# HyTBE: Hyperbolic Target-Background Expert Model for Cross-Domain Infrared Small Target Detection

**Aohua Li**[1], **Jin Kuang**[2], **Yubing Lu**[3,4], **Pingping Liu**[3,4]
[1]College of Software, Jilin University, Changchun, Jilin 130000, China
[2]Hunan Engineering Research Center of Advanced Embedded Computing and Intelligent Medical Systems, Xiangnan University, Chenzhou, 423300, China
[3]College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China
[4]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin 130012, China
liah24@mails.jlu.edu.cn, gasquue@gmail.com, luyb24@mails.jlu.edu.cn, liupp@jlu.edu.cn

## Abstract

Infrared small target detection (IRSTD) has achieved substantial progress under domain-consistent evaluation, yet detector performance often degrades markedly when generalizing to unseen infrared domains. Existing methods primarily improve detection by enhancing target responses and suppressing background interference. However, when trained on only a limited set of source domains, their learned decision rules are inevitably established from a restricted range of source-domain target-background relation patterns. We formulate this cross-domain failure as target-background relation shift: unseen domains may exhibit relation patterns that are not observed during training, thereby weakening the discriminative capability learned from the source domains. To address this problem, we propose HyTBE, a Hyperbolic Target-Background Expert model that expands source-domain relation patterns and adaptively adjusts visual representations using explicit relation cues. The Target-Background Relation Intervention selectively perturbs either targets or backgrounds, broadening the observable relation patterns during training while maintaining valid supervision. Subsequently, the Hyperbolic Relation Modeling maps multi-scale visual cues into a Poincaré ball and characterizes the target-background relation of each feature token according to its relative distances to the target and background anchors. The Hyperbolic-guided MoE Adapter further uses these hyperbolic relation representations to calibrate multi-scale visual features and aggregate expert-specific feature corrections for different relation patterns. Leave-one-domain-out experiments on NUAA-SIRST, NUDT-SIRST, and IRSTD-1K demonstrate that HyTBE achieves stronger cross-domain generalization than competitive baselines.

**Code** — https://github.com/PepperCS/HyTBE

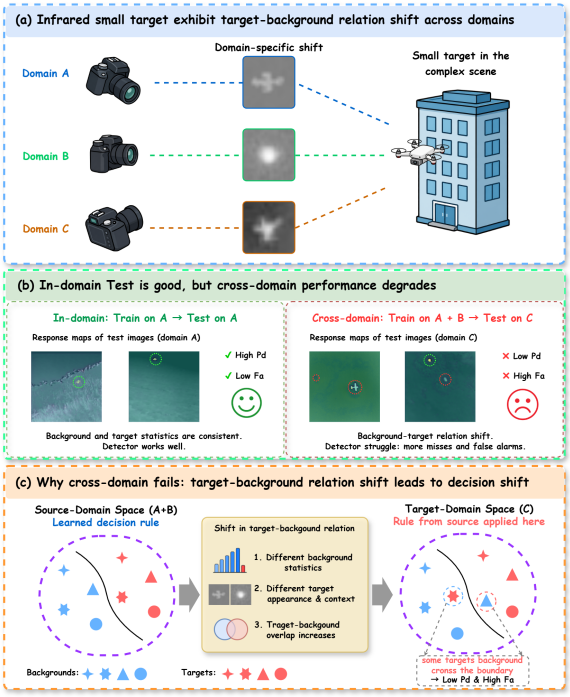


Figure 1: Motivation for cross-domain IRSTD: variations in target appearance and background statistics induce target-background relation shift, weakening source-domain decision rules in unseen domains.

## Introduction

Infrared small target detection (IRSTD) aims to locate tiny and dim targets embedded in complex infrared backgrounds (Kou et al. 2023), with important applications in remote sensing (Li et al. 2026a), maritime surveillance (Zhang et al. 2022), and early-warning systems (Lu et al. 2026b). Unlike generic objects (Huang et al. 2024) with recognizable textures, contours, and semantic structures, infrared small targets usually occupy only a few pixels and provide limited appearance information. Their weak responses can therefore be easily overwhelmed by background clutter and sensor noise. Recent deep-learning-based methods have substantially improved IRSTD performance through target-aware feature enhancement (Liu et al. 2024; Yang et al. 2025), multi-scale feature fusion (Wu, Hong, and Chanussot 2022; Li et al. 2022), background suppression (Liu et al. 2026a), model-driven unfolding (Wu et al. 2024), and Transformer-based representation learning (Liu et al. 2026b; Yuan et al. 2024). These advances have enabled increasingly accurate target localization under domain-consistent evaluation.

***Despite these advances, most existing methods have been developed and evaluated under domain-consistent settings, where the training and test data follow similar distributions.*** In practical deployment, detectors often encounter infrared domains that are not observed during training. Variations in sensors, imaging platforms, environmental conditions, and scene contents can jointly alter target appearances and background statistics (Yuan et al. 2025; Li et al. 2026b; Duan et al. 2026). As shown in Fig. 1(a), small targets from different infrared domains may appear as compact bright spots, weak blurred responses, or irregular structures, while their surrounding backgrounds exhibit distinct textures, noise levels, and clutter distributions. These observation differences further lead to a clear performance discrepancy. As shown in Fig. 1(b), a detector can produce concentrated target responses and effectively suppress background interference under domain-consistent evaluation, yet suffer from weakened responses, missed detections, and false alarms after being transferred to an unseen domain. This phenomenon indicates that a decision rule effective in the source domains does not necessarily generalize reliably to a new infrared domain.

***The central challenge of cross-domain IRSTD is that domain variations alter not only the appearances of targets and backgrounds individually, but also the discriminative relation between them.*** Because infrared small targets contain few pixels and lack stable texture and semantic structures, their detectability depends strongly on their relative contrast, saliency, morphology, and contextual difference from the surrounding background. Consequently, a target that is sufficiently distinctive against one background may become ambiguous against another. As conceptually illustrated in Fig. 1(c), a decision boundary learned from limited source-domain relation patterns may fail to reliably separate target and background features in an unseen domain. We refer to this variation in relation patterns and the resulting failure of the source-learned decision rule as target-background relation shift.

***A robust cross-domain detector should therefore both cover more relation patterns during training and adapt its features to different relations.*** On the one hand, the source-domain training data should expose the detector to sufficiently diverse target-background relations, thereby reducing its dependence on the limited patterns contained in the original datasets. On the other hand, the model should explicitly characterize whether visual features are more closely associated with targets or backgrounds and adjust their representations accordingly. The former broadens the range of relations observed during training, whereas the latter allows the model to respond adaptively to different relation patterns. Together, these two requirements provide a principled path from source-domain relation diversification to unseen-domain relation generalization.

Following this principle, we propose HyTBE, a Hyperbolic Target-Background Expert model for cross-domain IRSTD. HyTBE first introduces Target-Background Relation Intervention, which selectively perturbs either targets or backgrounds while retaining valid supervision. By varying one side of the relation at a time, this strategy broadens the target-background patterns observed during source-domain training without requiring access to the target domain. HyTBE then performs Hyperbolic Relation Modeling, which maps multi-scale visual cues into a Poincaré ball and characterizes each feature token through its relative distances to target and background anchors. The hyperbolic distance metric thereby provides explicit target-oriented and background-oriented relation representations. Finally, a Hyperbolic-guided MoE Adapter uses these relation representations to recalibrate multi-scale visual features and aggregate expert-specific feature corrections for different relation patterns. The three components form a progressive pipeline that diversifies target-background relations, explicitly represents them, and adapts visual features accordingly.

We evaluate HyTBE under a leave-one-domain-out protocol on three public IRSTD datasets: NUAA-SIRST, NUDT-SIRST, and IRSTD-1K. In each setting, two datasets are used as source domains for training, while the remaining dataset is held out as an unseen target domain without target-domain fine-tuning. Experimental results show that HyTBE achieves the best mIoU and F-measure across all three unseen target domains. Ablation studies further verify the individual contributions of Target-Background Relation Intervention, Hyperbolic Relation Modeling, and the Hyperbolic-guided MoE Adapter.

The main contributions of this work are summarized as follows:

- We formulate cross-domain IRSTD from the perspective of target-background relation shift, highlighting that decision rules established from limited source-domain relation patterns may become unreliable in unseen infrared domains.
- We introduce Target-Background Relation Intervention, which selectively perturbs targets or backgrounds to expand the relation patterns observed during source-domain training while retaining valid supervision.
- We develop HyTBE by combining Hyperbolic Relation Modeling with a Hyperbolic-guided MoE Adapter, allowing explicit hyperbolic target-background relation cues to guide adaptive multi-scale feature calibration for improved cross-domain generalization.

## Related Work

### IRSTD under Domain Shift

Early infrared small target detection (IRSTD) methods rely on hand-crafted priors, including local contrast (Bai and Zhou 2010; Chen et al. 2013), low-rank decomposition (Zhu et al. 2019), and background modeling (Gao et al. 2013). Deep networks have substantially improved in-domain performance through attention mechanisms (Dai et al. 2021b), multi-scale and U-shaped architectures (Wu, Hong, and Chanussot 2022; Li et al. 2022; Lu et al. 2026b; Zhang et al. 2022), Transformer modeling (Yuan et al. 2024), model-driven unfolding (Wu et al. 2024; Liu et al. 2026a), and query-guided representations (Liu et al. 2026b). However, these methods generally assume consistent training and test distributions, making them vulnerable to domain-specific target and background cues. Recent studies improve cross-domain robustness through semantic adaptation (Chi et al.

2024), perturbed training with test-time adaptation (Chen et al. 2024), or frequency-domain representations (Fu et al. 2026). Unlike these approaches, we address shifts in target-background discriminative relations through relation intervention and hyperbolic expert modeling.

### Hyperbolic Representation Learning

Hyperbolic representation learning exploits negatively curved manifolds to represent structured relations with low distortion. Foundational studies introduced Poincaré embeddings (Nickel and Kiela 2017) and extended neural operations to hyperbolic space (Ganea, Bécigneul, and Hofmann 2018). Subsequent works demonstrated its effectiveness in visual recognition, segmentation, contrastive learning, and domain generalization (Liu et al. 2020; Atigh et al. 2022; Ge et al. 2023; Bi et al. 2025). Recently, hyperbolic geometry has been applied to IRSTD. HyperISTD improves target-background separability through Poincaré embeddings and angular constraints (Lu et al. 2026a), while LoHGNet combines Lorentz encoding with high-order contextual modeling (Ma et al. 2026). These methods mainly employ hyperbolic geometry for feature enhancement or optimization under domain-consistent settings. In contrast, HyTBE characterizes each feature token through its relative hyperbolic distances to target and background anchors, and uses the resulting relation cues to guide MoE-based multi-scale feature adaptation for cross-domain IRSTD.

### Mixture-of-Experts Models

Mixture-of-Experts (MoE) employs input-dependent routing to combine specialized expert networks (Jacobs et al. 1991). Sparse MoE enables conditional computation (Shazeer et al. 2017), with later studies improving routing stability and load balancing (Fedus, Zoph, and Shazeer 2022). Its effectiveness has also been demonstrated in visual recognition (Riquelme et al. 2021), multimodal learning (Mustafa et al. 2022), and differentiable soft expert aggregation (Puigcerver et al. 2024).

MoE has recently been introduced into IRSTD to capture domain-dependent patterns (Duan et al. 2026). Unlike domain-level routing, HyTBE uses hyperbolic target-background relation cues to guide multi-scale MoE adapters, enabling relation-aware feature adaptation under domain shifts.

## Method

### Problem Formulation

**Domain Generalization Setting.** We study IRSTD under the leave-one-domain-out setting. Given multiple infrared domains $\{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_K\}$, the model is trained on source domains $\mathcal{D}_s = \{\mathcal{D}_k\}_{k \neq t}$ and directly evaluated on an unseen target domain $\mathcal{D}_t$. During training, images and annotations from $\mathcal{D}_t$ are not accessible. The detector predicts a response map $\hat{y} = f_\theta(x)$ for each infrared image $x$, where high-response regions indicate potential small targets.

**Target-Background Relation Shift.** Different from general object detection, IRSTD relies heavily on the discriminative relation between tiny targets and their surrounding backgrounds. Since infrared small targets occupy only a few pixels and lack stable texture or semantic structures, their detectability is determined not only by target appearance, but also by local saliency, target morphology, background clutter, and imaging style. We denote such target-background relation as

$$r = \Phi(x_t, x_b),$$

where $x_t$ and $x_b$ represent the target candidate and its surrounding background context, respectively. A model trained on source domains learns a decision rule from the source relation distribution $P_s(r, y)$. However, in an unseen target domain, target appearance and background context may change jointly, leading to

$$P_s(r, y) \neq P_t(r, y).$$

As a result, the source-domain decision rule may become unreliable, causing missed detections and false alarms.

**Therefore, the key to cross-domain IRSTD lies in learning a robust target-background discrimination rule that remains reliable under relation shifts.** This motivates us to move beyond source-domain target enhancement and explicitly consider how target-background relations vary and can be adapted across infrared domains.

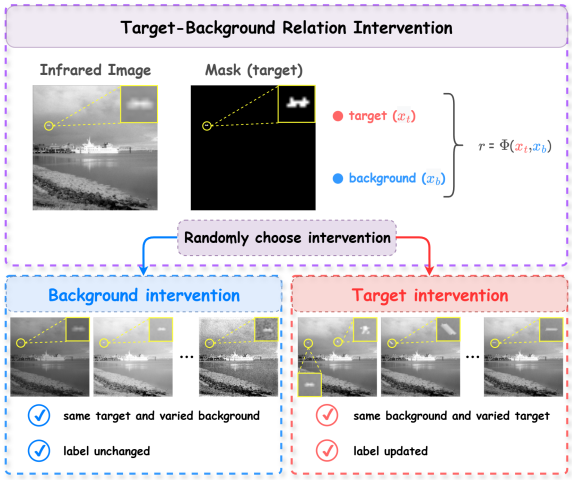


Figure 2: Target-Background Relation Intervention (TBRI) selectively perturbs targets or backgrounds to diversify relation patterns while retaining valid supervision.

### Target-Background Relation Intervention

To improve robustness to unseen domains, we introduce Target-Background Relation Intervention (TBRI) during source-domain training. Unlike conventional image-level augmentation, TBRI selectively modifies either the target or its background context while maintaining consistent supervision. As illustrated in Fig. 2, TBRI consists of target intervention and background intervention, from which only one operator is selected for each intervened sample.

Given an infrared image $x$ and its binary mask $y$, the target and background masks are defined as

$$M_t = y, \; M_b = 1 - y.$$

For each training sample, TBRI is activated with probability $p = 0.5$. Once activated, exactly one intervention operator is sampled:

$$k \sim \text{Categorical}(\boldsymbol{\pi}),\ \pi_k = \frac{1}{K},\ \sum_{k=1}^{K} \pi_k = 1,$$

$$(\tilde{x}, \tilde{y}) = \mathcal{T}_k(x, y).$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ denotes the uniform sampling probabilities, and $\mathcal{T}_k$ is the selected target or background intervention operator. Here, $\tilde{x}$ and $\tilde{y}$ denote the intervened image and its corresponding annotation, respectively.

**Background intervention.** Background intervention diversifies the imaging context while preserving the target support. We consider variations in global intensity response, background contrast, high-frequency components, and noise patterns. Let $\mathcal{B}_k(\cdot)$ denote the selected background intervention operator and $x_b^k$ its generated background candidate:

$$x_b^k = \mathcal{B}_k(x).$$

The intervened sample is constructed as

$$\tilde{x} = M_t \odot x + M_b \odot x_b^k,\ \tilde{y} = y.$$

Therefore, the target region and its annotation are retained, whereas the surrounding background is replaced by the perturbed candidate. Global imaging-style operators adjust the response of the entire image consistently without changing its binary annotation.

**Target intervention.** Target intervention preserves the surrounding background while jointly transforming target appearance and target support. We consider variations in target saliency, brightness, morphology, and scale, together with random target sampling from the source domains. Let $\mathcal{A}_k(\cdot)$ denote the selected target intervention operator. It generates the transformed target response $x_t^k$ and its corresponding binary mask $M_t^k$:

$$(x_t^k, M_t^k) = \mathcal{A}_k(x_t, M_t).$$

The intervened sample is constructed as

$$\tilde{x} = (1 - M_t^k) \odot x + M_t^k \odot x_t^k,\ \tilde{y} = M_t^k.$$

Therefore, the surrounding background is preserved, whereas the target response and its corresponding annotation are jointly updated according to the selected intervention. For random target sampling, an additional source-domain target is transformed and inserted into a valid background location, with its support incorporated into the updated binary annotation.

## Overview of HyTBE

The overall architecture of HyTBE is illustrated in Fig. 3. Given an infrared image $x$, an encoder composed of ResBlocks (He et al. 2016) extracts multi-scale visual features:

$$\{F_i\}_{i=1}^{I} = \mathcal{E}(x),$$

where $\mathcal{E}(\cdot)$ denotes the visual encoder, $I$ is the number of feature levels, and $F_i$ represents the visual feature at the $i$-th level.

Hyperbolic Relation Modeling (HRM) then maps the multi-scale features into a Poincaré ball and characterizes their target-background relations according to their relative distances to the target and background anchors:

$$\{z_i\}_{i=1}^{I} = \left\{\mathcal{F}_{\text{HRM}}^{i}(F_i)\right\}_{i=1}^{I},$$

where $\mathcal{F}_{\text{HRM}}^{i}(\cdot)$ denotes HRM at the $i$-th level, and $z_i$ is the resulting hyperbolic relation representation.

The Hyperbolic-guided MoE Adapter (HMA) subsequently uses $z_i$ to guide feature fusion and expert-based feature correction:

$$\{A_i\}_{i=1}^{I} = \left\{\mathcal{F}_{\text{HMA}}^{i}(F_i, z_i)\right\}_{i=1}^{I},$$

where $\mathcal{F}_{\text{HMA}}^{i}(\cdot)$ denotes HMA at the $i$-th level, and $A_i$ represents the corresponding relation-adapted feature.

Finally, a decoder composed of RPCABlocks (Wu et al. 2024; Liu et al. 2026a) progressively fuses the adapted multi-scale features and predicts the target response map:

$$\hat{y} = \mathcal{D}\left(\{A_i\}_{i=1}^{I}\right),$$

where $\mathcal{D}(\cdot)$ denotes the decoder and $\hat{y}$ is the predicted target response map. Overall, HyTBE follows a relation-modeling and relation-guided adaptation pipeline: HRM explicitly characterizes the target-background relation of visual features, while HMA uses these relation cues to adapt feature responses to diverse target-background patterns.

**Hyperbolic Relation Modeling** Given multi-scale encoder features $\{F_i\}_{i=1}^{I}$, we employ Patch Embedding (PE) to project features of different resolutions onto a common spatial grid and construct a target-background relation representation:

$$u_i = \mathcal{PE}\left(F_i\right)$$

where $\mathcal{PE}(\cdot)$ denotes PE and $u_i \in \mathbb{R}^D$ is the $i$-th relation token. PE combines learned patch features with max pooled and average pooled responses to capture target saliency and background context.

Following (Ganea, Bécigneul, and Hofmann 2018), we map each relation token from the tangent space at the origin onto the Poincaré ball using the exponential map, where (c) is fixed to (1) throughout all experiments:

$$z_i = \text{Exp}_0^c(u_i).$$

To establish target-background reference directions, we construct two opposite anchors:

$$a_t = \text{Exp}_0^c(\rho),\ a_b = \text{Exp}_0^c(-\rho),$$

where $a_t$ and $a_b$ denote the target and background anchors, respectively, and $\rho$ controls the anchor scale. The relation score of each token is defined as

$$s_i = d_c(z_i, a_b) - d_c(z_i, a_t),$$

where $d_c(\cdot, \cdot)$ denotes the geodesic distance in the Poincaré ball (Ganea, Bécigneul, and Hofmann 2018). Thus, $s_i >$

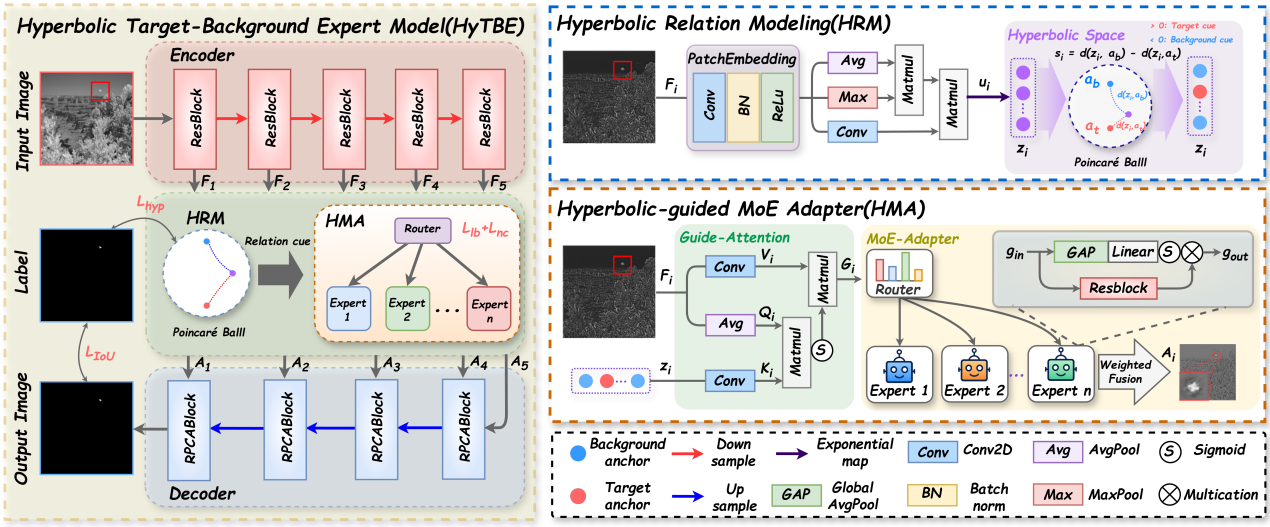


Figure 3: Overall architecture of the Hyperbolic Target-Background Expert model (HyTBE). Its two main components are Hyperbolic Relation Modeling (HRM) for target-background relation representation and the Hyperbolic-guided MoE Adapter (HMA) for relation-guided multi-scale feature adaptation.

0 indicates a target relation, whereas $s_i < 0$ indicates a background relation.

The ground-truth mask is resized to the token resolution using adaptive max pooling, yielding the target and background token sets $\Omega_t$ and $\Omega_b$. We define the target and background relation losses as

$$\mathcal{L}_t = \frac{1}{|\Omega_t|} \sum_{i \in \Omega_t} [m - s_i]_+$$

$$\mathcal{L}_b = \frac{1}{|\Omega_b|} \sum_{i \in \Omega_b} [m + s_i]_+$$

where $m$ is the relation margin and $[v]_+ = \max(0, v)$. The complete hyperbolic relation loss is

$$\mathcal{L}_{\text{hyp}} = \mathcal{L}_t + \mathcal{L}_b.$$

This loss encourages target tokens to satisfy $s_i \geq m$ and background tokens to satisfy $s_i \leq -m$, while preserving the diversity of relation patterns within each side. The resulting hyperbolically constrained representation $\{z_i\}_{i=1}^N$ is subsequently used as the shared relation cue for HMA.

**Hyperbolic-guided MoE Adapter** HMA consists of Guide-Attention and a MoE adapter. For the $i$-th visual feature $F_i$, its avg pooled representation provides the query, the HRM relation representation $z_i$ provides the key, and the visual feature itself provides the value:

$$Q_i = P(F_i), K_i = \phi_k(z_i), V_i = \phi_v(F_i),$$

where $P(\cdot)$ denotes average pooling and $\phi_k(\cdot)$ and $\phi_v(\cdot)$ are convolutional projections. The relation-guided feature is computed as

$$G_i = \text{Sigmoid}\left(\frac{\bar{Q}_i \bar{K}_i^\top}{\sqrt{C}}\right) V_i,$$

where $\bar{Q}_i$ and $\bar{K}_i$ are normalized features. Thus, HRM relation cues recalibrate the channel responses of each encoder feature.

The guided feature is subsequently fed to a scale-specific MoE adapter. Each expert applies a residual block and modulates its response with a lightweight channel gate:

$$\mathcal{E}_{i,e}(G_i) = \mathcal{B}_{i,e}(G_i) \otimes \text{Sigmoid}\left(W_{i,e}\,\text{GAP}(G_i)\right),$$

where $\mathcal{B}_{i,e}(\cdot)$ is the residual block of the $e$-th expert, $\text{GAP}(\cdot)$ denotes global average pooling, $W_{i,e}$ is a linear projection, and $\otimes$ multiplication.

In parallel, the router predicts input-dependent soft weights over the $E$ experts:

$$g_i = \text{Softmax}\left(\text{MLP}(\text{GAP}(G_i))\right).$$

The expert corrections are then aggregated by weighted fusion and added back to the guided feature:

$$A_i = G_i + \alpha_i \sum_{e=1}^{E} g_{i,e} \mathcal{E}_{i,e}(G_i),$$

where $g_{i,e}$ is the routing weight of the $e$-th expert and $\alpha_i$ is a learnable residual scale for the $i$-th feature level.

We employ a load-balancing loss ($\mathcal{L}_{\text{lb}}$) (Fedus, Zoph, and Shazeer 2022) to encourage balanced expert usage and a non-consistency loss ($\mathcal{L}_{\text{nc}}$) (Dai et al. 2021a) to promote diverse expert corrections.

**Training Loss** The model is jointly optimized by $\mathcal{L} = \mathcal{L}_{\text{IoU}} + \mathcal{L}_{\text{hyp}} + \mathcal{L}_{\text{lb}} + \mathcal{L}_{\text{nc}}$, where the four terms supervise target segmentation, hyperbolic relation modeling, balanced expert routing, and expert diversity, respectively.

# Experiments

## Experimental Settings

**Datasets.** We conduct experiments on three widely used public infrared small target detection datasets, including

Table 1: Cross-domain comparisons with SOTA methods on NUAA-SIRST, NUDT-SIRST, and IRSTD-1K in $mIoU$ (%), $F$-measure (%), $P_d$ (%), $F_a$ ($10^{-6}$), parameters (M), and FLOPs (G).

| Method | Source: NUAA-SIRST, NUDT-SIRST | | | | Source: NUDT-SIRST, IRSTD-1K | | | | Source: NUAA-SIRST, IRSTD-1K | | | | Params (M) | FLOPs (G) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Target: IRSTD-1K | | | | Target: NUAA-SIRST | | | | Target: NUDT-SIRST | | | | | |
| | $mIoU$ | $F$ | $P_d$ | $F_a$ | $mIoU$ | $F$ | $P_d$ | $F_a$ | $mIoU$ | $F$ | $P_d$ | $F_a$ | | |
| ALCNet (TGRS 2021) | 45.40 | 62.45 | 81.09 | 90.48 | 68.05 | 80.99 | 97.22 | 16.51 | 49.66 | 66.36 | 84.23 | 78.68 | **0.42** | **0.37** |
| DNANet (TIP 2021) | 49.67 | 66.37 | 79.38 | **32.79** | 68.89 | 81.58 | 94.44 | 24.59 | 51.54 | 68.02 | 81.79 | 98.88 | 4.69 | 14.26 |
| UIUNet (TIP 2022) | 51.83 | 68.27 | 83.50 | 100.89 | 72.83 | 84.28 | 95.37 | 14.54 | 55.28 | 71.20 | 83.80 | 43.06 | 50.54 | 54.42 |
| MSHNet (CVPR 2024) | 49.59 | 66.30 | 84.53 | 56.17 | 70.63 | 82.79 | 96.29 | 29.08 | 54.62 | 70.65 | 84.02 | 68.25 | 4.06 | 6.10 |
| SCTransNet (TGRS 2024) | 48.26 | 65.10 | 84.53 | 75.07 | 70.20 | 82.49 | 96.29 | 15.61 | 59.38 | 74.51 | 85.71 | **37.84** | 11.19 | 10.11 |
| DRPCANet (TGRS 2025) | 25.82 | 41.05 | 84.87 | 368.79 | 70.46 | 82.67 | 94.44 | **7.89** | 44.64 | 61.72 | 77.67 | 158.93 | 1.16 | 73.83 |
| PConv (AAAI 2025) | 48.14 | 64.99 | **92.09** | 126.16 | 69.28 | 81.85 | 96.29 | 14.72 | 56.75 | 72.41 | 84.65 | 53.91 | 2.93 | 5.24 |
| MLPNet (TGRS 2025) | 31.93 | 48.41 | 81.78 | 284.67 | 69.86 | 82.26 | 95.37 | 23.51 | 46.83 | 63.78 | 84.44 | 205.18 | 8.26 | 7.01 |
| PQGNet (TGRS 2026) | 50.12 | 66.77 | 87.62 | 99.97 | 73.66 | 84.83 | 98.69 | 58.70 | 57.66 | 73.15 | 84.12 | 55.61 | 1.20 | 9.89 |
| NS_FPN (CVPR 2026) | 51.18 | 67.71 | 88.65 | 64.07 | 65.51 | 79.16 | 94.44 | 29.62 | 49.37 | 66.11 | 82.22 | 81.41 | 4.16 | 7.96 |
| HyTBE (Ours) | **52.31** | **68.69** | 86.59 | 96.03 | **78.58** | **88.00** | **99.56** | 13.10 | **64.83** | **78.66** | **87.08** | 52.02 | 1.13 | 8.42 |

Table 2: Ablation study on the NUDT-SIRST + IRSTD-1K $\rightarrow$ NUAA-SIRST setting.

| TBRI | HRM | HMA | $mIoU \uparrow$ | $F \uparrow$ | $P_d \uparrow$ | $F_a \downarrow$ |
|---|---|---|---|---|---|---|
| | | | 68.93 | 81.60 | 95.37 | 14.89 |
| ✓ | | | 71.86 | 83.63 | 97.22 | 15.61 |
| ✓ | ✓ | | 74.48 | 85.37 | 98.25 | **12.92** |
| ✓ | ✓ | ✓ | **78.58** | **88.00** | **99.56** | 13.10 |

Table 3: Ablation study of target and background interventions in TBRI on the NUDT-SIRST + IRSTD-1K $\rightarrow$ NUAA-SIRST setting.

| Target | Background | $mIoU \uparrow$ | $F \uparrow$ | $P_d \uparrow$ | $F_a \downarrow$ |
|---|---|---|---|---|---|
| | | 71.03 | 83.06 | 97.22 | 31.05 |
| ✓ | | 73.22 | 84.54 | 96.29 | 13.46 |
| | ✓ | 71.92 | 83.67 | 98.14 | 21.36 |
| ✓ | ✓ | **78.58** | **88.00** | **99.56** | **13.10** |

NUAA-SIRST (Dai et al. 2021b), NUDT-SIRST (Li et al. 2022), and IRSTD-1K (Zhang et al. 2022), which contain 427, 1327, and 1000 images, respectively. These datasets are collected from different infrared imaging scenarios and exhibit variations in target scale, target morphology, background content, and imaging style, making them suitable for evaluating cross-dataset generalization. For each dataset, we adopt the standard training-test split configuration used in (Yuan et al. 2024). Following common IRSTD practice, pixel-level annotations are used for training and test.

**Cross-domain protocol.** To evaluate cross-domain robustness, we adopt a leave-one-domain-out protocol. Specifically, two datasets are used as source domains for training, while the remaining dataset is held out as an unseen target domain for testing. This yields three cross-domain settings: NUAA-SIRST + NUDT-SIRST $\rightarrow$ IRSTD-1K, NUAA-SIRST + IRSTD-1K $\rightarrow$ NUDT-SIRST, and NUDT-SIRST + IRSTD-1K $\rightarrow$ NUAA-SIRST.

**Evaluation metrics.** Following common evaluation protocols in IRSTD, we report four widely used metrics: mean Intersection over Union ($mIoU$), F-measure ($F$), probability of detection ($P_d$), and false alarm rate ($F_a$). Higher mIoU, F-measure, and $P_d$ indicate better detection performance, while lower $F_a$ indicates fewer false alarms.

Table 4: Comparison between Euclidean and hyperbolic relation modeling on the NUDT-SIRST + IRSTD-1K $\rightarrow$ NUAA-SIRST setting. All other components and training settings are identical.

| Geometry | Relation metric | $mIoU \uparrow$ | $F \uparrow$ | $P_d \uparrow$ | $F_a \downarrow$ |
|---|---|---|---|---|---|
| Euclidean | $\ell_2$ distance | 78.08 | 87.69 | 99.03 | 17.77 |
| Hyperbolic | Poincaré distance | **78.58** | **88.00** | **99.56** | **13.10** |

Table 5: Ablation study of the HMA module on the NUDT-SIRST + IRSTD-1K $\rightarrow$ NUAA-SIRST setting.

| Guide-Attn | MoE | $mIoU \uparrow$ | $F \uparrow$ | $P_d \uparrow$ | $F_a \downarrow$ |
|---|---|---|---|---|---|
| | | 74.48 | 85.37 | 98.25 | 12.92 |
| ✓ | | 75.83 | 86.25 | 99.13 | 21.90 |
| | ✓ | 76.15 | 86.46 | 98.12 | **6.82** |
| ✓ | ✓ | **78.58** | **88.00** | **99.56** | 13.10 |

**Implementation details.** All experiments are implemented with PyTorch on an NVIDIA GeForce RTX 4070 Ti SUPER GPU. The proposed HyTBE is trained from scratch without using any pretrained weights. For each input image, we first normalize it and then resize it to $256 \times 256$. Random flipping and rotation are adopted for data augmentation. We train the model for 200 epochs with a batch size of 4 using the Adam optimizer. The initial learning rate is set to 0.001.

## Comparison with State-of-the-Art Methods

As shown in Table 1, we compare HyTBE with ten representative methods, including ALCNet (Dai et al. 2021c), DNANet (Li et al. 2022), UIUNet (Wu, Hong, and Chanussot 2022), MSHNet (Liu et al. 2024), SCTransNet (Yuan et al. 2024), DRPCANet (Xiong et al. 2025), PConv (Yang et al. 2025), MLPNet (Wang et al. 2025), PQGNet (Liu et al. 2026b), and NS_FPN (Yuan et al. 2026). HyTBE achieves the best mIoU and $F$-measure across all three cross-domain set-

Table 6: Ablation study of the auxiliary objectives. $\mathcal{L}_{\mathrm{IoU}}$ is used in all variants.

| $\mathcal{L}_{\mathrm{hyp}}$ | $\mathcal{L}_{\mathrm{lb}}$ | $\mathcal{L}_{\mathrm{nc}}$ | $mIoU\uparrow$ | $F\uparrow$ | $P_d\uparrow$ | $F_a\downarrow$ |
|---|---|---|---|---|---|---|
| | | | 75.47 | 86.02 | 98.07 | 26.20 |
| ✓ | | | 77.18 | 87.12 | <u>99.12</u> | **9.51** |
| | ✓ | | 74.19 | 85.18 | 98.24 | 17.77 |
| | | ✓ | 75.31 | 85.92 | 97.22 | 15.56 |
| ✓ | ✓ | | 76.88 | 86.93 | 98.15 | <u>11.43</u> |
| ✓ | | ✓ | <u>77.72</u> | <u>87.46</u> | 98.84 | 11.61 |
| | ✓ | ✓ | 76.18 | 86.48 | 98.16 | 17.53 |
| ✓ | ✓ | ✓ | **78.58** | **88.00** | **99.56** | 13.10 |

tings, outperforming the corresponding second-best methods by 0.48/0.42, 4.92/3.17, and 5.45/4.15 percentage points on IRSTD-1K, NUAA-SIRST, and NUDT-SIRST, respectively. It also obtains the highest $P_d$ on NUAA-SIRST and NUDT-SIRST. Although HyTBE does not achieve the lowest $F_a$ in every setting, its consistent improvements in mIoU and F-measure demonstrate a better overall balance between target preservation and false-alarm suppression. Moreover, HyTBE contains only 1.13M parameters with 8.42G FLOPs, demonstrating strong cross-domain performance with a compact parameter footprint and moderate computational cost.

## Ablation Studies

All ablation experiments are conducted under the NUDT-SIRST + IRSTD-1K $\rightarrow$ NUAA-SIRST setting.

**Effectiveness of the main components.** As shown in Table 2, the baseline achieves 68.93% mIoU. Introducing TBRI improves mIoU and $F$-measure by 2.93 and 2.03 percentage points, respectively. Adding HRM further increases mIoU to 74.48%, while the complete model with HMA reaches 78.58% mIoU, 88.00% $F$-measure, and 99.56% $P_d$. These progressive improvements verify the complementary contributions of relation diversification, relation modeling, and relation-guided feature adaptation.

**Effect of target-background interventions.** Table 3 separately evaluates the two intervention branches while keeping the remaining architecture fixed. Target intervention provides a larger mIoU improvement and substantially reduces $F_a$, whereas background intervention improves $P_d$ by exposing the detector to diverse background contexts. Combining both branches achieves the best overall performance, improving mIoU from 71.03% to 78.58%. This result indicates that target-side and background-side variations provide complementary relation patterns.

**Geometry and HMA design.** As reported in Table 4, hyperbolic modeling consistently outperforms its Euclidean counterpart, particularly reducing $F_a$ from 17.77 to 13.10. Table 5 further shows that both guide-attention and MoE independently improve mIoU and $F$-measure. Their combination achieves the best mIoU, $F$-measure, and $P_d$, confirming that relation-guided fusion and expert adaptation work complementarily. Although MoE alone produces the lowest $F_a$, the complete HMA provides a better overall balance across the four evaluation metrics.

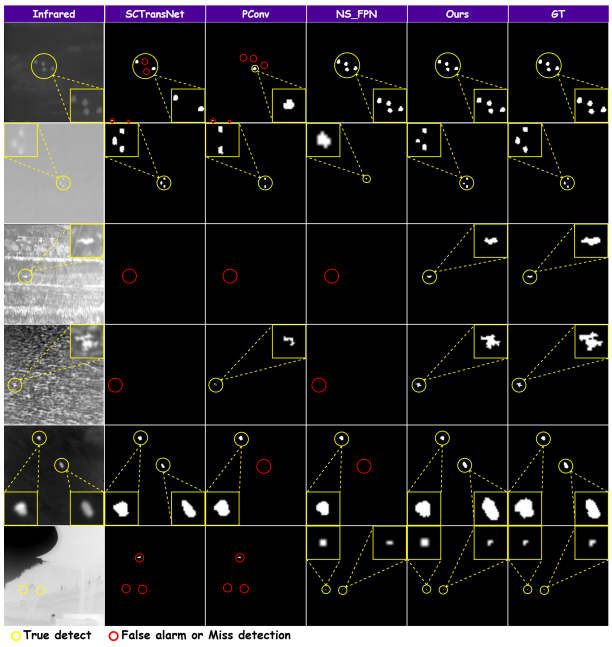


Figure 4: Qualitative cross-domain comparison with representative methods. From top to bottom, every two rows show results on the unseen target domains NUAA-SIRST, IRSTD-1K, and NUDT-SIRST, respectively.

**Effect of auxiliary loss.** As shown in Table 6, using $\mathcal{L}_{\mathrm{hyp}}$ alone $mIoU$ and $F$-measure by 1.71 and 1.10 percentage points, respectively, while reducing $F_a$ from 26.20 to 9.51. When used independently, $\mathcal{L}_{\mathrm{lb}}$ and $\mathcal{L}_{\mathrm{nc}}$ mainly reduce false alarms but do not improve the overlap-based metrics. Combining $\mathcal{L}_{\mathrm{nc}}$ with relation supervision further increases mIoU to 77.72%. Using all three objectives achieves the best $mIoU$, $F$-measure, and $P_d$ of 78.58%, 88.00%, and 99.56%, respectively, demonstrating their overall effectiveness in jointly constraining relation representation and expert adaptation.

## Visualization

Fig. 4 presents qualitative comparisons on three unseen target domains. Existing methods often miss dim targets or produce inaccurate masks under low contrast, background clutter, and scale variations. In contrast, HyTBE better preserves the target number, location, and shape, producing predictions closer to the ground truth and demonstrating stronger cross-domain generalization.

## Conclusion

This paper proposed a Hyperbolic Target-Background Expert model (HyTBE) to improve the cross-domain generalization of IRSTD. Considering that unseen infrared domains may exhibit target-background relation patterns beyond those observed during training, we introduced Target-Background Relation Intervention (TBRI) to diversify source-domain relations by selectively perturbing targets or backgrounds. Moreover, we developed Hyperbolic Relation Modeling

(HRM) to explicitly characterize target-background relations in hyperbolic space, together with a Hyperbolic-guided MoE Adapter (HMA) that uses these relation cues to adapt multi-scale visual features. These three components are closely integrated to expand, represent, and adapt target-background relations for more transferable discrimination. Experiments across three unseen target domains demonstrate the superiority of HyTBE, highlighting its potential for robust and practical IRSTD.

## References


Atigh, M. G.; Schoep, J.; Acar, E.; Van Noord, N.; and Mettes, P. 2022. Hyperbolic image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4453–4462.

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

Bi, Q.; Yi, J.; Zhan, H.; Ji, W.; and Xia, G.-S. 2025. Learning fine-grained domain generalization via hyperbolic state space hallucination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1853–1861.

Chen, C. P.; Li, H.; Wei, Y.; Xia, T.; and Tang, Y. Y. 2013. A local contrast method for small infrared target detection. *IEEE transactions on geoscience and remote sensing*, 52(1): 574–581.

Chen, G.; Wang, W.; Wang, Z.; Li, X.; and Wu, H. 2024. Enhanced generalization ability of infrared small target detection via perturbed training and adaptive test. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–17.

Chi, W.; Liu, J.; Wang, X.; Ni, Y.; and Feng, R. 2024. A semantic domain adaption framework for cross-domain infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–13.

Dai, Y.; Li, X.; Liu, J.; Tong, Z.; and Duan, L.-Y. 2021a. Generalizable person re-identification with relevance-aware mixture of experts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16145–16154.

Dai, Y.; Wu, Y.; Zhou, F.; and Barnard, K. 2021b. Asymmetric contextual modulation for infrared small target detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 950–959.

Dai, Y.; Wu, Y.; Zhou, F.; and Barnard, K. 2021c. Attentional local contrast networks for infrared small target detection. *IEEE transactions on geoscience and remote sensing*, 59(11): 9813–9824.

Duan, W.; Ji, L.; Huang, J.; Zhu, S.; and Ye, M. 2026. Cross-domain Joint Learning with Prototype-guided Mixture-of-Experts for Infrared Moving Small Target Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 3732–3740.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120): 1–39.

Fu, Y.; Wang, S.; Wu, F.; Lyu, J.; Liu, Z.; and Ng, M. K. 2026. Rethinking Representations for Cross-Domain Infrared Small Target Detection: A Generalizable Perspective from the Frequency Domain. *arXiv preprint arXiv:2604.01934*.

Ganea, O.; Bécigneul, G.; and Hofmann, T. 2018. Hyperbolic neural networks. *Advances in neural information processing systems*, 31.

Gao, C.; Meng, D.; Yang, Y.; Wang, Y.; Zhou, X.; and Hauptmann, A. G. 2013. Infrared patch-image model for small target detection in a single image. *IEEE transactions on image processing*, 22(12): 4996–5009.

Ge, S.; Mishra, S.; Kornblith, S.; Li, C.-L.; and Jacobs, D. 2023. Hyperbolic contrastive learning for visual representations beyond objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6840–6849.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, Y.-X.; Liu, H.-I.; Shuai, H.-H.; and Cheng, W.-H. 2024. Dq-detr: Detr with dynamic query for tiny object detection. In *European Conference on Computer Vision*, 290–305. Springer.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Kou, R.; Wang, C.; Peng, Z.; Zhao, Z.; Chen, Y.; Han, J.; Huang, F.; Yu, Y.; and Fu, Q. 2023. Infrared small target segmentation networks: A survey. *Pattern recognition*, 143: 109788.

Li, B.; Xiao, C.; Wang, L.; Wang, Y.; Lin, Z.; Li, M.; An, W.; and Guo, Y. 2022. Dense nested attention network for infrared small target detection. *IEEE Transactions on Image Processing*, 32: 1745–1758.

Li, R.; An, W.; Wang, Y.; Ying, X.; Dai, Y.; Wang, L.; Li, M.; Guo, Y.; and Liu, L. 2026a. Probing deep into temporal profile makes the infrared small target detector much better. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, Y.; Lu, Y.; Wu, H.; Zhang, S.; Lin, L.; and Shi, Y. 2026b. Ivan-ISTD: Rethinking cross-domain heteroscedastic noise perturbations in infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*.

Liu, P.; Li, A.; Lu, Y.; Kuang, J.; Zhang, T.; and Zhou, Q. 2026a. RPCASSM: Robust PCA State Space Model For Infrared Small Target Detection. *arXiv preprint arXiv:2606.01689*.

Liu, P.; Li, A.; Lu, Y.; Zhang, T.; Yang, M.; and Zhou, Q. 2026b. PQGNet: Perceptual Query Guided Network for Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*.

Liu, Q.; Liu, R.; Zheng, B.; Wang, H.; and Fu, Y. 2024. Infrared small target detection with scale and location sensitivity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17490–17499.

Liu, S.; Chen, J.; Pan, L.; Ngo, C.-W.; Chua, T.-S.; and Jiang, Y.-G. 2020. Hyperbolic visual embedding learning for zero-shot recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9273–9281.

Lu, Y.; Liu, P.; Zhang, T.; Li, A.; and Zhou, Q. 2026a. HyperISTD: Angular-Consistent Modeling in Hyperbolic Space for Infrared Small Target Detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

Lu, Y.; Liu, P.; Zhang, T.; Li, A.; and Zhou, Q. 2026b. Physics-Driven Feature Decoupling for Infrared Small Targets: A Dual Geometry-Guided Experts Network. *Knowledge-Based Systems*, 115268.

Ma, Q.; Xu, Y.; Deng, S.; Li, X.; and Hu, H. 2026. LoHGNet: Infrared Small Target Detection through Lorentz Geometric Encoding with High-Order Relation Learning. *arXiv preprint arXiv:2605.07213*.

Mustafa, B.; Riquelme, C.; Puigcerver, J.; Jenatton, R.; and Houlsby, N. 2022. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35: 9564–9576.

Nickel, M.; and Kiela, D. 2017. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30.

Puigcerver, J.; Riquelme Ruiz, C.; Mustafa, B.; and Houlsby, N. 2024. From sparse to soft mixtures of experts. In *International Conference on Learning Representations*, volume 2024, 28435–28445.

Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; and Houlsby, N. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34: 8583–8595.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Wang, Z.; Wang, C.; Li, X.; Xia, C.; and Xu, J. 2025. MLP-Net: Multi-Layer Perceptron Fusion Network for Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*.

Wu, F.; Zhang, T.; Li, L.; Huang, Y.; and Peng, Z. 2024. RPCANet: Deep unfolding RPCA based infrared small target detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4809–4818.

Wu, X.; Hong, D.; and Chanussot, J. 2022. UIU-Net: U-Net in U-Net for infrared small object detection. *IEEE Transactions on Image Processing*, 32: 364–376.

Xiong, Z.; Zhou, F.; Wu, F.; Yuan, S.; Fu, M.; Peng, Z.; Yang, J.; and Dai, Y. 2025. DRPCA-Net: Make robust PCA great again for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*.

Yang, J.; Liu, S.; Wu, J.; Su, X.; Hai, N.; and Huang, X. 2025. Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9202–9210.

Yuan, M.; Meng, D.; Xi, Z.; Zhao, T.; Zhao, S.; Dai, Y.; and Wei, X. 2026. Seeing Through the Noise: Improving Infrared Small Target Detection and Segmentation from Noise Suppression Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27783–27792.

Yuan, S.; Qin, H.; Yan, X.; Akhtar, N.; and Mian, A. 2024. SCTransNet: Spatial-channel cross transformer network for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–15.

Yuan, S.; Qin, H.; Yan, X.; Yang, S.; Yang, S.; Akhtar, N.; and Zhou, H. 2025. ASCNet: Asymmetric sampling correction network for infrared image destriping. *IEEE Transactions on Geoscience and Remote Sensing*.

Zhang, M.; Zhang, R.; Yang, Y.; Bai, H.; Zhang, J.; and Guo, J. 2022. ISNet: Shape matters for infrared small target detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 877–886.

Zhu, H.; Liu, S.; Deng, L.; Li, Y.; and Xiao, F. 2019. Infrared small target detection via low-rank tensor completion with top-hat regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 58(2): 1004–1016.

SUPPLEMENTARY MATERIAL

# HyTBE: Hyperbolic Target-Background Expert Model for Cross-Domain Infrared Small Target Detection

Aohua Li, Jin Kuang, Yubing Lu, and Pingping Liu

## Supplementary Material Overview

The supplementary material is organized as follows.

| Part | Content |
|---|---|
| A | Evidence of Target-Background Relation Shift. |
| B | Additional details of Target-Background Relation Intervention (TBRI). |
| C | Implementation Details and Hyperparameter Analysis. |

## A Evidence of Target-Background Relation Shift

This section provides empirical evidence for the target-background relation shift formulated in the main paper. Given a target candidate $x_t$ and its surrounding background context $x_b$, their relation is expressed as

$$r = \Phi(x_t, x_b). \tag{1}$$

A detector trained on the source domains establishes its decision rule from the source relation distribution $P_s(r, y)$. When the relation patterns in an unseen domain are not sufficiently represented by the source domains, the corresponding distribution changes:

$$P_s(r, y) \neq P_t(r, y). \tag{2}$$

Following this formulation, we organize our analysis around two questions.

**Q1: Does the target-background relation shift across domains?**

We assess its existence and magnitude through metric-wise distribution comparisons, joint PCA visualization, and standardized Wasserstein distances.

**Q2: Is relation deviation associated with source-model degradation?**

We evaluate a fixed source-trained baseline and examine whether target instances farther from the source-domain relation patterns exhibit higher target-pixel miss rates.

The following subsections address these two questions in turn.

### A.1 Analysis Protocol

All datasets adopt the publicly available splits from their original works, with no re-partitioning applied.

**Target and Local Background Extraction.** For a consistent measurement scale, we convert all images to grayscale, scale their intensities to $[0, 1]$, and resize them to $256 \times 256$; masks are resized using nearest-neighbor interpolation. Each connected component in the ground-truth mask is regarded as an individual target $T_i$. To determine how much surrounding context should be associated with this target, we first compute its equivalent diameter

$$d_i = 2\sqrt{|T_i|/\pi}. \tag{3}$$

We then expand the target mask outward by $\rho_i = \lceil d_i \rceil$ pixels and define its local background as

$$B_i = [\mathrm{Dilate}(T_i; \rho_i) \cap \Omega] \setminus \bigcup_j T_j, \tag{4}$$

where $\Omega$ denotes the valid image region and $\bigcup_j T_j$ contains all annotated targets in the image. Thus, $B_i$ contains only valid background pixels around $T_i$, without including the target itself or other nearby targets. The corresponding observations are $x_{t,i} = x|_{T_i}$ and $x_{b,i} = x|_{B_i}$. We use the same construction for all datasets. This step only constructs a target and its corresponding local background pair for subsequent measurements.

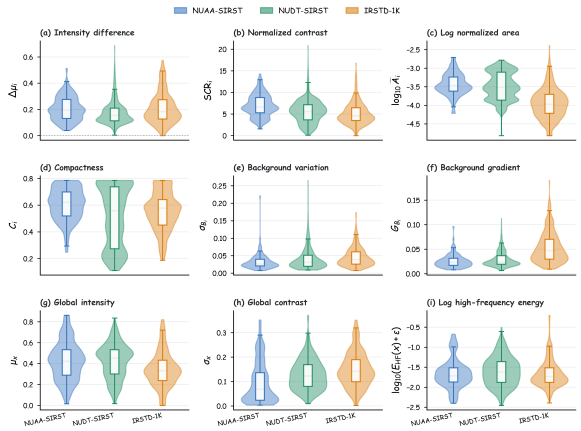


Figure 5: Cross-domain distributions of the observable target–background relation metrics. Panels (a)–(b), (c)–(d), (e)–(f), and (g)–(i) correspond to local saliency, target morphology, background clutter, and imaging style, respectively. Panels (c) and (i) use logarithmic mappings for normalized area and $E_{\mathrm{HF}}(x)$, respectively.

**Target-Background Relation Metrics.** Following the main formulation, the relation between target $T_i$ and its local background $B_i$ depends on local saliency, target morphology, background clutter, and imaging style. Accordingly, for empirical analysis, we construct a descriptor of $r_i = \Phi(x_{t,i}, x_{b,i})$ by combining measurements of these four factors:

$$r_{\mathrm{obs},i} = [r_{\mathrm{sal},i}, r_{\mathrm{mor},i}, r_{\mathrm{clu},i}, r_{\mathrm{sty},i}], \tag{5}$$

where $r_{\mathrm{sal},i}$ measures the intensity contrast between the target and its local background; $r_{\mathrm{mor},i}$ describes the target scale and

shape; $r_{\mathrm{clu},i}$ measures the intensity variation and structural complexity of the local background; and $r_{\mathrm{sty},i}$ summarizes the image-level intensity and frequency characteristics under which the pair is observed. Each term is a group of metrics whose exact components are defined below. We use $\epsilon = 10^{-6}$ throughout.

**Local Saliency.** Let $\mu_{T_i}$ and $\mu_{B_i}$ denote the mean intensities of target $T_i$ and its local background $B_i$, respectively. We characterize local saliency using the signed intensity difference $\Delta\mu_i$ and its background-normalized contrast $\mathrm{SCR}_i$:

$$r_{\mathrm{sal},i} = [\Delta\mu_i, \mathrm{SCR}_i]. \tag{6}$$

The two components are defined as

$$\Delta\mu_i = \mu_{T_i} - \mu_{B_i},\ \mathrm{SCR}_i = \frac{\Delta\mu_i}{\sigma_{B_i} + \epsilon}. \tag{7}$$

Here, the background standard deviation is

$$\sigma_{B_i} = \sqrt{\frac{1}{|B_i|}\sum_{p\in B_i}(x(p) - \mu_{B_i})^2}. \tag{8}$$

A positive $\Delta\mu_i$ indicates that the target is brighter than its local background, while a larger $\mathrm{SCR}_i$ indicates stronger target saliency relative to background fluctuations.

**Target Morphology.** Let $A_{T_i} = |T_i|$ and $P_{T_i}$ denote the area and perimeter of target $T_i$, respectively. We characterize target morphology using normalized area $\widetilde{A}_i$ and compactness $C_i$:

$$r_{\mathrm{mor},i} = [\widetilde{A}_i, C_i], \tag{9}$$

where

$$\widetilde{A}_i = \frac{A_{T_i}}{HW},\ C_i = \frac{4\pi A_{T_i}}{P_{T_i}^2 + \epsilon}, \tag{10}$$

and $H = W = 256$. Here, $\widetilde{A}_i$ measures the relative target scale, whereas $C_i$ measures shape compactness. A larger $C_i$ indicates a more regular and spatially concentrated target shape.

**Background Clutter.** Background clutter reflects both intensity fluctuations and spatial structures around the target. We characterize it using the background standard deviation $\sigma_{B_i}$ in Eq. 8 and the average gradient strength $G_{B_i}$:

$$r_{\mathrm{clu},i} = [\sigma_{B_i}, G_{B_i}], \tag{11}$$

where

$$G_{B_i} = \frac{1}{|B_i|}\sum_{p\in B_i}\|\nabla x(p)\|_2. \tag{12}$$

The gradient $\nabla x$ is computed using the Sobel operator. Thus, $\sigma_{B_i}$ measures intensity fluctuations, whereas $G_{B_i}$ measures the strength of local edges and textures. Larger values indicate a less uniform and more cluttered local background.

**Imaging Style.** We characterize the image-level appearance under which each target–background pair is observed using global intensity $\mu_x$, global contrast $\sigma_x$, and high-frequency energy ratio $E_{\mathrm{HF}}(x)$:

$$r_{\mathrm{sty},i} = [\mu_x, \sigma_x, E_{\mathrm{HF}}(x)], \tag{13}$$

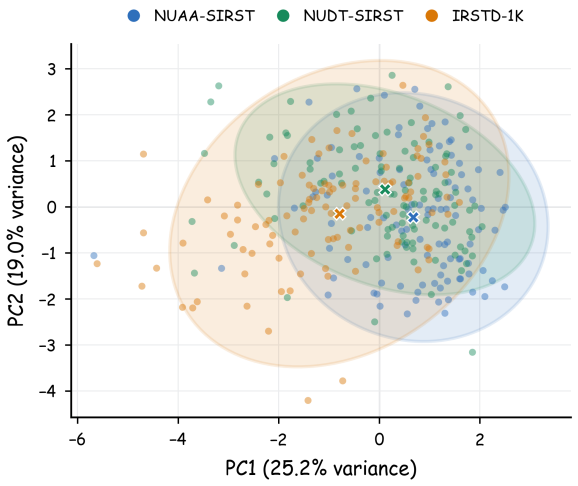


Figure 6: PCA visualization of the joint target-background relation distribution. Points, crosses, and ellipses denote target-background pairs, dataset centroids, and 90% probability regions, respectively.

where $\mu_x$ and $\sigma_x$ are the mean and standard deviation of all intensities in image $x$, respectively. Let $F_x$ be the centered Fourier spectrum of $x - \mu_x$. We compute the high-frequency energy ratio as

$$E_{\mathrm{HF}}(x) = \frac{\sum_{(u,v)\notin\mathcal{L}}|F_x(u,v)|^2}{\sum_{u,v}|F_x(u,v)|^2 + \epsilon}, \tag{14}$$

where $\mathcal{L}$ is the centered low-frequency region whose width and height are one quarter of the spectrum size. All targets in the same image therefore share the same imaging-style descriptor.

## A.2 Relation Distributions in Unseen Domains

**Marginal Distributions.** Fig. 5 shows clear cross-domain differences across all four relation factors. NUAA-SIRST exhibits higher target–background contrast, whereas IRSTD-1K generally contains smaller targets and stronger local background gradients. Differences in imaging style further show that the shift spans multiple relation properties.

**Joint Distribution.** We concatenate the metrics in Eq. 5 into $\mathbf{r}_{\mathrm{obs},i}$ and standardize each scalar dimension as

$$\widetilde{r}_{i,j} = \frac{r_{i,j} - \mu_j}{\sigma_j + \epsilon}, \tag{15}$$

where $\mu_j$ and $\sigma_j$ are estimated from an equally weighted mixture of the three test domains. PCA is fitted with the same dataset-level weighting to avoid dataset-size bias. The first two components explain $44.2\%$ of the variance. As shown in Fig. 6, IRSTD-1K and NUAA-SIRST shift toward negative and positive PC1, respectively, while NUDT-SIRST has a higher center on PC2. Despite partial overlap, these systematic displacements provide joint-distribution evidence of relation shift.

Table 7: Pairwise standardized 1-Wasserstein distances between relation descriptors. Brackets report image-cluster bootstrap 95% confidence intervals; larger values indicate stronger shifts.

| Relation factor | NUAA–NUDT | NUAA–IRSTD | NUDT–IRSTD |
|---|---|---|---|
| Local saliency | 0.45 [0.32, 0.61] | 0.45 [0.37, 0.62] | 0.33 [0.26, 0.42] |
| Target morphology | 0.42 [0.38, 0.53] | 0.54 [0.44, 0.68] | 0.64 [0.57, 0.71] |
| Background clutter | 0.23 [0.16, 0.36] | 0.81 [0.63, 0.99] | 0.61 [0.49, 0.75] |
| Imaging style | 0.26 [0.24, 0.36] | 0.43 [0.34, 0.58] | 0.31 [0.24, 0.40] |
| Average | 0.34 [0.31, 0.42] | 0.56 [0.50, 0.66] | 0.47 [0.42, 0.54] |

**Quantitative Distribution Distance.** For each relation-factor group $g$, we average the standardized 1-Wasserstein distances of its constituent metrics:

$$D_g(\mathcal{D}_a, \mathcal{D}_b) = \frac{1}{|g|} \sum_{j \in g} W_1\left(P_a(\widetilde{r}_j), P_b(\widetilde{r}_j)\right), \tag{16}$$

where the four groups follow Eq. 5; the Average row weights them equally. Larger values indicate stronger shifts. Distances are computed from the original standardized metrics, rather than the logarithmic values used only for visualization. We obtain 95% confidence intervals from 2,000 image-cluster bootstrap resamples, retaining all targets from each resampled image. Table 7 shows the largest average discrepancy for NUAA–IRSTD (0.56), followed by NUDT–IRSTD (0.47) and NUAA–NUDT (0.34). Background clutter dominates both comparisons involving IRSTD-1K, while target morphology contributes strongly to NUDT–IRSTD. Together with the marginal and PCA results, these distances establish a systematic, multi-factor relation shift across domains.

### A.3 Relation Shift and Source-Model Failures

**Relation-Deviation Score.** We train SCTransNet (Yuan et al. 2024) on IRSTD-1K and NUDT-SIRST and evaluate it on the unseen NUAA-SIRST test set. Each relation descriptor is standardized as in Eq. 15, using statistics estimated only from the two source training sets.

Let $\widehat{\mathbf{r}}^{s}_{i,g}$ denote the standardized subvector of relation factor $g$, where $\mathcal{G} = \{\mathrm{sal}, \mathrm{mor}, \mathrm{clu}, \mathrm{sty}\}$. We define the factor-balanced distance between instances $i$ and $q$ as

$$d(i,q) = \left[ \sum_{g \in \mathcal{G}} \frac{\left\| \widehat{\mathbf{r}}^{s}_{i,g} - \widehat{\mathbf{r}}^{s}_{q,g} \right\|_2^2}{|g|} \right]^{1/2}, \tag{17}$$

where division by $|g|$ gives each relation factor equal weight.

We then build a source reference bank $\mathcal{R}_s$ containing equal numbers of instances from the two source domains. The relation-deviation score is the mean distance to the $k = 20$ nearest source instances:

$$s_i = \frac{1}{k} \sum_{q \in \mathcal{N}_k(i; \mathcal{R}_s)} d(i,q), \; k = 20. \tag{18}$$

A larger $s_i$ indicates stronger deviation from the source relation patterns.

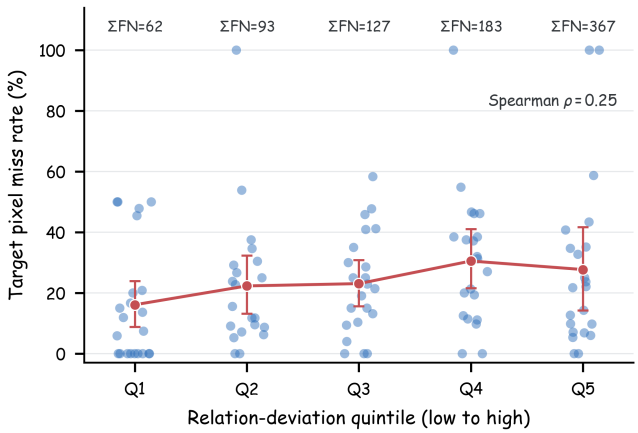


Figure 7: Relation deviation versus target-pixel miss rate on unseen NUAA-SIRST. SCTransNet is trained on IRSTD-1K and NUDT-SIRST. Q1–Q5 are equal-count groups ordered by deviation. Blue points show individual targets; red markers show group means with image-cluster bootstrap 95% confidence intervals; top labels report false-negative pixels.

**Detection Failure Measure.** For each target $T_i$, we measure detection failure by the proportion of its region not recovered by the binary prediction $\widehat{M}$:

$$\mathrm{FN}_i = |T_i \setminus \widehat{M}|, \; m_i = \frac{\mathrm{FN}_i}{|T_i|} = 1 - \frac{|T_i \cap \widehat{M}|}{|T_i|}. \tag{19}$$

Here, $\mathrm{FN}_i$ is the number of missed target pixels and $m_i$ is the corresponding miss rate. A larger $m_i$ indicates a more severe detection failure.

**Results.** Across the 108 NUAA-SIRST targets, SCTransNet misses 832 of 3,334 target pixels, yielding an overall miss rate of 24.95%. We rank these targets by their relation-deviation scores and divide them into five equal-count groups. As shown in Fig. 7, the mean miss rate increases from 16.12% in Q1 to 30.53% in Q4 and remains high at 27.71% in Q5.

At the instance level, the relation-deviation score is positively correlated with the miss rate (Spearman $\rho = 0.25$, image-cluster bootstrap 95% CI $[0.05, 0.45]$). The correlation remains after controlling for normalized target area (partial Spearman $\rho = 0.26$, 95% CI $[0.07, 0.45]$) and changes little for $k = 10, 20, 50$ ($\rho = 0.24, 0.25, 0.26$). These results indicate that test targets farther from the source-domain target-background relation patterns generally suffer higher detection miss rates.

## B Additional Details of Target-Background Relation Intervention

This section specifies the seven operators used in Target-Background Relation Intervention (TBRI).

### B.1 Operator Overview

Let $x \in [0,1]^{H \times W}$ denote a normalized infrared image and $y \in \{0,1\}^{H \times W}$ its binary annotation. The target and back-

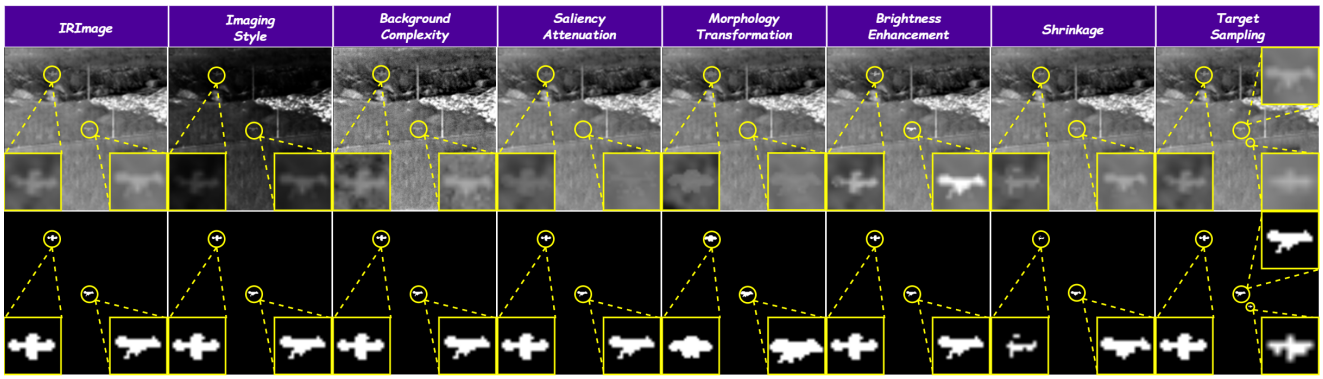


Figure 8: Qualitative examples of the seven TBRI operators. The upper and lower rows show the intervened images and corresponding annotations.

ground masks are $M_t = y$ and $M_b = 1 - y$, respectively. We organize the seven TBRI operators into two branches:

$$\begin{aligned} \mathcal{T} &= \mathcal{T}_{\text{bg}} \cup \mathcal{T}_{\text{tar}}, \\ \mathcal{T}_{\text{bg}} &= \{\mathcal{B}_{\text{sty}}, \mathcal{B}_{\text{clu}}\}, \\ \mathcal{T}_{\text{tar}} &= \{\mathcal{A}_{\text{sal}}, \mathcal{A}_{\text{mor}}, \mathcal{A}_{\text{bri}}, \mathcal{A}_{\text{shr}}, \mathcal{A}_{\text{sam}}\}. \end{aligned} \tag{20}$$

The subscripts sty, clu, sal, mor, bri, shr, and sam denote imaging style, background complexity, saliency attenuation, morphology transformation, brightness enhancement, shrinkage, and target sampling, respectively. For each activated sample, one valid operator is selected from $\mathcal{T}$. Appearance-only interventions preserve $y$, whereas support-changing interventions update it accordingly. Representative image-annotation pairs are shown in Fig. 8.

## B.2 Background Interventions

The background branch changes the context in which a target is observed while preserving its annotation.

**Imaging Style.** This operator jointly perturbs global brightness, contrast, and nonlinear intensity response:

$$x_{\text{sty}} = \text{clip}_{[0,1]}\left(\text{clip}_{[0,1]}[\mu_x + c_s(x - \mu_x) + b_s]^{\gamma_s}\right) \tag{21}$$

where $\mu_x$ is the mean image intensity, and $c_s$, $b_s$, and $\gamma_s$ control contrast, brightness, and gamma response, respectively. The transformation is applied to the entire image, yielding $(\widetilde{x}, \widetilde{y}) = (x_{\text{sty}}, y)$.

**Background Complexity.** This operator modifies local fluctuations and structures only within $M_b$. Let $\overline{x}_{15} = \text{AvgPool}_{15}(x)$ and $h_5 = x - \text{AvgPool}_5(x)$. We construct

$$x_{\text{clu}} = \overline{x}_{15} + c_b(x - \overline{x}_{15}) + \lambda_h h_5 + n_w + n_s, \tag{22}$$

where $c_b$ and $\lambda_h$ control local contrast and high-frequency enhancement, while $n_w$ and $n_s$ denote white and spatially smoothed noise. The original target is retained:

$$\widetilde{x} = M_t \odot x + M_b \odot x_{\text{clu}}, \ \widetilde{y} = y. \tag{23}$$

## B.3 Target Interventions

The target branch modifies target intensity, support, or occurrence using the surrounding background as a local reference. We define the reference ring as

$$R_b = \text{Dilate}(M_t; r_{\text{loc}}) \setminus \text{Dilate}(M_t; 1), \tag{24}$$

after excluding all target pixels. Its mean and standard deviation are denoted by $\mu_{R_b}$ and $\sigma_{R_b}$. If the ring is too small, these statistics are estimated from $M_b$.

**Saliency Attenuation.** This operator reduces the target–background contrast while preserving the target support. We measure the mean and peak contrasts as

$$\begin{aligned} \text{CNR} &= \frac{|\mu_{M_t} - \mu_{R_b}|}{\sigma_{R_b} + \epsilon}, \\ \text{PSNR}_t &= \frac{|p_t - \mu_{R_b}|}{\sigma_{R_b} + \epsilon}, \end{aligned} \tag{25}$$

where $p_t$ is the maximum target intensity for a bright target and the minimum target intensity for a dark target.

Given the sampled contrast levels $\tau_c$ and $\tau_p$, the retained contrast factor is

$$\begin{aligned} \alpha_0 &= \min\left\{\frac{\tau_c}{\text{CNR} + \epsilon}, \frac{\tau_p}{\text{PSNR}_t + \epsilon}\right\}, \\ \alpha &= \text{clip}_{[\alpha_{\min}, \alpha_{\max}]}(\alpha_0). \end{aligned} \tag{26}$$

The target residual relative to the local background is then scaled by $\alpha$:

$$\begin{aligned} \widetilde{x} &= M_b \odot x + M_t \odot [\mu_{R_b} + \alpha(x - \mu_{R_b})], \\ \widetilde{y} &= y. \end{aligned} \tag{27}$$

A smaller $\alpha$ moves the target intensity closer to its local background, thereby producing stronger saliency attenuation.

**Morphology Transformation.** This operator changes target shape using a sampled offset set $\mathcal{O}_m$, including elongated, curved, broken, asymmetric, and block-like patterns:

$$M_t' = \mathbb{I}\left[\max_{(\Delta u, \Delta v) \in \mathcal{O}_m} \text{Shift}(M_t; \Delta u, \Delta v) > 0\right]. \tag{28}$$

Target intensities are propagated with the same offsets and averaged in overlapping regions to obtain $x_{\text{mor}}$. Image and annotation are then updated jointly:

$$\widetilde{x} = (1 - M_t') \odot x + M_t' \odot x_{\text{mor}}, \ \widetilde{y} = M_t'. \tag{29}$$

**Brightness Enhancement.** This operator strengthens the positive target response relative to the local background:

$$x_{\text{bri}} = \mu_{R_b} + g_b|x - \mu_{R_b}|, \tag{30}$$

where $g_b > 1$ is a sampled gain. The target support is unchanged: $\widetilde{x} = M_b \odot x + M_t \odot x_{\text{bri}}$ and $\widetilde{y} = y$.

**Shrinkage.** This operator retains target pixels with strong local contrast and high centrality. For each $p \in M_t$, we first compute its absolute residual from the local-background mean and normalize it within the target:

$$\delta(p) = |x(p) - \mu_{R_b}|, \ \widehat{\delta}(p) = \frac{\delta(p) - \delta_{\min}}{\delta_{\max} - \delta_{\min} + \epsilon}, \tag{31}$$

where $\delta_{\min}$ and $\delta_{\max}$ are the minimum and maximum of $\delta(p)$ over $M_t$, respectively. Each target pixel is then ranked by

$$s(p) = 0.65\,\widehat{\delta}(p) + 0.35\left(1 - \frac{\|p - c_t\|_2}{d_{\max} + \epsilon}\right), \tag{32}$$

where $c_t$ is the target centroid and $d_{\max}$ is the maximum distance from $c_t$ to a target pixel. Given a sampled keep ratio $\eta$, the number of retained pixels is

$$n_{\text{keep}} = \min\{|M_t|, \max[1, \text{round}(\eta|M_t|)]\}. \tag{33}$$

The $n_{\text{keep}}$ pixels with the highest scores define the reduced support $M_t'$. Removed pixels are filled by

$$x_{\text{fill}} = 0.65\ \text{AvgPool}_k(x) + 0.35\,\mu_{R_b}. \tag{34}$$

With $M_r = M_t - M_t'$, the resulting pair is

$$\widetilde{x} = (1 - M_r) \odot x + M_r \odot x_{\text{fill}}, \ \widetilde{y} = M_t'. \tag{35}$$

**Target Sampling.** This operator samples an annotated source-domain target, rescales it, and places it at a valid background location $\ell$. Let $x_s^\ell$, $M_s^\ell$, and $\overline{M}_s^\ell$ denote its response, binary support, and soft blending mask, respectively. The resulting pair is

$$\widetilde{x} = (1 - \overline{M}_s^\ell) \odot x + \overline{M}_s^\ell \odot x_s^\ell, \ \widetilde{y} = y \vee M_s^\ell. \tag{36}$$

Candidate locations overlapping existing targets are rejected, and the sampled support is added to the annotation.

# C Implementation Details and Hyperparameter Analysis

This section supplements the implementation details omitted from the main paper, covering the model architecture and the operations used for Poincaré-ball relation modeling. We then conduct hyperparameter analyses.

## C.1 Implementation Details

**Encoder.** The encoder contains five feature levels with spatial resolutions $\{256, 128, 64, 32, 16\}$, each containing 32 channels. Let $X_1 = x$. At level $i$, a ResBlock(He et al. 2016) produces

$$F_i = \mathcal{R}_i(X_i), \ X_{i+1} = \text{MaxPool}_2(F_i), \tag{37}$$

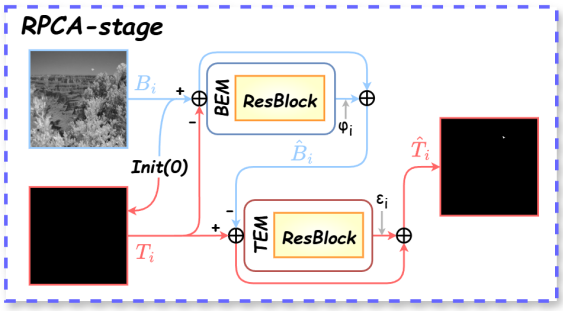


Figure 9: Architecture of the RPCA stage within the RPCA-Block.

where pooling is omitted after the final level. Each block first projects its input as $\widetilde{X} = \delta(\text{BN}[\text{Conv}_3(X)])$ and then applies a $5 \times 5$–$3 \times 3$ residual branch:

$$\begin{aligned} \widehat{X} &= \delta\Big(\text{BN}[\text{Conv}_5(\widetilde{X})]\Big), \\ \mathcal{R}_i(X) &= \delta\Big(\widetilde{X} + \text{BN}[\text{Conv}_3(\widehat{X})]\Big). \end{aligned} \tag{38}$$

Here, $\widehat{X}$ is the intermediate branch feature and $\delta$ denotes ReLU.

**Decoder.** The decoder reconstructs target representations from coarse to fine. At level $i$, the preceding target feature $\widehat{T}_{i+1}$ is bilinearly upsampled and fused with the adapted encoder feature $A_i$:

$$B_i = \text{Conv}_1\Big([A_i, \text{Up}(\widehat{T}_{i+1})]\Big). \tag{39}$$

At the coarsest level $I = 5$, we set $B_I = A_I$.

Each decoder level contains one RPCABlock, as shown in Fig. 9. Following (Liu et al. 2026a; Wu et al. 2024), the block performs one background update followed by one target update.

With $T_i = 0$, the two updates are

$$\begin{aligned} \widehat{B}_i &= \mathcal{R}_{B,i}(B_i - T_i) + \varphi_i(B_i - T_i), \\ \widehat{T}_i &= \mathcal{R}_{T,i}(T_i - \widehat{B}_i) + \varepsilon_i(T_i - \widehat{B}_i). \end{aligned} \tag{40}$$

Here, $\mathcal{R}_{B,i}$ and $\mathcal{R}_{T,i}$ denote two ResBlock. The learnable coefficients $\varphi_i$ and $\varepsilon_i$ are initialized to 0.01.

**Poincaré Mapping and Distance.** We detail the two Poincaré-ball operations used by HRM. Following Ganea et al. (Ganea, Bécigneul, and Hofmann 2018), each relation token $u_i$ is mapped from the tangent space at the origin to $z_i$ in the Poincaré ball:

$$z_i = \text{Exp}_0^c(u_i) = \tanh\big(\sqrt{c}\|u_i\|_2\big)\,\frac{u_i}{\sqrt{c}\|u_i\|_2}. \tag{41}$$

The target and background anchors follow the notation of the main paper:

$$a_t = \text{Exp}_0^c(\rho), \ a_b = \text{Exp}_0^c(-\rho), \tag{42}$$

Table 8: Ablation of the number of experts $E$ on the NUDT-SIRST + IRSTD-1K $\rightarrow$ NUAA-SIRST setting. All other components and training settings are kept unchanged. Results are reported in $mIoU$ (%), $F$-measure (%), $P_d$ (%), and $F_a$ ($10^{-6}$).

| | Detection performance | | | | Complexity | |
|---|---|---|---|---|---|---|
| $E$ | $mIoU \uparrow$ | $F \uparrow$ | $P_d \uparrow$ | $F_a \downarrow$ | Params (M) | FLOPs (G) |
| 1 | 75.74 | 86.19 | 99.07 | 15.25 | 1.06 | 7.99 |
| 2 | 77.57 | 87.37 | 99.23 | 12.20 | 1.08 | 8.13 |
| $4^\star$ | 78.58 | 88.00 | 99.56 | 13.10 | 1.13 | 8.42 |
| 8 | 76.54 | 86.71 | 99.12 | 8.07 | 1.22 | 9.01 |

where $a_t$ and $a_b$ are the target and background anchors, respectively, and $\rho$ controls their scale. For $a \in \{a_t, a_b\}$, the geodesic distance is

$$d_c(z_i, a) = \frac{2}{\sqrt{c}} \operatorname{artanh}\left(\sqrt{c}\,\|(-z_i) \oplus_c a\|_2\right). \tag{43}$$

Here, $\oplus_c$ denotes Möbius addition. The relation score used in the main paper is therefore

$$s_i = d_c(z_i, a_b) - d_c(z_i, a_t). \tag{44}$$

We use $c = 1$ and $\rho = 0.05$ in the default configuration.

## C.2 Hyperparameter Ablation

**Number of Experts.** We examine the number of experts $E$ in HMA under the same NUDT-SIRST + IRSTD-1K $\rightarrow$ NUAA-SIRST setting used for the ablation studies in the main paper. We vary $E \in \{1, 2, 4, 8\}$ while keeping all other training and architectural settings unchanged. The default configuration uses $E = 4$.

Performance improves as $E$ increases from 1 to 4, with $E = 4$ achieving the highest $mIoU$, $F$-measure, and $P_d$. Increasing $E$ to 8 further reduces $F_a$, but degrades the overlap-based metrics and increases the computational cost. We therefore use $E = 4$ as the default setting, which provides the best overall accuracy–complexity trade-off.

**Intervention Probability and Anchor Scale.** We further examine the activation probability $p$ of TBRI and the anchor scale $\rho$ of HRM. We vary one hyperparameter at a time while retaining the default value of the other and keeping all remaining settings unchanged. For TBRI, $p$ controls the proportion of training samples exposed to the intervention operators. For HRM, $\rho$ controls the separation between the target and background anchors in the tangent space. The default configuration uses $p = 0.5$ and $\rho = 0.05$.

Table 9: Hyperparameter sensitivity under the NUDT-SIRST + IRSTD-1K $\rightarrow$ NUAA-SIRST setting. Each parameter is varied independently; a star indicates the default. $mIoU$, $F$, and $P_d$ are reported in %, and $F_a$ in $10^{-6}$.

| Value | $mIoU \uparrow$ | $F \uparrow$ | $P_d \uparrow$ | $F_a \downarrow$ |
|---|---|---|---|---|
| *TBRI: activation probability $p$* | | | | |
| 0.25 | 77.47 | 87.30 | 99.37 | 15.61 |
| $0.50^\star$ | 78.58 | 88.00 | 99.56 | 13.10 |
| 0.75 | 76.77 | 86.86 | 98.67 | 21.18 |
| 1.00 | 77.60 | 87.39 | 99.07 | 13.48 |
| *HRM: anchor scale $\rho$* | | | | |
| 0.01 | 74.79 | 85.57 | 96.29 | 31.59 |
| 0.03 | 76.77 | 86.86 | 98.07 | 24.05 |
| $0.05^\star$ | 78.58 | 88.00 | 99.56 | 13.10 |
| 0.07 | 75.52 | 86.05 | 97.96 | 26.56 |
| 0.10 | 76.62 | 86.76 | 98.14 | 31.41 |

Table 9 shows that TBRI performs best at the moderate activation probability $p = 0.50$, achieving the highest $mIoU$, $F$-measure, and $P_d$, together with the lowest $F_a$. Reducing $p$ to 0.25 weakens the overall performance, whereas increasing it to 0.75 or 1.00 provides no further improvement. In particular, $p = 0.75$ increases $F_a$ to 21.18.

HRM exhibits a similar intermediate optimum. The default scale $\rho = 0.05$ achieves the best result across all four metrics. Both smaller scales (0.01 and 0.03) and larger scales (0.07 and 0.10) reduce the detection accuracy and increase false alarms. These results support the default configuration $p = 0.50$ and $\rho = 0.05$.