Domain-Auxiliary Infrared Moving Small Target Detection by Learning to Overlook Domain Discrepancy

Shengjia Chen, Luping Ji*, Shuang Peng, Sicheng Zhu, Mao Ye

School of Computer Science and Engineering, University of Electronic Science and Technology of China, China
shengjiachen@std.uestc.edu.cn, jiluping@uestc.edu.cn, {shuangpeng, sichengzhu}@std.uestc.edu.cn, maoye@uestc.edu.cn

Abstract

Currently, almost all traditional infrared small target detection methods work on the assumption that training and test sets always belong to the same domain, and training samples are sufficient. However, in real applications, a new detection task could often have no sufficient training samples from a special domain. In this situation, adopting the auxiliary data from big-sample domains is usually believed to be one of the most potential solutions. However, exceeding expectations, it is found that simply adding auxiliary samples cannot often be always effective, even causing performance decline, due to existing infrared domain shift. To overcome this unexpected problem, we propose the first infrared moving small target detection framework with domain-auxiliary supports by *Learning to Overlook Domain Discrepancy* (**Loddis**). This framework consists of three primary processing stages: correlation weakening, domain confusing, and target consistency contrastive learning. Breaking through traditional learning paradigm, through auxiliary data, it enables the model to focus more on targets themselves, and less on image backgrounds, minimizing the sensitivity to domain discrepancy. The extensive experiments on 6 different-domain datasets show the effectiveness and superiority of the proposed Loddis framework for infrared small target detection.

Code — https://github.com/UESTC-nnLab/Loddis

Introduction

Different from general object detection, infrared small target detection (ISTD), benefiting from its independence from external light sources and its all-weather visibility capability, has found increasingly wide and important applications (Zhao et al. 2022), including forest fire monitoring, satellite remote sensing and military use (Chen et al. 2025b). It has garnered extensive research attention over the past decades (Zeng, Li, and Peng 2006; Dai et al. 2021).

Compared to the general objects of visible-light (Chen, Li, and Tang 2020), infrared small targets possess two typical characteristics: *Small* and *Dim* (Chen et al. 2023). Although their actual physical sizes could be large (Deshpande et al. 1999), *e.g.*, the big airplanes in infrared images (Sun et al. 2023), the targets shown in these infrared images often have a small area relative to image backgrounds, due to the impacts of imaging distances and imaging means. They may even lack clear shape, texture, semantic features or distinct brightness (Zeng, Li, and Peng 2006; Zhang et al. 2022). Given these two characteristics above, it is often an extremely challenging issue for ISTD methods to effectively learn and accurately detect small targets in highly complex infrared application scenarios (Zhang et al. 2024).

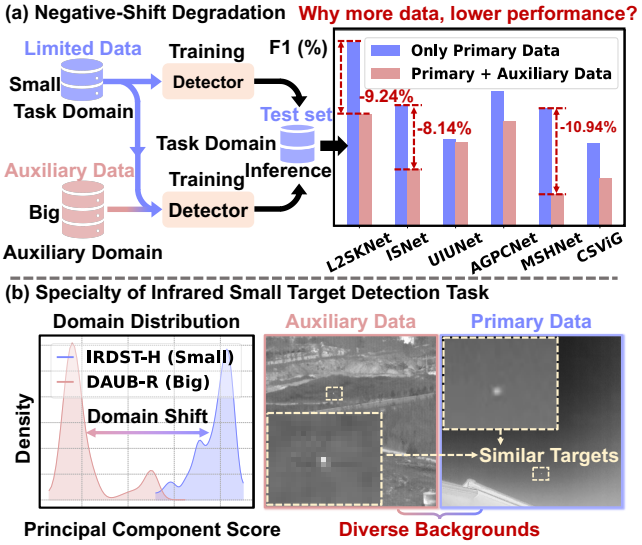

Figure 1: The negative-shift degradation problem in ISTD.

At present, beyond traditional methods (Chen et al. 2013), data-driven ones have been becoming one of research hot spots in ISTD (Chen et al. 2024b). Usually, this category of methods could achieve impressive detection performance (Zhu et al. 2023), when training and test are performed on a same task domain, and there are sufficient training samples. Once no sufficient samples, they cannot often work.

In real applications, some new tasks, such as the detection for the unseen targets before, could often only have much-limited samples. In this situation, how to effectively train a special detection model seems very challenging. For this challenge, besides classic few-shot learning schemes (Ravi and Larochelle 2017), the additional adoption of auxiliary

*Corresponding Author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

data is also believed to be one of the most potential solutions. It means that, for the sample-limited (small domain) training task, we could utilize the plentiful samples of other domains as the auxiliary data to enhance the training. However, in exploration experiments, inconsistent with our expectations, many methods, $e.g.$, MSHNet (Liu et al. 2024), were found that only simply adding auxiliary samples could often not be always effective, even causing decline, although more samples are used, as shown in Figure 1(a).

One of possible reasons is that this degradation is caused by the domain shift from auxiliary domain to task domain (Wu et al. 2024), as illustrated in Figure 1(b). Influenced by the domain shift, a similar infrared small target in different backgrounds would often not correctly recognized (Zhu et al. 2024), because domain shift often weakens the feature learning to small targets in task domain.

Domain shift is mainly brought by infrared background. To address the influence of domain shift ($i.e.$, Negative-Shift Degradation) on the training by auxiliary data, in model training, it is necessary to guide the detection model paying more attention to the targets themselves, rather than the backgrounds. In view of this, this paper explores the first framework (*i.e.,* Loddis) to effectively learn how to overlook domain discrepancy. This framework mainly consists of three parts: Correlation Weakening, Domain Confusing, and Target Consistency Contrastive Learning. In detail, the first part is designed to weaken the feature correlation between targets and backgrounds. The second one is designed to reduce the sensitivity of model to domains. In addition, the last one is for enhancing the feature consistency of targets in different-domain backgrounds.

In summary, the primary contributions of our work for infrared small target detection field are as follows:

(I) A new phenomenon of negative-shift degradation is discovered. To address it, we propose a new framework with auxiliary data ($i.e.$, Loddis). It's an original work to tackle the domain shift in infrared moving small target detection.

(II) In our Loddis, a) Correlation Weakening is designed to reduce the feature correlation of target and background from strong to weak, b) Domain Confusing is designed to minimize the domain-related information in target features, and c) Target Consistency Contrastive Learning is proposed to enhance the feature consistency of the same target in different-domain backgrounds.

(III) The extensive experiments on 6 datasets with domain shift are performed to demonstrate the superiority of our Loddis. It could consistently improve the base performance of detection model on task-domain datasets, by the training supports of auxiliary data from other domains.

Related Works

Infrared Moving Small Target Detection

Current mainstream schemes could be divided into two categories: model-driven (Moradi, Moallem, and Sabahi 2020) and data-driven (Yuan et al. 2024). Model-driven approaches include human visual systems (Chen et al. 2013), optimization-based methods (Zhang and Peng 2019) and background consistency (Deshpande et al. 1999). Representative methods separate targets utilizing the low-rank characteristics of backgrounds (Liu et al. 2023).

With the development of deep neural networks, data-driven schemes have advanced significantly and become a mainstream paradigm (Wu et al. 2025). Existing methods mainly rely on extracting spatial visual features for detection (Yang et al. 2025). For example, ISNet (Zhang et al. 2022) improves small target detection capabilities by implementing cross-level feature fusion and extracting valuable edge features. Similarly, UIUNet (Wu, Hong, and Chanussot 2022) introduces an attention-based nested network that preserves target resolution and enhances both global and local context using an interactive attention module.

In addition, to address the issue of features being easily lost in deep layers, DNANet (Li et al. 2023) proposes a dense nested convolutional network. However, these methods have a large number of parameters and slow inference speed. In view of this, RDIAN (Sun et al. 2023) proposes an efficient method with low complexity. Furthermore, to implement a small parameter number model, RPCANet (Wu et al. 2024) presents a deep unfolding-based method that achieves high performance with only a minimal number of parameters.

However, the above single-frame methods could be easily affected by domain shifts in mixed-domain data due to their reliance on global visual features, leading to a loss of effectiveness in challenging video scenarios (Yan et al. 2023). Because the target-related visual cues extracted from individual frames are typically weak and could be inadequate to support accurate detection by models (Chen et al. 2024b).

Motion Modeling-based Detection

To extract discriminative target features from complex backgrounds and guide the model to focus on target-specific representations, a series of multi-frame methods based on motion modeling have emerged (Chen et al. 2025a). In contrast to single-frame schemes, multi-frame ones could extract additional motion features, providing them with better detection performance (Du and Hamdulla 2019).

For example, model-driven inter-frame difference method (Kim, Sun, and Kim 2014), detects targets by calculating the difference between two successive frames. In addition, some tensor-optimized methods, 4-D STT model (Wu et al. 2023), have demonstrated the high performance in ISTD.

Even though these model-driven methods based on motion modeling have made significant advancements in ISTD, they may struggle to effectively handle the real-world scenarios with complex noise (Zhu et al. 2023).

Therefore, to overcome these limitations, some data-driven methods based on motion modeling have begun to emerge continuously (Zhu et al. 2024). For instance, DTUM (Li et al. 2025) proposes an auxiliary module to assist single-frame methods in utilizing spatio-temporal features.

However, these existing motion modeling-based methods typically capture motion features across entire image range, and these features are heavily affected by domain shifts, limiting performance improvement (Peng et al. 2025). Unlike these methods, our method focuses solely on target features, guiding the model to learn to overlook domain discrepancy and thereby alleviating the impact of domain shift.

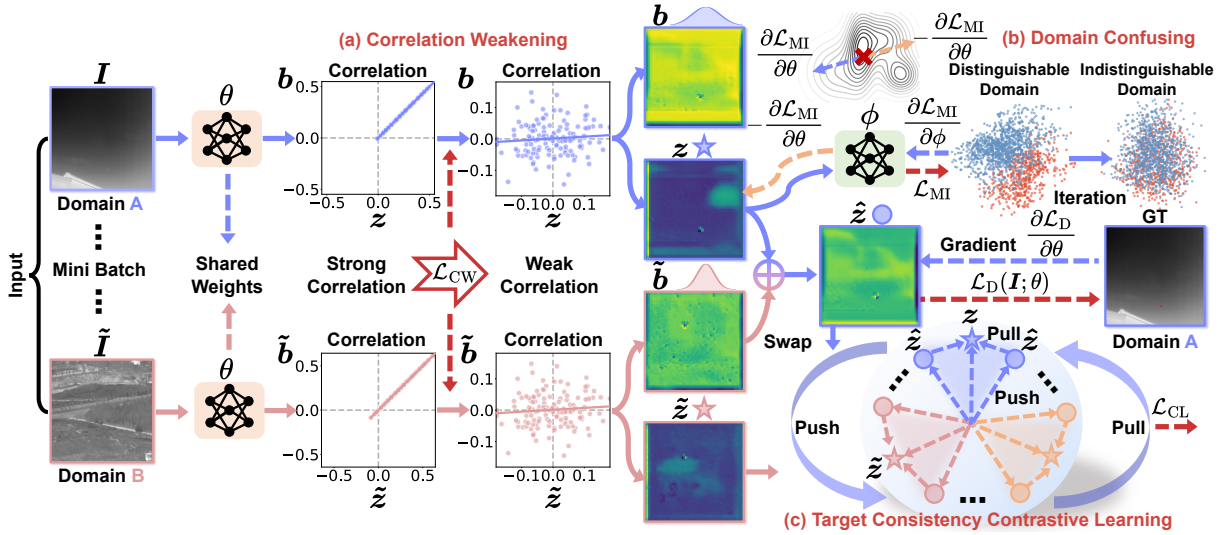

Figure 2: Our Loddis framework. A batch of mixed-domain frames (*e.g.*, images from two or three different domains), serves as original input. First, output visual feature set, generated by *Feature Extractor* θ with shared weights, gains target features $\boldsymbol{z}$ and background features $\boldsymbol{b}$ through minimizing the dependency through **Correlation Weakening**. Then, **Domain Confusing** is introduced using target features $\boldsymbol{z}$ to make *Feature Extractor* θ indistinguishable to domain discrepancy. Finally, **Target Consistency Contrastive Learning** is proposed to minimize feature differences of the same target across different domains (*e.g.,* $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$) and to maximize differences between target features from different domains (*e.g.,* $\boldsymbol{z}$ and $\tilde{\boldsymbol{z}}$).

Proposed Method

Overall Loddis Pipeline

An end-to-end Loddis framework is proposed for learning to overlook domain discrepancy, as illustrated in Figure 2.

Training phase. First, the visual input to the entire framework is a set of images with a batch size of n, denoted as $\boldsymbol{F} \in \mathbb{R}^{n\times 3\times w\times h}$, where w and h represent the height and width of input images, respectively. Although these images are single-channel data, they are processed as three-channel data in our implementation. For *Feature Extractor* θ, a simple and effective CSPDarknet (Ge et al. 2021) is used as the backbone network with an improved feature fusion layer (Chen et al. 2024a). This process results in a visual feature tensor $\boldsymbol{Z} \in \mathbb{R}^{n\times t\times u\times v}$, where t, u and v denote the channel, height and width of visual features, respectively.

Second, two new visual features $\boldsymbol{z} \in \mathbb{R}^{n\times c}$ and $\boldsymbol{b} \in \mathbb{R}^{n\times c}$ are obtained by projecting the output $\boldsymbol{Z}$ of *Feature Extractor* into a semantic subspace and passed through a pooling layer, where c is the channel of these new features. These two new features, $\boldsymbol{z}$ and $\boldsymbol{b}$, represent the target features and background features, respectively. They process $\boldsymbol{z}$ and $\boldsymbol{b}$ through **Correlation Weakening** to obtain statistically independent target and background features. This process is achieved by minimizing correlation weakening loss $\mathcal{L}_{\text{CW}}$.

Then, $\boldsymbol{z}$ is fed into **Domain Confusing**, where adversarial learning (Ganin et al. 2016) is performed by computing detection loss $\mathcal{L}_{\text{D}}$ and mutual information loss $\mathcal{L}_{\text{MI}}$ losses to reduce the ability of the *Feature Extractor* θ to distinguish domain discrepancy, thereby gradually eliminating domain-specific information in target features $\boldsymbol{z}$.

Finally, to further enhance the consistency and discriminability of target representations under different domain backgrounds, **Target Consistency Contrastive Learning** is performed. Specifically, the background feature $\boldsymbol{b}$ is replaced with $\tilde{\boldsymbol{b}}$ to construct a positive pair of the same target feature under real and swapped backgrounds (*e.g.,* $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$), while the features of other targets in the same batch serve as negative samples (*e.g.,* $\tilde{\boldsymbol{z}}$). Accordingly, the loss $\mathcal{L}_{\text{CL}}$ is computed by maximizing the similarity between positive pairs and minimizing the similarity with negative pairs.

In our scheme, on a base detection-head loss $\mathcal{L}_{\text{D}}$ (Chen et al. 2024a), we incorporate the loss from our proposed method. The total loss function $\mathcal{L}$ is defined by

$$\mathcal{L} = \mathcal{L}_{\text{D}} + \alpha\mathcal{L}_{\text{CW}} + \beta\mathcal{L}_{\text{MI}} + \eta\mathcal{L}_{\text{CL}}, \tag{1}$$

where $\mathcal{L}_{\text{CW}}$, $\mathcal{L}_{\text{MI}}$ and $\mathcal{L}_{\text{CL}}$ are our proposed new correlation weakening loss, mutual information loss and target consistency contrastive loss, respectively. α, β and η are hyperparameters to balance loss terms.

Inference phase. Only the loss $\mathcal{L}_{\text{D}}$ from basic detection head is computed, while three proposed losses $\mathcal{L}_{\text{CW}}$, $\mathcal{L}_{\text{MI}}$ and $\mathcal{L}_{\text{CL}}$ are not applied. The visual features obtained from the *Feature Extractor* θ, optimized by our method, are directly fed into the detection head as the final output.

Correlation Weakening

In our scheme, the goal is to explicitly disentangle visual features into two independent representations: target features and background features. If there exists a strong statistical correlation between these two types of features, it could negatively affect model learning under mixed-domain datasets, leading to performance degradation. Therefore, it

is necessary to design a method to quantify and minimize the correlation between target and background features.

In view of this, we introduce the Hilbert–Schmidt norm (Gretton et al. 2005) based on the covariance operator to measure the statistical dependence between two random variables. Specifically, given target features $\boldsymbol{z}$ and background features $\boldsymbol{b}$, their degree of dependence is defined as:

$$\mathcal{D}(\boldsymbol{z}, \boldsymbol{b}) = \|C_{zb}\|_{\text{HS}}^2, \tag{2}$$

$$C_{zb} = \mathbb{E}\big[(\boldsymbol{z} - \bar{\boldsymbol{z}})\,(\boldsymbol{b} - \bar{\boldsymbol{b}})^\top\big], \tag{3}$$

where C_{zb} denotes the cross-covariance operator between $\boldsymbol{z}$ and $\boldsymbol{b}$, and $\|\cdot\|_{\text{HS}}$ represents the Hilbert–Schmidt norm. μ_z denotes the mean of $\boldsymbol{z}$. A lower $\mathcal{D}(\boldsymbol{z}, \boldsymbol{b})$ value indicates weaker statistical dependence, implying a higher degree of independence. By substituting this covariance matrix C_{zb}, it could be further formulated as:

$$\mathcal{D}(\boldsymbol{z}, \boldsymbol{b}) = \mathbb{E}_{z,b}\,\mathbb{E}_{\acute{z},\acute{b}}\Big[(\boldsymbol{z}-\bar{\boldsymbol{z}})^\top(\acute{\boldsymbol{z}}-\bar{\boldsymbol{z}})\;(\boldsymbol{b}-\bar{\boldsymbol{b}})^\top(\acute{\boldsymbol{b}}-\bar{\boldsymbol{b}})\Big], \tag{4}$$

where $\boldsymbol{z}$ and $\acute{\boldsymbol{z}}$ are two independent but identically distributed samples of target feature vectors. Therefore, correlation weakening loss could be generally described as:

$$\mathcal{L}_{\text{CW}} = \frac{1}{(n-1)^2}\sum_{i=1}^{n}\mathcal{D}(\boldsymbol{z}_i, \boldsymbol{b}_i), \tag{5}$$

where n is the number of the samples in a batch. By optimizing this loss, the model could be optimized to capture statistically independent $\boldsymbol{z}$ and $\boldsymbol{b}$.

Domain Confusing

The *Feature Extractor* θ is enforced to lose its discriminative ability with respect to domain, enabling domain-irrelevant target representations. To achieve this purpose, adversarial training is employed to pursue a dual objective: enhancing detection performance while reducing sensitivity to domain discrepancy. The overall objective is formulated as follows:

$$\min_{\theta}\max_{\phi}\mathcal{J}(\theta, \phi) = \mathbb{E}[\mathcal{L}_{\text{D}}(\boldsymbol{I}, \boldsymbol{g}; \theta) - \lambda\mathcal{L}_{\text{MI}}(\boldsymbol{z}, \boldsymbol{d}; \phi)], \tag{6}$$

where ϕ denotes a *Domain Classifier* and θ represents a *Feature Extractor*. $\boldsymbol{g} \in \mathbb{R}^{n\times g}$ and $\boldsymbol{d} \in \mathbb{R}^{n\times m}$ denote ground truth and domain category, respectively. Hyperparameter λ controls the strength of domain confusing. To maximize the discrepancy between $\boldsymbol{z}$ and $\boldsymbol{d}$, the gradient of ϕ is multiplied by $-\lambda$ during backpropagation to influence θ. Therefore, for θ, minimizing the shared information between $\boldsymbol{z}$ and $\boldsymbol{d}$ is equivalent to maximizing the shared information of ϕ. Specifically, the gradient θ is described as follows:

$$\nabla_\theta\mathcal{J} = \mathbb{E}\left[\nabla_\theta\mathcal{L}_{\text{D}}(\boldsymbol{I}, \boldsymbol{g}; \theta) - \lambda\nabla_\theta\mathcal{L}_{\text{MI}}(\boldsymbol{z}, \boldsymbol{d}; \phi)\right]. \tag{7}$$

Thus, by applying chain rule, we obtain:

$$\frac{\partial \boldsymbol{z}}{\partial\theta}\left(\nabla_{\boldsymbol{z}}\mathcal{L}_{\text{D}}(\boldsymbol{I}, \boldsymbol{g}; \theta) - \lambda\nabla_{\boldsymbol{z}}\mathcal{L}_{\text{MI}}(\boldsymbol{z}, \boldsymbol{d}; \phi)\right). \tag{8}$$

Through this adversarial learning, ϕ maximization and θ minimization are achieved within a single backpropagation.

To intuitively characterize the shared information between target features $\boldsymbol{z}$ and domain $\boldsymbol{d}$, mutual information (Belghazi et al. 2018) could be employed as a metric. The mutual information $\mathcal{M}(\boldsymbol{z}; \boldsymbol{d})$ between $\boldsymbol{z}$ and $\boldsymbol{d}$ is defined as:

$$\mathcal{M}(\boldsymbol{z}; \boldsymbol{d}) = \mathbb{E}_{(z,d)\sim p(z,d)}\left[\log\frac{p(z,d)}{p(z)p(d)}\right]. \tag{9}$$

To estimate it, we utilize Donsker-Varadhan representation and reformulate it as a supremum over functions $T(\cdot)$:

$$\mathcal{M}(\boldsymbol{z}; d) = \sup_{T}\Big\{\mathbb{E}_{p(\boldsymbol{z},\boldsymbol{d})}\big[T(\boldsymbol{z}, \boldsymbol{d})\big] - \log\mathbb{E}_{p(\boldsymbol{z})\,p(\boldsymbol{d})}\big[e^{T(\boldsymbol{z},\boldsymbol{d})}\big]\Big\}. \tag{10}$$

To overcome its computational cost, an auxiliary classifier $q_\phi(\boldsymbol{d}|\boldsymbol{z})$ is used as a variational proxy for the true posterior $p(\boldsymbol{d}|\boldsymbol{z})$, thereby obtaining a tractable approximation. Its upper bounded could be described as follows:

$$\mathcal{M}(\boldsymbol{z}; \boldsymbol{d}) \le \mathbb{E}_{p(\boldsymbol{z},\boldsymbol{d})}\left[-\log q_\phi(d|\boldsymbol{z})\right] + \log m, \tag{11}$$

where m is the number of domain classes. Consequently, we could define the mutual information loss $\mathcal{L}_{\text{MI}}$ as:

$$\mathcal{L}_{\text{MI}} = \mathbb{E}_{p(\boldsymbol{z},\boldsymbol{d})}\left[-\log q_\phi(d|\boldsymbol{z})\right] \tag{12}$$

In summary, minimizing $\mathcal{L}_{\text{MI}}$ encourages the target features $\boldsymbol{z}$ learned by θ to be invariant to domain discrepancy.

Target Consistency Contrastive Learning

Explicitly disentangling features alone is insufficient to ensure that the learned target features $\boldsymbol{z}$ remain robust to domain-varying backgrounds. To further enhance this robustness, a new learning method is proposed.

First, by constructing domain-confused samples, the spurious association between the target and background domain is disrupted, forcing the model to rely solely on target-specific cues. Specifically, we define two index sets: $\mathcal{I}_0 = \{i|d_i = 0\}$ and $\mathcal{I}_1 = \{i|d_i = 1\}$, where $\mathcal{I}_0$ denotes the set of indices for all samples from domain 0. Next, the maximum number of possible background exchanges is determined by taking the minimum cardinality of these two index sets: $k = \min(|\mathcal{I}_0|, |\mathcal{I}_1|)$, where k denotes the maximum number of pairs that could be formed for background swapping Then, these backgrounds could be are swapped:

$$\hat{\boldsymbol{b}}_i = \begin{cases} \boldsymbol{b}_{\mathcal{I}_1[j]} & \text{if } i = \mathcal{I}_0[j],\ j < k \\ \boldsymbol{b}_{\mathcal{I}_0[j]} & \text{if } i = \mathcal{I}_1[j],\ j < k \\ \boldsymbol{b}_i & \text{otherwise} \end{cases}. \tag{13}$$

That is, the first k samples from domain 0 and domain 1 are paired and undergo background swapping.

Second, through contrastive learning, the model is encouraged to ensure that target features remain consistent under different background domain conditions, as follows:

$$\mathcal{L}_{\text{CL}} = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\exp(\text{sim}(\boldsymbol{z}_i, \hat{\boldsymbol{z}}_i)/\tau)}{\sum_{j=1}^{n}\exp(\text{sim}(\boldsymbol{z}_i, \hat{\boldsymbol{z}}_j)/\tau)}, \tag{14}$$

where $\text{sim}(\cdot)$ is the cosine similarity and τ is the temperature parameter. $\hat{\boldsymbol{z}}$ is the feature obtained by recombining the target with a swapped background. This enables the target features to focus solely on intrinsic target characteristics while suppressing the influence of background domains.

	Methods	(3IR+7DA) → IR				(3DA+7IR) → DA				(3TS+7DA) → TS			
		mAP$_{50}$	**Pr**	**Re**	**F1**	**mAP$_{50}$**	**Pr**	**Re**	**F1**	**mAP$_{50}$**	**Pr**	**Re**	**F1**
Vision	ISNet (Zhang et al. 2022)	13.73	36.32	38.36	37.31	45.85	48.19	**96.87**	64.36	36.40	57.37	64.36	60.66
	UIUNet (Wu et al. 2022)	16.74	35.97	47.00	40.75	50.91	62.60	82.58	71.22	30.75	54.40	57.92	56.11
	AGPCNet (Zhang et al. 2023)	18.48	47.65	39.77	43.35	55.63	60.93	92.24	73.39	35.21	57.27	62.50	59.77
	RDIAN (Sun et al. 2023)	21.01	32.57	65.86	43.58	40.03	44.88	90.41	59.98	42.10	60.67	70.77	65.33
	DNANet (Li et al. 2023)	16.59	33.33	50.32	40.10	44.42	58.14	77.34	66.38	51.92	70.79	**74.53**	72.61
	SCTransNet (Yuan et al. 2024)	26.52	49.40	54.54	51.84	46.89	53.71	88.52	66.86	48.26	71.56	68.82	70.16
	SIRST5K (Lu et al. 2024)	11.68	30.13	39.04	34.01	42.06	63.24	67.08	65.10	44.59	65.20	69.42	67.24
	CSViG (Lin et al. 2024)	14.10	27.81	51.71	36.17	58.17	67.45	87.49	76.17	46.00	66.31	70.64	68.41
	RPCANet (Wu et al. 2024)	14.53	25.24	58.79	35.32	40.09	44.41	91.56	59.81	27.82	46.77	60.77	52.86
	MSHNet (Liu et al. 2024)	11.37	32.55	35.86	34.12	43.95	71.47	62.50	66.68	49.65	68.63	73.11	70.80
	MLPNet (Wang et al. 2025)	10.17	16.75	61.30	26.32	37.83	44.83	85.69	58.87	32.99	52.84	63.39	57.63
	PConv (Yang et al. 2025)	25.27	43.21	59.39	50.02	53.00	58.09	92.75	71.44	38.76	55.09	71.16	62.10
	L2SKNet (Wu et al. 2025)	20.69	35.32	59.21	44.24	49.66	54.37	92.19	68.40	47.66	67.72	71.08	69.36
Motion	TMP (Zhu et al. 2024)	33.20	53.83	62.79	57.96	46.92	**96.35**	49.96	65.81	54.21	85.14	64.60	73.46
	SSTNet (Chen et al. 2024a)	39.02	56.91	69.25	62.47	57.31	93.19	62.29	74.67	56.84	**85.34**	67.42	75.33
	STME (Peng et al. 2025)	35.33	57.12	62.89	59.87	56.41	81.27	70.80	75.67	53.94	83.44	65.90	73.64
	DTUM (Li et al. 2025)	38.72	55.60	70.04	61.99	58.76	78.77	75.92	77.32	53.16	75.47	71.78	73.58
	Loddis (Ours)	**42.30**	**60.78**	**70.23**	**65.16**	**64.33**	92.48	70.42	**79.96**	**58.66**	81.00	73.39	**77.00**

Table 1: The quantitative detection performance comparisons on three mixed-domain datasets. For example, (3IR+7DA)→IR, indicates that the training set consists of 30% of task-domain dataset IRDST-H and 70% of auxiliary-domain dataset DAUB-R, with evaluated on the 100% test set of IRDST-H. The best metric is marked in green, and the second-best one is yellow.

Experiments

Implementation Details

All comparison methods are evaluated on three infrared moving small target datasets: IRDST-H (Chen et al. 2024b), DAUB-R(Chen et al. 2024a) and TSIRMT-M (Huang et al. 2024). Based on these three datasets, six new mixed-domain datasets are constructed. Two examples of these datasets are shown in Figure 3. A standard evaluation paradigm (Chen et al. 2024a) is strictly followed in experiments. Precision *(Pr)*, Recall *(Re)*, F1 score and mean Average Precision (mAP$_{50}$, with an Intersection over Union (IoU) threshold of 0.5) are used as evaluation metrics.

For all comparison methods, input images are resized to a resolution of 512×512 using a letterbox resizing (without data augmentation). Furthermore, all comparison methods are trained for 100 epochs with a batch size of 4. In each epoch, only 20% of the data is randomly selected for training. Initial learning rate is set to 0.01, using SGD as the optimizer, with a momentum of 0.937. The hyperparameters of our method α, β, η, λ and τ are set to 1, 1, 1, 1 and 0.2, respectively. In test, letterbox resizing is also used.

Comparisons with SOTA Ones

Quantitative Comparisons. Table 1 presents the quantitative comparisons across 18 recent methods, revealing two notable observations. One is that our Loddis establishes new SOTA benchmarks across most metrics, consistently securing the top performance indicators on three datasets. For example, on IRDST-H, Loddis could achieve the highest mAP$_{50}$ 42.30%, *Pr* 60.78%, *Re* 70.23% and F1 score 65.16%. The other is that our Loddis has better data adaptability than other approaches. For instance,

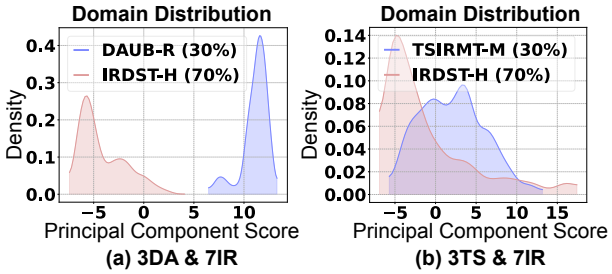

Figure 3: The distribution of two different-domain datasets.

on DAUB-R, DTUM (Li et al. 2025) achieves a SOTA F1 score of 77.32%, whereas Loddis sets a new benchmark with 79.96%, outperforming DTUM by 2.64%. However, on TSIRMT-M, DTUM attains an F1 score of 73.58%, while Loddis reaches 77.00%, reflecting a 3.42% improvement. These results suggest that our method demonstrates stronger adaptability across different datasets compared to DTUM.

Negative-Shift Degradation Comparisons. Figure 4 presents negative-shift degradation comparisons for representative methods, revealing two notable observations. One is that our Loddis establishes a new SOTA benchmark in terms of the F1 score, consistently securing the top performance indicators on six mixed-domain datasets. For example, on **(3DA+7IR)→DA** setting, our method achieves an F1 score of over 80%, surpassing all other approaches. The other is that our method consistently enhances the performance of a weaker baseline across all six datasets using auxiliary datasets, whereas other methods may lead to performance degradation. For example, on **(3DA+7TS)→DA** setting, our method improves from fourth place on dataset **P**

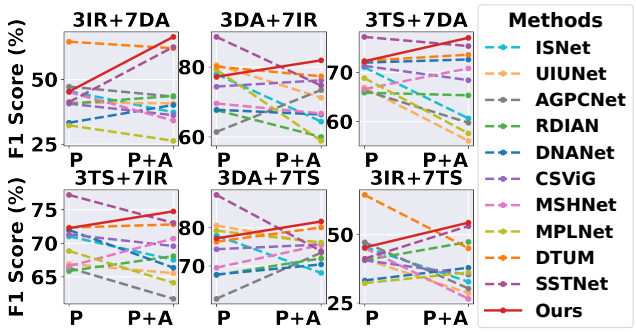

Figure 4: Negative-shift degradation comparisons on six mixed-domain datasets. **P:** primary data, **A:** auxiliary data.

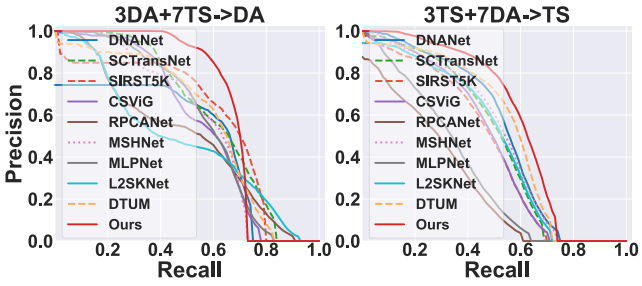

Figure 5: PR curve comparisons of representative methods.

to first place on **P+A**, while SSTNet drops from first to fifth.

PR Curve Comparisons. As usual, we adopt *Precision-Recall* (PR) curves to visually compare the comprehensive performance of methods on two mixed-domain datasets, as illustrated in Figure 5. The proximity of a method to top-right corner directly indicates its effectiveness. These figures clearly demonstrate that our PR curves outperform those of competing methods. Specifically, on **(3DA+7TS)→DA**, our PR curve consistently occupies higher positions, particularly near top-right. This pattern continues on **(3TS+7DA)→TS**, where our curve exceeds others toward top-right. Therefore, these PR curves highlight the superior balance of our method in precision and recall compared to other methods.

Inference Cost Comparisons. The comparisons of inference cost are presented in Table 2. From these comparisons, two clear findings emerge. First, although our proposed Loddis relies only on single-frame images, it still achieves a new SOTA performance. For example, SSTNet uses 5 frames and achieves an F1 score of only 62.47%, whereas our method achieves 65.16%. Second, our method achieves the highest inference speed while maintaining high performance. For example, CSViG reaches a high FPS of 40.48, but its mAP_{50} is only 14.10%. In contrast, our method achieves the highest FPS of 40.79 while maintaining an mAP_{50} of 42.30%.

Ablation Study

Effects of Different Assemblies. We undertake a series of ablation studies to evaluate the impact of various configurations within our method, as demonstrated in Table 3. Through comparisons, it is apparent that the individual

Methods	Frames	mAP_{50}↑	F1↑	FLOPs↓	Params↓	FPS↑
ISNet	1	13.73	37.31	265.74G	3.48M	20.02
UIUNet	1	16.74	40.75	456.70G	53.06M	19.28
AGPCNet	1	18.48	43.35	366.15G	14.88M	18.33
RDIAN	1	21.01	43.58	50.44G	**2.74M**	34.20
DNANet	1	16.59	40.10	135.24G	7.22M	7.21
SCTransNet	1	26.52	51.84	101.61G	13.71M	10.34
SIRST5K	1	11.68	34.01	182.61G	11.48M	6.17
CSViG	1	14.10	36.17	117.56G	5.81M	40.48
RPCANet	1	14.53	35.32	382.69G	3.21M	14.81
MSHNet	1	11.37	34.12	69.49G	6.59M	16.37
MLPNet	1	10.17	26.32	93.87G	23.79M	11.62
PConv	1	25.27	50.02	**7.89G**	2.91M	40.24
L2SKNet	1	20.69	44.24	76.00G	3.42M	30.54
TMP	5	33.20	57.96	92.85G	16.41M	12.75
SSTNet	5	39.02	62.47	123.59G	11.95M	10.38
STME	5	35.33	59.87	42.09G	9.93M	14.55
DTUM	5	38.72	61.99	128.16G	9.64M	12.77
Loddis	1	**42.30**	**65.16**	32.65G	7.75M	**40.79**

Table 2: The inference cost comparisons on 3IR+7DA → IR.

Settings	C	D	T	3IR+7DA		3DA+7IR		3TS+7DA	
				mAP_{50}	F1	mAP_{50}	F1	mAP_{50}	F1
w/o All	-	-	-	36.23	60.53	57.84	76.44	53.13	73.12
w/ C	✓	-	-	37.47	61.32	59.94	77.23	54.53	74.26
w/ D	-	✓	-	37.52	61.59	59.81	77.73	53.41	73.25
w/ T	-	-	✓	38.17	61.79	61.44	78.83	55.22	74.68
w/ C & D	✓	✓	-	38.05	61.95	60.99	77.83	56.61	74.53
w/ C & T	✓	-	✓	39.30	63.05	62.80	79.53	57.88	76.18
w/ D & T	-	✓	✓	41.39	64.57	62.04	78.76	57.01	75.63
w/ all	✓	✓	✓	**42.30**	**65.16**	**64.33**	**79.96**	**58.66**	**77.00**

Table 3: The ablation study on Loddis with different assemblies and settings. **C:** correlation weakening, **D:** domain confusing, **T:** target consistency contrastive learning.

components are consistently effective in enhancing detection capabilities. For instance, on **(3IR+7DA)→IR** setting, the baseline configuration, devoid of any specialized components, achieves mAP_{50} and F1 score of 36.23% and 60.53%, respectively. The integration of components (w/ T) improves these scores to 38.17% for mAP_{50} and 61.79% for F1. Ultimately, when all components are fully integrated, performance improves significantly, with an mAP_{50} of 42.30% and an F1 scores of 65.16%, reaching their best levels.

Effects of Correlation Weakening. Figure 6 presents a visualization of the correlation between target features $\boldsymbol{z}$ and background features $\boldsymbol{b}$. From it, we could observe that our proposed *Correlation Weakening* effectively weakens the statistical correlation between target and background features, enabling their separation. For example, in this case, sub-figure (a) shows that the original target and background features exhibit perfectly linear positive correlation, as they are both derived from the same features obtained by *Feature Extractor*. In contrast, our method establishes clear independence between target and background features.

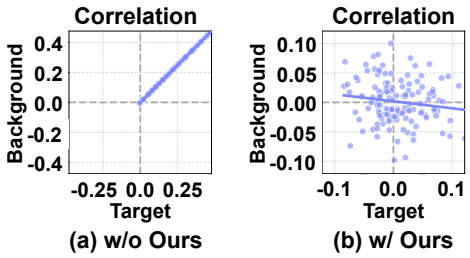

Figure 6: Visualization of the correlation between target features $\boldsymbol{z}$ and background features $\boldsymbol{b}$.

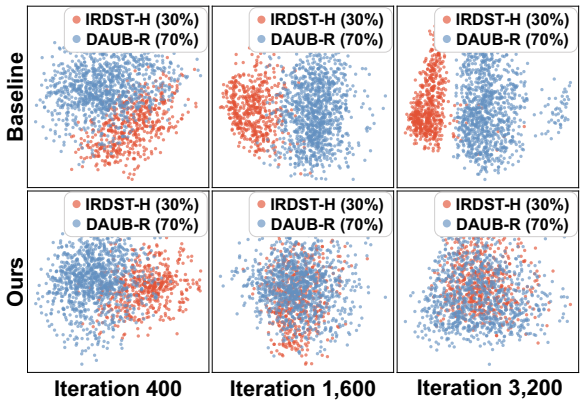

Figure 7: Domain classification visualization comparisons on the mixed-domain dataset configured as 3IR+7DA.

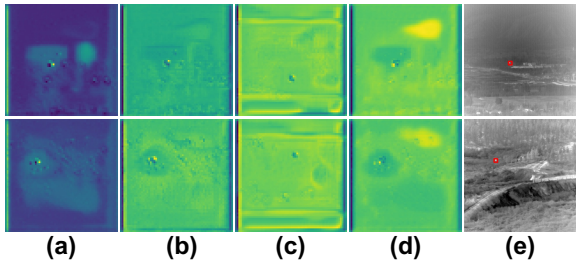

Figure 8: Visualization of feature maps. (a) $\boldsymbol{z}$, (b) $\boldsymbol{b}$, (c) $\tilde{\boldsymbol{b}}$, (d) $\hat{\boldsymbol{z}}$, (e) Ground Truth.

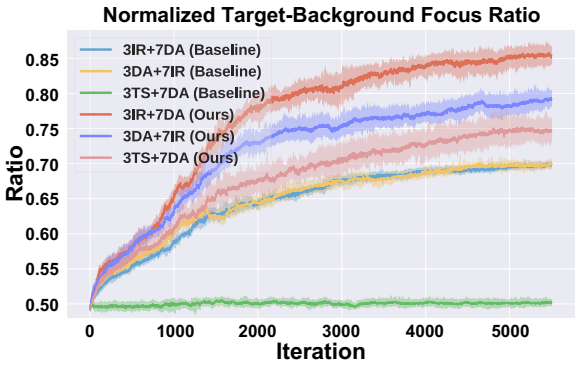

Figure 9: Target-background focus ratio curve. A ratio greater than 0.5 indicates more attention to target features, and vice versa indicates more attention to backgrounds.

Effects of Domain Confusing. Figure 7 presents classification visualization comparisons of two domains under the baseline (top) and our method (bottom) at early and late iterations. From it, we could observe that our *Domain Confusing* enables the model to learn features that avoid relying on domain discrepancy. For example, at the early stage of training (*i.e.* at iteration 400), both the baseline and our method are able to classify the domains effectively. However, by 1,600 iterations, while the baseline further improves its ability to distinguish between two domains, our method increasingly prevents the model from making such distinctions. Finally, baseline clearly separates two domains, whereas our method results in nearly complete domain confusing.

Effects of Target Consistency Contrastive Learning. To intuitively observe whether the target features $\hat{\boldsymbol{z}}$ with swapped backgrounds remain consistent with original target features $\boldsymbol{z}$, they are visualized as shown in Figure 8. From it, we could clearly find that the feature maps $\hat{\boldsymbol{z}}$ after background swapping retain the information of original targets $\boldsymbol{z}$. This indicates that our method could alter the background domain while preserving the consistency of target features.

Discussion. To further investigate how effectively our complete method overlook background discrepancy, we analyze the target-background focus ratio over training iterations, as illustrated in figure 9. This ratio serves as an indicator of the focus bias of the models towards targets over backgrounds, with higher values suggesting better target-specific representation. Compared to baselines, our method consistently achieves significantly higher focus ratios across all three cross-domain configurations. This indicates that our approach effectively guides the model to focus more on target-related features while overlooking domain discrepancy. These observations align well with our core motivation: learning to overlook domain discrepancy and enhancing target representation under domain shifts.

Conclusions

To address the issue of negative-shift degradation caused by the use of auxiliary data in training, this paper proposes the first paradigm of infrared small target detection that learns to overlook domain discrepancy ($i.e.$, Loddis). In detail, it firstly weakens the feature correlation of target and background by statistical distributions. Then, it suppresses the domain-related information in target features to achieve domain-indistinguishable representation learning. Finally, it combines cross-domain background swapping with contrastive learning to enforce the feature consistency of targets in different domains. The extensive experiments on six different-domain datasets demonstrate that our method could effectively alleviates infrared domain shift problem. On primary metrics ($e.g.$, mAP_{50} and F1), it not only obviously outperforms previous SOTA methods, but also refreshes a new FPS peak, with less FLOPs and Params. One of its limitations is that only two-domain scenario is explored. In the future, extending this paradigm to multi-domain scenario is worthy of further exploration.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant 62476049 and Grant 62276048.

References

Belghazi, M. I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, D. 2018. Mutual information neural estimation. In *International conference on machine learning*, 531–540. PMLR.

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, S.; Ji, L.; Duan, W.; Peng, S.; and Ye, M. 2025a. Motion Prior Knowledge Learning with Homogeneous Language Descriptions for Moving Infrared Small Target Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2186–2194.

Chen, S.; Ji, L.; Zhu, J.; Ye, M.; and Yao, X. 2024a. SSTNet: Sliced Spatio-Temporal Network With Cross-Slice ConvLSTM for Moving Infrared Dim-Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–12.

Chen, S.; Ji, L.; Zhu, S.; and Ye, M. 2025b. MICPL: Motion-Inspired Cross-Pattern Learning for Small-Object Detection in Satellite Videos. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4): 6437–6450.

Chen, S.; Ji, L.; Zhu, S.; Ye, M.; Ren, H.; and Sang, Y. 2024b. Toward Dense Moving Infrared Small Target Detection: New Datasets and Baseline. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–13.

Chen, S.; Li, Z.; and Tang, Z. 2020. Relation R-CNN: A graph based relation-aware network for object detection. *IEEE Signal Processing Letters*, 27: 1680–1684.

Chen, S.; Zhu, J.; Ji, L.; Pan, H.; and Xu, Y. 2023. AugTarget Data Augmentation for Infrared Small Target Detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Dai, Y.; Wu, Y.; Zhou, F.; and Barnard, K. 2021. Attentional local contrast networks for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 59(11): 9813–9824.

Deshpande, S. D.; Er, M. H.; Venkateswarlu, R.; and Chan, P. 1999. Max-mean and max-median filters for detection of small targets. In *Signal and Data Processing of Small Targets 1999*, volume 3809, 74–83. SPIE.

Du, P.; and Hamdulla, A. 2019. Infrared moving small-target detection using spatial–temporal local difference measure. *IEEE Geoscience and Remote Sensing Letters*, 17(10): 1817–1821.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59): 1–35.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

Gretton, A.; Bousquet, O.; Smola, A.; and Schölkopf, B. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*, 63–77. Springer.

Huang, Y.; Zhi, X.; Hu, J.; Yu, L.; Han, Q.; Chen, W.; and Zhang, W. 2024. LMAFormer: Local Motion Aware Transformer for Small Moving Infrared Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–17.

Kim, S.; Sun, S.-G.; and Kim, K.-T. 2014. Highly efficient supersonic small infrared target detection using temporal contrast filter. *Electronics Letters*, 50(2): 81–83.

Li, B.; Xiao, C.; Wang, L.; Wang, Y.; Lin, Z.; Li, M.; An, W.; and Guo, Y. 2023. Dense Nested Attention Network for Infrared Small Target Detection. *IEEE Transactions on Image Processing*, 32: 1745–1758.

Li, R.; An, W.; Xiao, C.; Li, B.; Wang, Y.; Li, M.; and Guo, Y. 2025. Direction-coded temporal U-shape module for multiframe infrared small target detection. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1): 555–568.

Lin, J.; Li, S.; Yang, X.; Niu, S.; Yan, B.; and Meng, Z. 2024. CS-ViG-UNet: Infrared small and dim target detection based on cycle shift vision graph convolution network. *Expert Systems with Applications*, 254: 124385.

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 Computer Vision and Pattern Recognition*.

Liu, T.; Yin, Q.; Yang, J.; Wang, Y.; and An, W. 2023. Combining deep denoiser and low-rank priors for infrared small target detection. *Pattern Recognition*, 135: 109184.

Lu, Y.; Lin, Y.; Wu, H.; Xian, X.; Shi, Y.; and Lin, L. 2024. SIRST-5K: Exploring Massive Negatives Synthesis with Self-supervised Learning for Robust Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*.

Moradi, S.; Moallem, P.; and Sabahi, M. F. 2020. Fast and robust small infrared target detection using absolute directional mean difference algorithm. *Signal Processing*, 177: 107727.

Peng, S.; Ji, L.; Chen, S.; Duan, W.; and Zhu, S. 2025. Moving infrared dim and small target detection by mixed spatio-temporal encoding. *Engineering Applications of Artificial Intelligence*, 144: 110100.

Ravi, S.; and Larochelle, H. 2017. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*.

Sun, H.; Bai, J.; Yang, F.; and Bai, X. 2023. 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.

Wang, Z.; Wang, C.; Li, X.; Xia, C.; and Xu, J. 2025. MLP-Net: Multilayer Perceptron Fusion Network for Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–13.

Wu, F.; Liu, A.; Zhang, T.; Zhang, L.; Luo, J.; and Peng, Z. 2025. Saliency at the Helm: Steering Infrared Small Target Detection with Learnable Kernels. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–14.

Wu, F.; Yu, H.; Liu, A.; Luo, J.; and Peng, Z. 2023. Infrared small target detection using spatiotemporal 4-D tensor train and ring unfolding. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–22.

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.

Yan, P.; Hou, R.; Duan, X.; Yue, C.; Wang, X.; and Cao, X. 2023. STDMANet: Spatio-Temporal Differential Multiscale Attention Network for Small Moving Infrared Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–16.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9): 9202–9210.

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*.

Zeng, M.; Li, J.; and Peng, Z. 2006. The design of top-hat morphological filter and application to infrared target detection. *Infrared physics & technology*, 48(1): 67–76.

Zhang, L.; and Peng, Z. 2019. Infrared small target detection based on partial sum of the tensor nuclear norm. *Remote Sensing*, 11(4): 382.

Zhang, M.; Yang, H.; Guo, J.; Li, Y.; Gao, X.; and Zhang, J. 2024. IRPruneDet: efficient infrared small target detection via wavelet structure-regularized soft channel pruning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7224–7232.

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.

Zhang, T.; Li, L.; Cao, S.; Pu, T.; and Peng, Z. 2023. Attention-Guided Pyramid Context Networks for Detecting Infrared Small Target Under Complex Background. *IEEE Transactions on Aerospace and Electronic Systems*, 1–13.

Zhao, M.; Li, W.; Li, L.; Hu, J.; Ma, P.; and Tao, R. 2022. Single-frame infrared small-target detection: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 10(2): 87–119.

Zhu, J.; Chen, S.; Li, L.; and Ji, L. 2023. Sanet: Spatial Attention Network with Global Average Contrast Learning for Infrared Small Target Detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Zhu, S.; Ji, L.; Zhu, J.; Chen, S.; and Duan, W. 2024. TMP: Temporal motion perception with spatial auxiliary enhancement for moving infrared dim-small target detection. *Expert Systems with Applications*, 255: 124731.