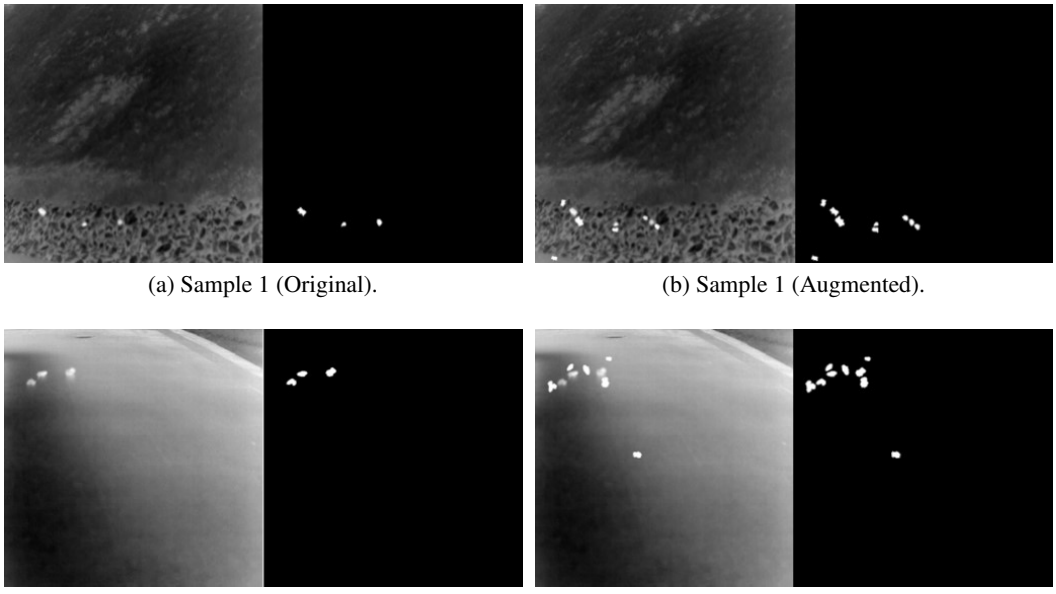


Figure 7: Cross-dataset segmentation error ratios corresponding to IRSTD1k$_{TE}$, SIRST$_{TE}$, and NUDT$_{TE}$ from bottom to top for the models trained on different datasets.



(a) Sample 1 (Original).

(b) Sample 1 (Augmented).

(c) Sample 2 (Original).

(d) Sample 2 (Augmented).

Figure 8: Visual demonstration of sample augmentation effects.

which lack explicit global modeling or deep semantic suppression, exhibit high $\mathbf{E}^{seg}_{ITF}$ and $\mathbf{E}^{loc}_{ITF}$ errors, often confusing textured backgrounds as targets. ISNet [41] and UIUNet [34], while integrating attention and multi-scale decoding, still suffer from structural inconsistencies and incorrect merging

Table 4: Performance and error analysis on the synthetic dataset IRSTD1k$_{TE}^{AUG}$ which utilize data augmentation strategies to significantly expand the number and morphology of targets on each image. Colors **red**, **green** and **blue** represent the first, second and third ranked results.

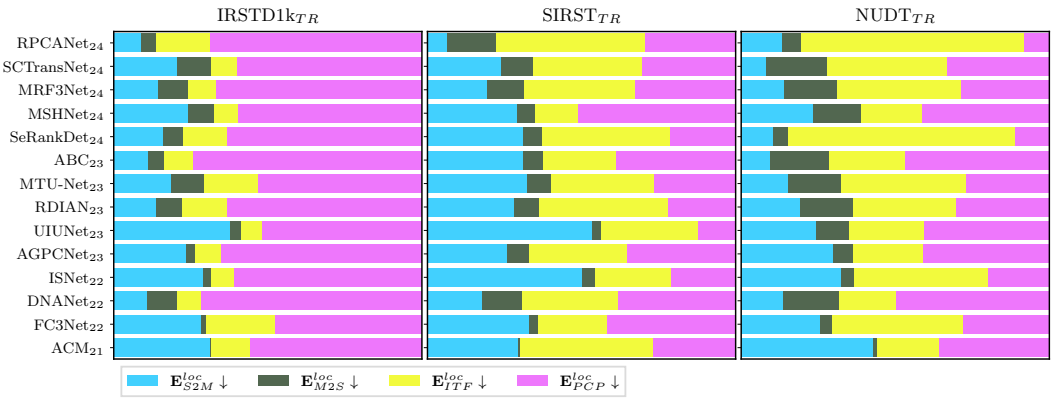
	ACM$_{21}$ [9]	FC3Net$_{22}$ [40]	DNANet$_{22}$ [17]	ISNet$_{22}$ [41]	AGPCNet$_{23}$ [42]	UIUNet$_{23}$ [34]	RDIAN$_{23}$ [29]	MTU-Net$_{23}$ [33]	ABC$_{23}$ [22]	SeRankDet$_{24}$ [6]	MSHNet$_{24}$ [20]	MRF3Net$_{24}$ [44]	SCTransNet$_{24}$ [37]	RPCANet$_{24}$ [32]
Params.	0.5M	6.9M	4.7M	1.1M	12.4M	50.5M	0.1M	4.1M	73.5M	108.9M	4.1M	0.5M	11.2M	0.7M
FLOPs	2.0G	10.6G	56.1G	121.9G	327.5G	217.9G	14.8G	24.4G	332.6G	568.7G	24.4G	33.2G	67.4G	179.7G
Trained on IRSTD1k$_{TR}$ [41].														
IoU$_{pix}$ ↑	0.272	0.315	0.398	0.307	0.372	**0.469**	0.372	0.417	0.361	0.456	**0.458**	**0.467**	0.431	0.371
nIoU$_{pix}$ ↑	0.327	0.414	0.491	0.359	0.425	0.489	0.468	0.495	0.446	**0.534**	**0.546**	**0.537**	0.492	0.369
F1$_{pix}$ ↑	0.428	0.479	0.569	0.470	0.543	**0.638**	0.542	0.588	0.531	0.626	**0.629**	**0.637**	0.602	0.541
Pd↑	0.313	0.451	0.633	0.424	0.526	0.598	0.622	**0.681**	0.588	0.663	**0.715**	0.667	**0.678**	0.552
+OPDC	0.407	0.541	0.669	0.502	0.577	0.664	0.666	**0.737**	0.629	0.695	**0.758**	0.700	**0.735**	0.605
Fa×10^6 ↓	267.446	388.321	**44.106**	127.954	82.120	122.450	66.596	71.056	**47.674**	68.703	**43.081**	57.752	67.811	74.548
+OPDC	59.365	234.556	**14.424**	14.727	14.784	26.475	24.103	22.034	**12.716**	31.542	**11.406**	19.909	15.126	24.976
hIoU↑	0.150	0.231	0.384	0.203	0.279	0.329	0.343	0.370	0.336	**0.408**	**0.426**	**0.406**	0.379	0.298
IoU$_{pix}^{seg}$ ↑	4.019e-01	4.898e-01	6.147e-01	4.276e-01	5.100e-01	5.144e-01	5.656e-01	5.525e-01	5.687e-01	6.344e-01	5.890e-01	6.199e-01	5.498e-01	5.493e-01
$\mathbf{E}_{MRG}^{seg}$ ↓	7.081e-02	6.669e-02	1.716e-02	5.151e-02	3.660e-02	4.259e-02	1.938e-02	2.138e-02	1.793e-02	2.042e-02	1.994e-02	1.603e-02	2.032e-02	1.983e-02
$\mathbf{E}_{IIF}^{seg}$ ↓	2.726e-01	1.513e-01	6.955e-02	1.085e-01	9.634e-02	7.270e-02	7.014e-02	8.126e-02	4.279e-02	7.339e-02	6.408e-02	1.050e-01	4.268e-02	5.748e-02
$\mathbf{E}_{PCP}^{seg}$ ↓	2.546e-01	2.922e-01	2.986e-01	4.124e-01	3.571e-01	3.703e-01	3.448e-01	3.448e-01	3.706e-01	2.718e-01	3.269e-01	2.591e-01	3.872e-01	3.734e-01
IoU$_{tgt}^{loc}$ ↑	3.739e-01	4.716e-01	6.250e-01	4.758e-01	5.469e-01	6.387e-01	6.057e-01	6.688e-01	5.913e-01	6.430e-01	7.236e-01	6.543e-01	6.900e-01	5.422e-01
$\mathbf{E}_{S2M}^{loc}$ ↓	1.955e-01	1.491e-01	4.130e-02	1.531e-01	1.069e-01	1.365e-01	5.391e-02	6.224e-02	4.590e-02	5.699e-02	6.659e-02	5.000e-02	6.332e-02	4.067e-02
$\mathbf{E}_{M2S}^{loc}$ ↓	2.137e-03	1.014e-02	3.587e-02	1.212e-02	1.213e-02	1.230e-02	3.383e-02	3.481e-02	2.076e-02	2.366e-02	2.331e-02	3.370e-02	3.493e-02	2.190e-02
$\mathbf{E}_{IIF}^{loc}$ ↓	7.906e-02	1.176e-01	2.935e-02	4.075e-02	3.969e-02	2.573e-02	5.708e-02	5.802e-02	3.934e-02	5.161e-02	2.220e-02	3.152e-02	2.620e-02	8.134e-02
$\mathbf{E}_{PCP}^{loc}$ ↓	3.494e-01	2.515e-01	2.685e-01	3.183e-01	2.944e-01	1.868e-01	2.495e-01	1.762e-01	3.027e-01	2.247e-01	1.643e-01	2.304e-01	1.856e-01	3.139e-01
Trained on SIRST$_{TR}$ [9].														
IoU$_{pix}$ ↑	0.213	0.472	0.520	0.430	0.496	0.517	0.518	**0.568**	0.504	**0.548**	0.459	**0.572**	0.540	0.492
nIoU$_{pix}$ ↑	0.323	0.445	0.535	0.442	0.490	0.523	0.513	**0.591**	0.555	**0.557**	0.497	**0.567**	0.554	0.468
F1$_{pix}$ ↑	0.351	0.641	0.685	0.601	0.663	0.681	0.682	**0.725**	0.670	**0.708**	0.629	**0.728**	0.701	0.659
Pd↑	0.324	0.515	**0.757**	0.458	0.658	0.519	0.651	0.735	0.721	0.745	0.665	**0.776**	**0.763**	0.755
+OPDC	0.452	0.605	**0.813**	0.584	0.726	0.643	0.728	0.792	0.770	0.802	0.719	**0.826**	**0.812**	**0.813**
Fa×10^6 ↓	1378.662	271.090	140.347	361.106	193.904	424.969	221.841	185.288	**118.388**	232.070	**99.695**	126.777	**124.348**	140.157
+OPDC	960.089	78.116	65.818	98.062	72.745	155.302	102.124	87.510	**52.628**	143.877	**36.287**	66.083	**65.097**	105.047
hIoU↑	0.124	0.275	0.414	0.235	0.339	0.298	0.317	**0.442**	**0.417**	0.415	0.368	**0.433**	0.415	0.347
IoU$_{pix}^{seg}$ ↑	3.914e-01	5.150e-01	5.840e-01	4.708e-01	5.480e-01	5.499e-01	5.546e-01	6.394e-01	5.956e-01	6.106e-01	5.478e-01	6.090e-01	5.916e-01	5.705e-01
$\mathbf{E}_{MRG}^{seg}$ ↓	7.867e-02	5.215e-02	1.875e-02	7.959e-02	3.263e-02	8.619e-02	4.768e-02	3.392e-02	3.214e-02	3.462e-02	2.758e-02	1.897e-02	2.815e-02	1.095e-02
$\mathbf{E}_{IIF}^{seg}$ ↓	3.441e-01	1.677e-01	6.384e-02	2.004e-01	1.463e-01	2.121e-01	1.053e-01	1.227e-01	7.095e-02	7.951e-02	9.872e-02	1.195e-01	6.383e-02	5.400e-02
$\mathbf{E}_{PCP}^{seg}$ ↓	1.858e-01	2.651e-01	3.334e-01	2.492e-01	2.730e-01	1.517e-01	2.924e-01	2.040e-01	3.013e-01	2.753e-01	3.259e-01	2.525e-01	3.164e-01	3.645e-01
IoU$_{tgt}^{loc}$ ↑	3.173e-01	5.333e-01	7.082e-01	4.995e-01	6.184e-01	5.422e-01	5.722e-01	6.914e-01	6.998e-01	6.805e-01	6.717e-01	7.114e-01	7.008e-01	6.078e-01
$\mathbf{E}_{S2M}^{loc}$ ↓	2.015e-01	1.549e-01	5.167e-02	2.517e-01	9.911e-02	2.451e-01	1.207e-01	1.005e-01	9.302e-02	9.961e-02	9.565e-02	5.611e-02	7.229e-02	2.522e-02
$\mathbf{E}_{M2S}^{loc}$ ↓	4.078e-03	1.231e-02	3.850e-02	2.090e-02	2.676e-02	1.274e-02	3.474e-02	2.335e-02	2.008e-02	1.972e-02	1.848e-02	3.507e-02	3.012e-02	6.261e-02
$\mathbf{E}_{IIF}^{loc}$ ↓	2.945e-01	1.056e-01	9.017e-02	1.234e-01	1.209e-01	1.441e-01	1.792e-01	1.036e-01	7.082e-02	1.321e-01	4.674e-02	1.032e-01	1.064e-01	1.896e-01
$\mathbf{E}_{PCP}^{loc}$ ↓	1.827e-01	1.938e-01	1.114e-01	1.045e-01	1.348e-01	5.588e-02	9.324e-02	8.122e-02	1.163e-01	6.805e-02	1.674e-01	9.419e-02	9.036e-02	1.148e-01
Trained on SIRST$_{TR}$ [9].														
IoU$_{pix}$ ↑	0.364	0.377	0.409	0.293	0.376	**0.440**	0.417	**0.445**	0.346	0.337	0.426	**0.439**	0.344	**0.439**
nIoU$_{pix}$ ↑	0.376	0.412	0.477	0.348	0.433	0.492	0.498	**0.512**	0.413	0.415	**0.507**	**0.535**	0.437	0.489
F1$_{pix}$ ↑	0.534	0.547	0.581	0.453	0.546	**0.611**	0.588	**0.616**	0.514	0.504	0.597	**0.610**	0.511	**0.610**
Pd↑	0.393	0.572	0.726	0.535	0.623	0.736	0.708	0.765	0.744	0.743	**0.780**	**0.772**	**0.771**	0.745
+OPDC	0.516	0.664	0.773	0.634	0.692	0.784	0.771	**0.820**	0.800	0.800	**0.823**	**0.820**	**0.821**	0.814
Fa×10^6 ↓	396.102	242.622	**87.909**	293.921	124.519	117.610	116.604	127.954	**93.166**	652.408	**74.700**	115.788	126.682	372.379
+OPDC	122.469	111.271	**51.717**	193.164	55.911	67.924	58.663	91.097	**49.610**	592.664	**31.485**	74.510	88.801	283.085
hIoU↑	0.196	0.251	0.368	0.192	0.300	**0.378**	0.340	0.369	0.313	0.271	**0.403**	**0.394**	0.321	0.282
IoU$_{pix}^{seg}$ ↑	4.295e-01	4.878e-01	5.378e-01	4.033e-01	4.885e-01	5.386e-01	5.448e-01	5.597e-01	4.524e-01	5.875e-01	5.369e-01	5.971e-01	4.912e-01	5.797e-01
$\mathbf{E}_{MRG}^{seg}$ ↓	7.636e-02	4.842e-02	1.391e-02	5.619e-02	3.714e-02	2.325e-02	2.890e-02	2.004e-02	1.073e-02	3.154e-02	1.715e-02	1.853e-02	9.236e-03	3.725e-02
$\mathbf{E}_{IIF}^{seg}$ ↓	2.678e-01	1.244e-01	4.590e-02	1.234e-01	9.219e-02	5.022e-02	5.870e-02	5.309e-02	2.186e-02	6.937e-02	5.065e-02	9.179e-02	2.851e-02	8.325e-02
$\mathbf{E}_{PCP}^{seg}$ ↓	2.264e-01	3.394e-01	4.024e-01	4.172e-01	3.821e-01	3.879e-01	3.676e-01	3.672e-01	5.150e-01	3.115e-01	3.953e-01	2.926e-01	4.710e-01	2.998e-01
IoU$_{tgt}^{loc}$ ↑	4.568e-01	5.144e-01	6.842e-01	4.752e-01	6.140e-01	7.021e-01	6.243e-01	6.589e-01	6.922e-01	4.617e-01	7.508e-01	6.601e-01	6.531e-01	4.858e-01
$\mathbf{E}_{S2M}^{loc}$ ↓	2.335e-01	1.243e-01	4.321e-02	1.700e-01	1.156e-01	7.292e-02	7.156e-02	5.234e-02	2.917e-02	5.570e-02	5.832e-02	4.775e-02	2.775e-02	6.870e-02
$\mathbf{E}_{M2S}^{loc}$ ↓	6.173e-03	1.892e-02	5.761e-02	2.267e-02	2.477e-02	3.125e-02	6.497e-02	5.794e-02	5.835e-02	2.617e-02	3.924e-02	5.805e-02	6.846e-02	3.192e-02
$\mathbf{E}_{IIF}^{loc}$ ↓	1.091e-01	2.063e-01	5.761e-02	2.276e-01	8.772e-02	7.292e-02	1.252e-01	1.383e-01	7.646e-02	3.966e-01	4.878e-02	1.367e-01	1.360e-01	3.713e-01
$\mathbf{E}_{PCP}^{loc}$ ↓	1.944e-01	1.360e-01	1.574e-01	1.046e-01	1.579e-01	1.208e-01	1.139e-01	9.252e-02	1.439e-01	5.973e-02	1.029e-01	9.738e-02	1.147e-01	4.233e-02

or splitting (*e.g.*, high $\mathbf{E}_{PCP}^{seg}$ or $\mathbf{E}_{S2M}^{loc}$/$\mathbf{E}_{M2S}^{loc}$ errors), reflecting limitations of shallow local attention. These observations highlight structural limitations and support the need for better context integration, stronger structural priors, and error-aware learning objectives.

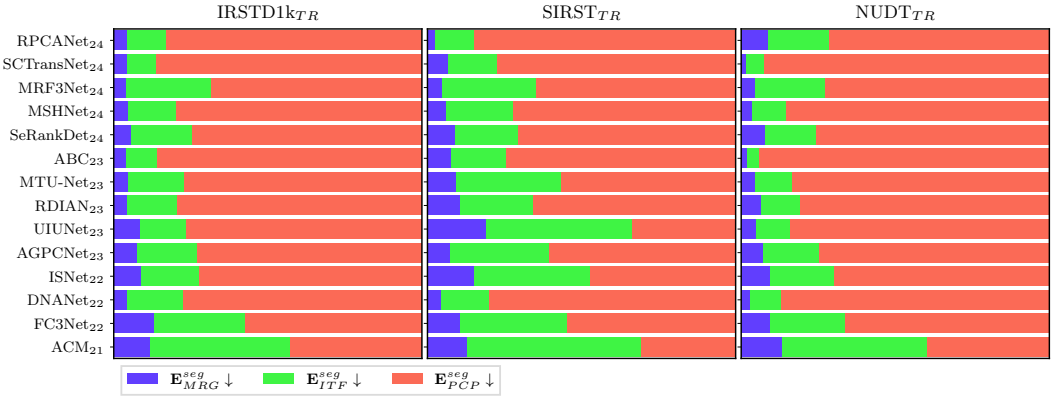
C.2 Why Use the Multiplicative Form?

We chose the multiplicative form $\text{hIoU} = \text{IoU}_{tgt}^{loc} \times \text{IoU}_{pix}^{seg}$ in Equ. 7 over the additive alternative $\text{aIoU} = 0.5(\text{IoU}_{tgt}^{loc} + \text{IoU}_{pix}^{seg})$ for modeling the interdependence of localization and segmentation in IRSTD:

1. Since $0 \leq \text{IoU}_{loc}, \text{IoU}_{seg} \leq 1$, both hIoU and aIoU lie in $[0, 1]$:
   (a) Both equal 1 only when both $\text{IoU}_{loc} = \text{IoU}_{seg} = 1$. $\text{hIoU} = 0$ if either component is 0, but $\text{aIoU} = 0$ only if both are 0.
   (b) $\text{hIoU} = \text{IoU}_{loc}\text{IoU}_{seg} > t \Rightarrow \text{IoU}_{loc} = \frac{t}{\text{IoU}_{seg}} > t$ and $\text{IoU}_{seg} = \frac{t}{\text{IoU}_{loc}} > t$, so any threshold t is enforced simultaneously. By contrast, $\text{aIoU} = 0.5(\text{IoU}_{loc} + \text{IoU}_{seg}) > t \Rightarrow \text{IoU}_{loc} > 2t - \text{IoU}_{seg}$ corresponds to the half-space, a much larger region that includes points where one coordinate can be far below t (*e.g.*, $(\text{IoU}_{loc}, \text{IoU}_{seg}) = (2t - 1, 1)$). Consequently, aIoU covers a broader area, over-approximating joint performance, whereas hIoU strictly requires both components to exceed t.
2. Both increase monotonically in each argument as follows:
   (a) $\frac{\partial \text{hIoU}}{\partial \text{IoU}_{loc}} = \text{IoU}_{seg}, \frac{\partial \text{hIoU}}{\partial \text{IoU}_{seg}} = \text{IoU}_{loc}$: The coupling means that when one component is low, the influence of the other is diminished, ensuring that isolated improvements cannot disproportionately boost the overall score.
   (b) $\frac{\partial \text{aIoU}}{\partial \text{IoU}_{loc}} = \frac{\partial \text{aIoU}}{\partial \text{IoU}_{seg}} = 0.5$: In contrast, aIoU assigns fixed, equal weight to each component, ignoring their interaction and thus diluting the impact of imbalance.
3. The total error can be written as follows:



(a) Localization errors.



(b) Segmentation errors.

Figure 9: Cross-dataset error ratios on the synthetic dataset IRSTD1k$^{AUG}_{TE}$ for the models trained on different datasets.

(a) For hIoU: $\mathbf{E}_{\text{hIoU}} = 1 - \text{hIoU} = 1 - (1 - \mathbf{E}^{loc})(1 - \mathbf{E}^{seg}) = \mathbf{E}^{loc} + \mathbf{E}^{seg} - \mathbf{E}^{loc}\mathbf{E}^{seg} =$
$\begin{cases} \mathbf{E}^{seg}(1 - \mathbf{E}^{loc}) + \mathbf{E}^{loc} \geq \mathbf{E}^{loc} \\ \mathbf{E}^{loc}(1 - \mathbf{E}^{seg}) + \mathbf{E}^{seg} \geq \mathbf{E}^{seg} \end{cases} \Rightarrow \max(\mathbf{E}^{loc}, \mathbf{E}^{seg}) \leq \mathbf{E}_{\text{hIoU}} \leq \min(1, \mathbf{E}^{loc} + \mathbf{E}^{seg})$

(b) For aIoU: $\mathbf{E}_{\text{aIoU}} = 1 - \text{aIoU} = 0.5(\mathbf{E}^{loc} + \mathbf{E}^{seg})$.

(c) $\mathbf{E}_{\text{hIoU}}$ intuitively reflects that "the shortcoming dictates the performance ceiling". In contrast, $\mathbf{E}_{\text{aIoU}}$ is always $0.5(\mathbf{E}^{loc}+\mathbf{E}^{seg})$, lacking the shortcoming effect. Thus, $\mathbf{E}_{\text{hIoU}}$ more faithfully captures the coupled relationship in IRSTD.

The multiplicative form naturally penalizes any weak link, making it a more faithful, robust metric for IRSTD than the additive aIoU. No high score can "hide" a bad component, whereas the additive form can still report misleadingly high scores even if one factor is near zero, thus failing to reflect genuine end-to-end performance.

C.3 Why Use the IoU-based Form?

Our choice to adopt an IoU-based formulation is intentional and motivated by both practical and conceptual considerations.

1. Using the IoU form ensures consistency with the segmentation IoU, thereby yielding a uniform metric structure across both pixel and target levels. This consistency simplifies the interpretation and facilitates a coherent hierarchical error decomposition framework. Specifically:
   (a) It enables pixel-level and target-level components to have the same variation trends and intrinsic meanings.

(b) It allows a unified performance and error analysis principle based on set intersection-over-union, to be applied at different levels.

(c) It can further simplify the final computation of the hIoU, as the term $\sum_{i=1}^{K} |\mathrm{TP}_{tgt}^{[i]}|$ can be cleanly canceled out due to structural consistency.

2. IoU has the added advantage of directly reflecting spatial overlap, which is particularly meaningful for tasks involving localization and segmentation. This makes it more intuitive and analytically convenient in the context of IRSTD.

C.4 Extended Discussion on F1-score

Target-level F1-score ($\mathbf{F1}_{tgt}$). As stated in the main text, the F1 we used follows the pixel-level formulation that has been widely adopted in IRSTD [37]. Therefore, our implementation of F1 is aligned with the precedent set by prior literature. F1/precision/recall can also be formulated at the target level. Notably, the recall in this context corresponds to the commonly used target-level IRSTD metric Pd. To provide a more comprehensive evaluation, we additionally supplement the target-level $\mathrm{F1}_{tgt}$ in Tab. 5.

Why not $\mathbf{F1}_{tgt} \times \mathbf{nIoU}_{pix}$? There are fundamental differences in the way they handle error attribution. The product of target-level $\mathrm{F1}_{tgt}$ and pixel-level nIoU_{pix} can lead to redundant penalization, as it implicitly assumes independence between segmentation and localization errors. However, the two types of errors are often correlated, for instance, inaccurate localization will simultaneously degrade whole segmentation performance. As a result, this form tends to double-count the impact of shared failure sources, and thus does not faithfully reflect the overall performance. In contrast, our hIoU is designed to disentangle these two layers of performance. Localization quality is measured first, and any performance loss due to missed or poorly localized targets is entirely attributed to the target-level. Segmentation quality is then evaluated only within the region of correctly matched targets, focusing solely on the spatial overlap and avoiding entanglement with localization failure. This design ensures that each layer is evaluated in a complementary manner and errors are attributed unambiguously to their true source. Consequently, hIoU offers a more accurate, interpretable, and fair assessment of the model's overall performance, avoiding the distortion introduced by overlapping error contributions.

C.5 Multi-Frame Infrared Small Target Detection

Some advances leverage temporal cues (*i.e.*, multi-frame IRSTD) to enhance robustness against clutter, including motion direction encoding [18] and recurrent refinement with motion compensation [35], effectively exploiting spatiotemporal dynamics. Recent work [19] explores the temporal-profile perspective by reformulating detection as a one-dimensional anomaly task, offering high efficiency. These works highlight the emerging shift from purely spatial designs toward spatio-temporal paradigms.

Our evaluation framework is agnostic to the detection paradigm and is designed to be equally applicable to both single-frame and multi-frame IRSTD methods. In particular, the proposed metrics and analysis mechanisms remain consistent across these settings without any need for structural adaptation. In Tab. 6, we conduct experiments on a representative multi-frame dataset [18] and evaluate several state-of-the-art multi-frame IRSTD methods [18, 35, 19] using our proposed framework.

C.6 Validation under Occlusion, Deformation, and Connectivity

Our work focuses on the IRSTD task, where the targets often exhibit significant boundary ambiguity and shape diversity, as shown in Fig. 5 and Fig. 8. Our OPDC strategy does not rely on the content of the input image (*e.g.*, weather conditions or image quality), but instead operates solely on the binary prediction and ground-truth masks. It incorporates both distance-based and region-overlap constraints, which helps reduce interference caused by target occlusion, deformation, or connectivity. We also supplement experiments of OPDC under these challenging conditions by creating three data subsets with random target occlusion, deformation, and connectivity. The proposed OPDC strategy is compared with the original distance-based approach [17]. The ratio of predefined target pairs that are successfully matched is summarized in Tab. 7. This experiment demonstrates that our OPDC performs better than the distance-based method [17].

Table 5: Cross-dataset performance analysis. Colors **red**, **green** and **blue** represent the first, second and third ranked results. **Besides the reported metrics, the target-level F1-score ($F1_{tgt}$) is included in the content of Tab. 1.**

		ACM$_{21}$ [9]	FC3Net$_{22}$ [40]	DNANet$_{22}$ [17]	ISNet$_{22}$ [41]	AGPCNet$_{23}$ [42]	UIUNet$_{23}$ [34]	RDIAN$_{23}$ [29]	MTU-Net$_{23}$ [33]	ABC$_{23}$ [22]	SeRankDet$_{24}$ [6]	MSHNet$_{24}$ [20]	MRF3Net$_{24}$ [44]	SCTransNet$_{24}$ [37]	RPCANet$_{24}$ [32]
Params.		0.5M	6.9M	4.7M	1.1M	12.4M	50.5M	0.1M	4.1M	73.5M	108.9M	4.1M	0.5M	11.2M	0.7M
FLOPs		2.0G	10.6G	56.1G	121.9G	327.5G	217.9G	14.8G	24.4G	332.6G	568.7G	24.4G	33.2G	67.4G	179.7G
Trained on IRSTD1K$_{TR}$ [41].															
IRSTD1K$_{TE}$ [41]	IoU$_{pix}$ ↑	0.439	0.358	0.637	0.578	0.605	0.570	0.603	0.610	0.624	0.642	0.650	0.636	0.644	0.608
	nIoU$_{pix}$ ↑	0.476	0.531	0.625	0.518	0.580	0.600	0.605	0.607	0.595	0.621	0.620	0.630	0.622	0.579
	F1$_{pix}$ ↑	0.610	0.527	0.778	0.733	0.754	0.726	0.753	0.757	0.768	0.782	0.788	0.777	0.783	0.756
	F1$_{tgt}$ ↑	0.761	0.757	0.905	0.888	0.890	0.906	0.869	0.851	0.885	0.863	0.913	0.891	0.887	0.830
	+OPDC	0.796	0.766	0.908	0.894	0.894	0.912	0.878	0.860	0.888	0.870	0.916	0.902	0.897	0.831
	Pd↑	0.798	0.865	0.912	0.919	0.916	0.906	0.902	0.929	0.916	0.926	0.933	0.899	0.912	0.886
	+OPDC	0.835	0.875	0.916	0.926	0.919	0.912	0.912	0.939	0.919	0.933	0.936	0.909	0.923	0.896
	Fa×10^6 ↓	95.178	237.365	13.854	13.266	15.354	51.147	21.503	28.012	16.815	44.638	11.539	17.441	16.834	28.145
	+OPDC	61.187	233.266	10.476	11.444	11.862	44.277	17.062	23.647	13.475	41.108	7.686	12.146	10.476	21.844
	hIoU↑	0.356	0.383	0.557	0.443	0.496	0.530	0.511	0.493	0.508	0.520	0.549	0.553	0.537	0.470
SIRST$_{TE}$ [9]	IoU$_{pix}$ ↑	0.472	0.234	0.676	0.712	0.763	0.696	0.658	0.701	0.708	0.734	0.649	0.752	0.629	0.543
	nIoU$_{pix}$ ↑	0.567	0.595	0.733	0.719	0.735	0.688	0.735	0.749	0.737	0.742	0.699	0.763	0.686	0.676
	F1$_{pix}$ ↑	0.642	0.380	0.807	0.832	0.866	0.821	0.794	0.824	0.829	0.847	0.787	0.858	0.772	0.704
	F1$_{tgt}$ ↑	0.786	0.688	0.963	0.973	0.969	0.950	0.921	0.939	0.942	0.955	0.929	0.951	0.917	0.890
	+OPDC	0.802	0.696	0.963	0.973	0.969	0.968	0.921	0.956	0.951	0.955	0.929	0.951	0.917	0.882
	Pd↑	0.908	0.872	0.963	0.982	0.991	0.963	0.963	0.982	0.972	0.982	0.954	0.982	0.963	0.927
	+OPDC	0.927	0.881	0.963	0.982	0.991	0.982	0.963	1.000	0.982	0.982	0.954	0.982	0.963	0.927
	Fa×10^6 ↓	127.650	932.797	3.754	5.632	2.560	86.351	17.407	42.152	22.526	17.236	11.946	4.608	20.820	124.066
	+OPDC	121.165	932.456	3.754	5.632	2.560	30.376	17.407	38.397	21.844	17.236	11.946	4.608	20.820	124.066
	hIoU↑	0.418	0.390	0.687	0.674	0.682	0.644	0.645	0.693	0.672	0.684	0.623	0.694	0.596	0.598
NUDT$_{TE}$ [17]	IoU$_{pix}$ ↑	0.331	0.288	0.504	0.443	0.468	0.450	0.441	0.416	0.469	0.494	0.463	0.512	0.377	0.291
	nIoU$_{pix}$ ↑	0.452	0.443	0.627	0.538	0.556	0.541	0.554	0.528	0.582	0.581	0.561	0.609	0.502	0.430
	F1$_{pix}$ ↑	0.498	0.448	0.670	0.614	0.638	0.620	0.612	0.588	0.638	0.661	0.633	0.678	0.548	0.451
	F1$_{tgt}$ ↑	0.588	0.648	0.806	0.710	0.749	0.797	0.713	0.641	0.776	0.783	0.732	0.771	0.689	0.680
	+OPDC	0.613	0.678	0.841	0.749	0.809	0.824	0.771	0.712	0.805	0.794	0.798	0.813	0.751	0.662
	Pd↑	0.757	0.752	0.848	0.759	0.785	0.836	0.804	0.769	0.808	0.820	0.771	0.815	0.748	0.701
	+OPDC	0.792	0.787	0.886	0.806	0.850	0.867	0.871	0.857	0.841	0.832	0.843	0.862	0.818	0.722
	Fa×10^6 ↓	253.092	273.488	121.206	71.920	40.690	114.695	71.869	96.690	133.464	63.883	65.562	43.844	110.728	190.989
	+OPDC	236.308	266.469	114.899	57.068	31.789	106.049	62.002	79.753	124.563	59.916	47.913	35.502	92.723	187.581
	hIoU↑	0.274	0.333	0.526	0.434	0.456	0.457	0.408	0.366	0.484	0.466	0.445	0.484	0.393	0.344
Trained on SIRST$_{TR}$ [9].															
IRSTD1K$_{TE}$ [41]	IoU$_{pix}$ ↑	0.104	0.456	0.564	0.498	0.518	0.444	0.382	0.545	0.574	0.549	0.581	0.581	0.550	0.492
	nIoU$_{pix}$ ↑	0.306	0.469	0.556	0.495	0.481	0.470	0.470	0.544	0.544	0.543	0.542	0.550	0.519	0.490
	F1$_{pix}$ ↑	0.188	0.626	0.721	0.665	0.683	0.615	0.552	0.705	0.730	0.709	0.735	0.735	0.710	0.659
	F1$_{tgt}$ ↑	0.406	0.720	0.810	0.747	0.749	0.713	0.441	0.768	0.822	0.771	0.858	0.764	0.777	0.656
	+OPDC	0.429	0.734	0.816	0.752	0.754	0.724	0.446	0.771	0.825	0.777	0.861	0.767	0.779	0.676
	Pd↑	0.818	0.865	0.912	0.909	0.912	0.912	0.879	0.909	0.902	0.909	0.892	0.899	0.912	0.879
	+OPDC	0.872	0.882	0.919	0.916	0.919	0.929	0.889	0.912	0.906	0.916	0.896	0.906	0.916	0.909
	Fa×10^6 ↓	1784.007	139.018	66.672	118.806	84.815	178.835	187.907	88.364	60.959	85.537	44.524	56.822	72.005	88.630
	+OPDC	1704.411	129.073	61.813	115.124	81.608	160.730	182.935	87.947	60.409	81.399	44.144	52.666	71.663	81.342
	hIoU↑	0.124	0.334	0.435	0.370	0.356	0.336	0.172	0.408	0.443	0.413	0.459	0.398	0.400	0.313
SIRST$_{TE}$ [9]	IoU$_{pix}$ ↑	0.607	0.651	0.708	0.746	0.787	0.738	0.670	0.805	0.775	0.785	0.715	0.777	0.764	0.690
	nIoU$_{pix}$ ↑	0.664	0.740	0.779	0.759	0.771	0.723	0.751	0.796	0.793	0.800	0.754	0.775	0.773	0.742
	F1$_{pix}$ ↑	0.756	0.788	0.829	0.854	0.881	0.849	0.803	0.892	0.873	0.879	0.834	0.875	0.866	0.817
	F1$_{tgt}$ ↑	0.904	0.924	0.963	0.960	0.965	0.960	0.914	0.952	0.969	0.969	0.955	0.956	0.952	0.930
	+OPDC	0.922	0.924	0.972	0.969	0.965	0.960	0.914	0.952	0.969	0.969	0.955	0.956	0.960	0.926
	Pd↑	0.954	1.000	0.963	0.982	1.000	0.991	0.972	1.000	1.000	1.000	0.972	1.000	0.991	0.972
	+OPDC	0.972	1.000	0.972	0.991	1.000	0.991	0.972	1.000	1.000	1.000	0.972	1.000	1.000	0.972
	Fa×10^6 ↓	92.836	121.506	4.778	35.155	8.362	13.994	75.771	14.335	31.571	31.571	25.598	13.482	7.679	51.367
	+OPDC	51.879	121.506	3.584	22.526	8.362	13.994	75.771	14.335	31.571	31.571	25.598	13.482	6.485	51.367
	hIoU↑	0.584	0.655	0.736	0.720	0.723	0.679	0.655	0.732	0.747	0.755	0.699	0.715	0.712	0.675
NUDT$_{TE}$ [17]	IoU$_{pix}$ ↑	0.121	0.361	0.521	0.463	0.458	0.482	0.226	0.502	0.599	0.586	0.539	0.543	0.474	0.355
	nIoU$_{pix}$ ↑	0.373	0.461	0.635	0.552	0.563	0.549	0.527	0.586	0.663	0.644	0.616	0.643	0.572	0.467
	F1$_{pix}$ ↑	0.215	0.530	0.685	0.633	0.628	0.650	0.369	0.668	0.749	0.739	0.701	0.704	0.643	0.524
	F1$_{tgt}$ ↑	0.357	0.604	0.641	0.716	0.637	0.708	0.172	0.695	0.765	0.760	0.797	0.695	0.701	0.615
	+OPDC	0.366	0.665	0.703	0.744	0.669	0.764	0.192	0.747	0.771	0.770	0.815	0.735	0.749	0.687
	Pd↑	0.855	0.757	0.841	0.832	0.846	0.846	0.825	0.846	0.888	0.867	0.864	0.890	0.827	0.750
	+OPDC	0.879	0.836	0.923	0.867	0.888	0.916	0.923	0.909	0.895	0.879	0.883	0.946	0.883	0.848
	Fa×10^6 ↓	2600.861	150.604	114.594	115.000	142.008	118.510	1013.184	100.352	58.492	42.979	70.089	97.198	91.553	139.211
	+OPDC	2571.971	139.567	100.352	111.084	137.126	109.049	996.348	82.855	56.661	41.453	66.427	90.078	84.686	124.003
	hIoU↑	0.125	0.317	0.379	0.419	0.340	0.409	0.070	0.404	0.492	0.474	0.481	0.401	0.406	0.328
Trained on NUDT$_{TR}$ [17].															
IRSTD1K$_{TE}$ [41]	IoU$_{pix}$ ↑	0.340	0.398	0.420	0.259	0.443	0.464	0.448	0.450	0.441	0.214	0.405	0.414	0.392	0.270
	nIoU$_{pix}$ ↑	0.420	0.425	0.506	0.342	0.479	0.522	0.523	0.506	0.492	0.341	0.527	0.511	0.484	0.409
	F1$_{pix}$ ↑	0.508	0.569	0.591	0.411	0.614	0.633	0.619	0.621	0.612	0.352	0.576	0.585	0.564	0.425
	F1$_{tgt}$ ↑	0.697	0.623	0.755	0.533	0.760	0.742	0.686	0.656	0.780	0.452	0.793	0.672	0.693	0.464
	+OPDC	0.716	0.630	0.761	0.541	0.775	0.748	0.705	0.666	0.789	0.448	0.796	0.674	0.701	0.475
	Pd↑	0.862	0.859	0.852	0.865	0.875	0.892	0.862	0.892	0.896	0.943	0.896	0.889	0.892	0.855
	+OPDC	0.886	0.879	0.859	0.882	0.892	0.899	0.886	0.906	0.906	0.943	0.902	0.896	0.902	0.879
	Fa×10^6 ↓	251.409	134.596	76.408	268.699	79.710	107.741	57.904	112.847	58.530	678.048	104.724	81.247	89.465	265.985
	+OPDC	226.149	113.872	73.504	257.919	70.695	92.255	49.325	106.508	54.810	678.048	101.346	79.843	87.112	256.970
	hIoU↑	0.306	0.279	0.383	0.186	0.366	0.382	0.342	0.322	0.375	0.192	0.408	0.325	0.321	0.192
SIRST$_{TE}$ [9]	IoU$_{pix}$ ↑	0.605	0.557	0.645	0.572	0.587	0.675	0.638	0.603	0.611	0.615	0.587	0.625	0.569	0.484
	nIoU$_{pix}$ ↑	0.668	0.627	0.716	0.621	0.660	0.720	0.720	0.687	0.678	0.704	0.670	0.709	0.632	0.560
	F1$_{pix}$ ↑	0.754	0.716	0.784	0.728	0.740	0.806	0.779	0.752	0.759	0.762	0.740	0.770	0.725	0.652
	F1$_{tgt}$ ↑	0.909	0.779	0.927	0.863	0.885	0.925	0.871	0.870	0.873	0.825	0.897	0.846	0.877	0.703
	+OPDC	0.909	0.779	0.927	0.872	0.885	0.934	0.871	0.879	0.873	0.833	0.897	0.846	0.877	0.690
	Pd↑	0.963	0.954	0.936	0.927	0.917	0.963	0.963	0.954	0.945	0.954	0.954	0.936	0.945	0.826
	+OPDC	0.963	0.963	0.936	0.936	0.917	0.972	0.963	0.963	0.945	0.963	0.954	0.936	0.945	0.826
	Fa×10^6 ↓	33.960	56.316	8.191	42.152	15.188	22.356	15.700	26.793	17.407	48.125	19.796	25.939	11.605	52.220
	+OPDC	33.960	55.463	8.191	20.820	15.188	19.967	15.700	25.939	17.407	47.613	19.796	25.939	11.605	53.586
	hIoU↑	0.575	0.441	0.637	0.491	0.557	0.648	0.566	0.557	0.555	0.526	0.559	0.554	0.511	0.372
NUDT$_{TE}$ [17]	IoU$_{pix}$ ↑	0.635	0.693	0.831	0.790	0.778	0.840	0.803	0.809	0.854	0.840	0.790	0.830	0.856	0.747
	nIoU$_{pix}$ ↑	0.668	0.723	0.850	0.813	0.796	0.844	0.826	0.821	0.862	0.854	0.817	0.847	0.856	0.772
	F1$_{pix}$ ↑	0.776	0.819	0.908	0.882	0.875	0.913	0.891	0.895	0.921	0.913	0.882	0.907	0.923	0.855
	F1$_{tgt}$ ↑	0.917	0.878	0.974	0.937	0.956	0.970	0.911	0.924	0.966	0.909	0.939	0.966	0.977	0.793
	+OPDC	0.920	0.885	0.976	0.941	0.956	0.973	0.917	0.924	0.968	0.909	0.946	0.968	0.982	0.794
	Pd↑	0.974	0.953	0.988	0.988	0.986	0.993	0.988	0.993	0.993	0.991	0.974	0.991	0.991	0.970
	+OPDC	0.977	0.960	0.991	0.993	0.986	0.995	0.995	0.993	0.995	0.991	0.981	0.993	0.995	0.972
	Fa×10^6 ↓	46.844	46.539	11.698	17.700	13.987	7.222	22.125	20.549	12.309	31.586	30.518	10.071	3.916	57.576
	+OPDC	42.979	36.621	9.003	16.886	13.987	5.442	19.786	20.091	10.325	31.586	20.142	7.782	2.391	57.271
	hIoU↑	0.600	0.620	0.817	0.742	0.744	0.793	0.704	0.714	0.814	0.728	0.748	0.803	0.824	0.547

Table 6: Results of multi-frame IRSTD methods.

	hIoU↑	IoU$^{seg}_{pix}$ ↑	$\mathbf{E}^{seg}_{MRG}$ ↓	$\mathbf{E}^{seg}_{ITF}$ ↓	$\mathbf{E}^{seg}_{PCP}$ ↓	IoU$^{loc}_{tgt}$ ↑	$\mathbf{E}^{loc}_{S2M}$ ↓	$\mathbf{E}^{loc}_{M2S}$ ↓	$\mathbf{E}^{loc}_{ITF}$ ↓	$\mathbf{E}^{loc}_{PCP}$ ↓
[18]	0.552	0.832	0.000	0.107	0.061	0.683	0.000	0.010	0.189	0.119
[35]	0.546	0.816	0.000	0.117	0.067	0.669	0.000	0.012	0.195	0.123
[19]	0.591	0.822	0.000	0.114	0.064	0.719	0.000	0.006	0.151	0.124

Table 7: Success rate of predefined target pair matching.

	Occlusion	Deformation	Connectivity
OPDC	1.000	1.000	0.949
Distance	0.420	0.320	0.379

C.7 Validation on Medical Small Object Detection

To verify its applicability, we transfer the proposed framework to the medical image domain, specifically, the polyp segmentation task, which also involves small and irregular targets. We evaluate the classic method [11] in Tab. 8 using our framework and observe clear localization-related errors, including large values in $\mathbf{E}^{loc}_{ITF}$ and $\mathbf{E}^{loc}_{PCP}$, corresponding to missed detections and incorrect background predictions. These issues are also visually evident in its predictions, confirming that established methods can also suffer from target-level limitations not reflected in traditional metrics. This extension further shows the generality and diagnostic value of our framework beyond IRSTD.

Table 8: Experiments on medical small object detection, *i.e.*, the polyp segmentation task.

	hIoU↑	IoU^{seg}_{pix} ↑	$\mathbf{E}^{seg}_{MRG}$ ↓	$\mathbf{E}^{seg}_{ITF}$ ↓	$\mathbf{E}^{seg}_{PCP}$ ↓	IoU^{loc}_{tgt} ↑	$\mathbf{E}^{loc}_{S2M}$ ↓	$\mathbf{E}^{loc}_{M2S}$ ↓	$\mathbf{E}^{loc}_{ITF}$ ↓	$\mathbf{E}^{loc}_{PCP}$ ↓
[11]	0.589	0.881	0.003	0.050	0.066	0.669	0.000	0.000	0.143	0.188

C.8 Validation on Different Target Attributes

Since our metrics are computed based solely on binary prediction and ground truth masks, they do not rely on the input image itself and the target contrast in the image has no influence on the evaluation process. To analyze the influence of target size and density attributes, we manually adjust the size and spatial density of targets and the results are reported in Tab. 9. The results show that while the metrics are generally stable, changes in target size or density do lead to observable variations in hIoU. This behavior is expected and reasonable: modifying target size or density can alter the relative distances and overlaps between targets, which directly affects localization accuracy, segmentation overlap, and ultimately the joint hIoU score.

Table 9: Experiments on different target attributes, including size and density.

	hIoU↑	IoU^{seg}_{pix} ↑	$\mathbf{E}^{seg}_{MRG}$ ↓	$\mathbf{E}^{seg}_{ITF}$ ↓	$\mathbf{E}^{seg}_{PCP}$ ↓	IoU^{loc}_{tgt} ↑	$\mathbf{E}^{loc}_{S2M}$ ↓	$\mathbf{E}^{loc}_{M2S}$ ↓	$\mathbf{E}^{loc}_{ITF}$ ↓	$\mathbf{E}^{loc}_{PCP}$ ↓
Original	0.430	0.632	0.000	0.182	0.186	0.687	0.000	0.000	0.207	0.106
Smaller (Erosion)	0.422	0.614	0.002	0.192	0.192	0.686	0.000	0.000	0.206	0.108
Larger (Dilation)	0.437	0.633	0.001	0.180	0.186	0.690	0.000	0.000	0.209	0.101
Sparser	0.424	0.622	0.001	0.188	0.189	0.682	0.000	0.000	0.208	0.110
Denser	0.441	0.634	0.000	0.176	0.190	0.696	0.000	0.000	0.207	0.097

In addition, we conduct experiments in Sec. B by creating synthetic scenarios with more and denser small targets. As shown in the comparison between the original dataset (Tab. 1) and the synthetic dataset (Tab. 4), the average differences between other metrics and hIoU are listed as Tab. 10. These results show that, except for Pd and Fa, the performance gap between hIoU and other metrics remains relatively stable. The changes in Pd and Fa may be related to their strong dependence on target matching, which is itself sensitive to variations in target number and density.

Table 10: Average differences between other metrics and hIoU.

IoU_{pix} ↑	nIoU_{pix} ↑	F1_{pix} ↑	Pd↑	+OPDC↑	Fa×10^6 ↓	+OPDC↑
-0.003	0.014	0.006	-0.195	-0.142	58.519	-28.424

C.9 Experiments on NUDT-SIRST-Sea

The selected datasets in the main text are the most commonly-used benchmarks in IRSTD. As shown in Tab. 1, existing methods still exhibit large performance variations across these datasets, suggesting inherent distribution differences in data. We additionally introduce the NUDT-SIRST-Sea dataset [33], which features space-based infrared imagery of tiny ships and drastically different with existing data. Results of RPCANet [32] on four datasets as listed in Tab. 11 show a significant performance drop under this distribution shift, confirming both the sensitivity of current methods and the usefulness of our framework in revealing such robustness issues.

To better characterize the differences in data distributions, we analyze and compare the four testing datasets, *i.e.*, IRSTD1k_{TE} [41], SIRST_{TE} [9], NUDT_{TE} [17], and $\text{NUDT-SIRST-Sea}_{TE}$ [33]. A set of hand-crafted statistical features from each image, including image attributes (*i.e.*, brightness, contrast, and noise estimation) and target attributes (*i.e.*, count, size, contrast, and area ratio) are

Table 11: Results of RPCANet [32] on four datasets.

	IRSTD1k$_{TE}$ [41]	NUDT$_{TE}$ [17]	SIRST$_{TE}$ [9]	NUDT-SIRST-Sea$_{TE}$ [33]
IoU$_{pix}$ ↑	0.608	0.291	0.543	0.001
F1$_{pix}$ ↑	0.756	0.451	0.704	0.001
Pd↑	0.886	0.701	0.927	0.303
Fa×10^6 ↓	28.145	190.989	124.066	56666.629
hIoU↑	0.470	0.344	0.598	0.017

Table 12: Attribute statistics for four datasets.

	Image Attributes				Target Attributes			
	Brightness Mean	Brightness Standard Deviation	Contrast Root Mean Square Contrast	Laplacian-based Noise Estimation	Average Target Count	Average Target Size	Target-Background Contrast	Foreground-Background Area Ratio
IRSTD1k$_{TE}$ [41]	0.344	0.149	0.149	43.436	1.477	51.148	131.865	1.786×10^4
SIRST$_{TE}$ [9]	0.428	0.098	0.098	40.068	1.267	30.289	1.936	3.795×10^4
NUDT$_{TE}$ [17]	0.419	0.127	0.127	110.505	1.426	34.703	1.525	4.263×10^4
NUDT-SIRST-Sea$_{TE}$ [33]	0.244	0.076	0.076	66.729	2.238	10.011	0.881	1.761×10^4

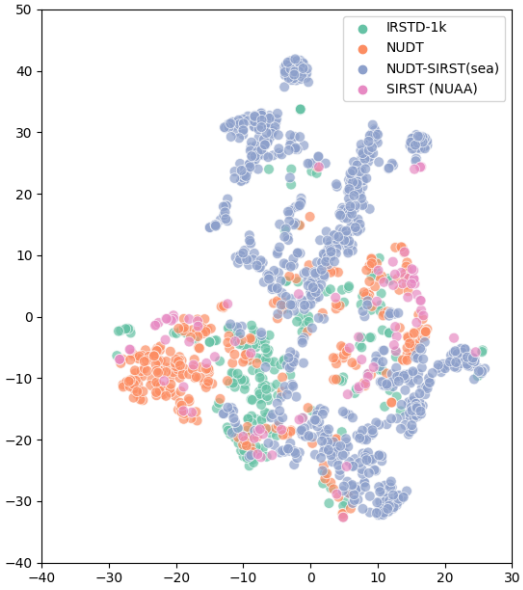


Figure 10: Illustration of attribute distribution for four datasets.

extracted and summarized in Tab. 12. They can be used to visualize the overall data distribution of these datasets based on t-SNE, as shown in Fig. 10. The visualization offers an interpretable and model-agnostic way to clear distribution differences through images and target attributes. While partial overlaps exist, each dataset occupies distinct regions, highlighting cross-dataset heterogeneity. This underscores the necessity of cross-dataset analysis to avoid bias and improve generalization in the IRSTD research.

D Societal Impacts

This work rethinks the evaluation protocols in IRSTD, aiming to enhance research transparency and fairness through the hierarchical analysis framework and open-source benchmarking tool, ultimately supporting the development of more reliable systems for real-world deployment. Furthermore, through systematic error analysis, this work identifies critical failure modes, which may help facilitate the exploration of necessary risk mitigation strategies in safety-sensitive applications.