Cross-domain Joint Learning with Prototype-guided Mixture-of-Experts for Infrared Moving Small Target Detection

Weiwei Duan, Luping Ji*, Jianghong Huang, Sicheng Zhu, Mao Ye

School of Computer Science and Engineering, University of Electronic Science and Technology of China, China
dww@std.uestc.edu.cn, jiluping@uestc.edu.cn, {jianghong, sichengzhu}@std.uestc.edu.cn, cvlab.uestc@gmail.com

Abstract

Infrared small target detection often faces significant domain gaps across datasets due to varying sensors and scene distributions. Currently, most existing methods are typically based on single-domain learning ($i.e.$, training and test are on the same dataset), requiring training separate detectors when considering different datasets. However, they overlook the valuable public knowledge across domains and limit the applicability in multiple infrared scenarios. To break through single-domain learning, implementing only one universal detector simultaneously on multiple datasets, as the first exploration, we propose a cross-domain joint learning task framework with prototype-guided Mixture-of-Experts (CoMoE). Specifically, it designs a hyperspherical prototype learning to adaptively maintain both domain-specific prototypes and global prototypes, enhancing cross-domain feature representation. Meanwhile, a domain-aware Mixture-of-Experts with Top-K routing strategy is proposed to assign the optimal domain experts. Moreover, to enhance cross-domain feature alignment, we design an adaptive cross-domain feature modulation with noise-guided contrastive learning. The extensive experiments on a newly constructed benchmark comprising three datasets verify the superiority of our CoMoE, even under limited data settings. It could often surpass general joint learning methods, and state-of-the-art (SOTA) single-domain ones.

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

Introduction

Infrared targets are typically small and dim, with a low signal-to-clutter ratio (Bai and Zhou 2010). Infrared small target detection (ISTD) has the unique advantages of independence from external lighting and all-weather visibility, making it highly valuable in wide and critical applications, $e.g.$, military surveillance, autonomous driving, and maritime rescue (Peng et al. 2025). The primary goal of ISTD is to accurately detect and locate small targets within complex backgrounds (Duan et al. 2025a). As a fundamental technology in computer vision, it has garnered significant research attention over the past decades (Zhu et al. 2025).

For effectively detecting small targets in infrared images, researchers have proposed various methods specifically for ISTD. They can be divided into two categories: traditional scheme and learning-based scheme. Early ISTD schemes typically utilize traditional image processing technologies, $e.g.$, filters (Deshpande et al. 1999; Bai and Zhou 2010), human visual systems (Chen et al. 2013) and data structures (Wang et al. 2021). These methods usually heavily rely on the prior knowledge of infrared images and intricate handcrafted features, lacking sample learning capabilities. Therefore, they often struggle to adapt to dynamically evolving real-world scenarios, causing missed and false detections.

Recently, with the progress of machine learning, many learning-based schemes have been proposed. Depending on the frame number to be processed, they can be further grouped into *Single-frame* and *Multi-frame* (Duan et al. 2025b). The former only utilizes the visual features of an individual image, no more information is available between adjacent frames, $e.g.$, ACM (Dai et al. 2021), DNANet (Li et al. 2022) and MSHNet (Liu et al. 2024). Recent studies have begun to explore some multi-frame detection schemes, such as infrared moving small target detection (IMSTD) (Chen et al. 2025). They often capture small target features from both visual and motion patterns, $e.g.$, ST-Trans (Tong et al. 2024) and DTUM (Li et al. 2025).

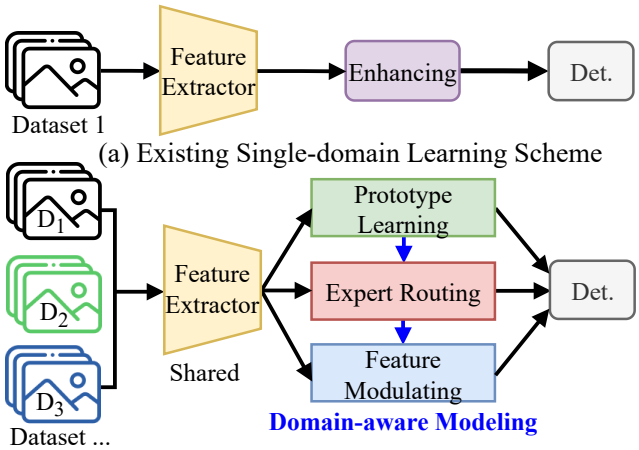

Figure 1: The comparisons between typically single-domain learning and our new cross-domain joint learning schemes.

*Corresponding Author

Totally, almost all existing learning-based methods belong to a single-domain scheme, $i.e.$, training and test are conducted on the same dataset, as shown in Figure 1 (a). This type of schemes could usually acquire good performance only if training and test have an almost-same domain distribution. However, in practical scenarios, the domain shift between training and test is often inevitable. As such, it could cause significant performance degradation when applied to different scenario domains. To address these problems above, several *Domain Adaptation* methods (Zhang et al. 2023b; Chi et al. 2024) have been presented for single-frame infrared small targets. However, they usually aim to adapt the detector trained on a labeled source domain to an unlabeled target domain via distribution alignment or self-training, which limits their applicability in multiple complex scenarios. Different from *Domain Adaptation* schemes, we propose a new task framework, the cross-domain joint learning for IMSTD, as shown in Figure 1 (b). Rather than training multiple detectors, it aims to construct a general one to detect multi-scenario infrared small targets effectively by the cross-domain joint training on multiple datasets.

In our framework, to implement cross-domain joint detection, requiring to address three critical issues. The first is how to capture domain-specific private features and domain-irrelevant public knowledge. The second is how to identify different detection domains, and the last is how to treat domain gaps. For the first issue, we propose a hyperspherical prototype learning mechanism, maximizing inter-domain distances and minimizing intra-domain distance. For the second one, concerning the effectiveness of Mixture-of-Experts (MoE) (Jacobs et al. 1991) in image classification (Riquelme et al. 2021) and multi-task learning (Li et al. 2024), we design a new domain-aware MoE to differentiate detection domains. Besides, to address the problem of domain gaps, an adaptive cross-domain feature modulation by noise-guided contrastive learning is presented in our cross-domain joint learning framework. The extensive experiments on a new benchmark comprising three datasets demonstrate the effectiveness and superiority of our scheme, even under limited data settings.

In summary, our primary contributions include: (**i**) breaking through traditional single-domain learning, the first cross-domain joint learning framework is proposed; (**ii**) rather than general feature space, a hyperspherical prototype learning mechanism is proposed to capture feature prototypes in a hypersphere space, promoting cross-domain feature representation; (**iii**) a domain-aware MoE with Top-K routing strategy is developed to assign optimal domain experts through prototypes and motion cues; (**iv**) an adaptive cross-domain feature modulation with noise-guided contrastive learning is presented to tackle domain gaps, enhancing the cross-domain consistency of infrared small targets.

Related Work

Infrared Small Target Detection

According to the number of input frames, ISTD can be classified into single-frame and multi-frame schemes (Duan et al. 2024). Since single-frame methods (Li et al. 2022; Zhang et al. 2023a; Liu et al. 2024; Wang et al. 2025) have no available information between adjacent frames, rendering them ineffective in challenging video scenes. For multi-frame ones (Tong et al. 2024; Zhu et al. 2024; Chen et al. 2024; Duan et al. 2024; Li et al. 2025), they often model the spatial-temporal features to promote detection accuracy. For example, ST-Trans (Tong et al. 2024) uses a spatial-temporal transformer to extract motion dependencies between successive frames. Recently, DTUM (Li et al. 2025) uses a direction-coded temporal U-shape module and a direction-coded convolution block to encode the motion direction of targets. However, almost all current methods are conducted within single-domain learning, requiring training multiple detectors when concerning different datasets.

Cross-domain Joint Learning

Cross-domain joint learning has emerged as a promising approach, as demonstrated in general object detection (Chen et al. 2023; Jain et al. 2023; Wang et al. 2024). It aims to leverage multiple datasets from various domains in training, enabling the model to process multi-domain data during the inference stage. For example, Plain-Det (Shi, Zhu, and Yang 2024) enhances the emergent property by utilizing few-iteration, dataset-specific training to address the challenges of multi-dataset object detection. However, substantial differences inherently exist in infrared images themselves due to the non-uniformity of different infrared sensors, making multi-domain joint learning more challenging.

Mixture-of-Experts

MoE (Jacobs et al. 1991) is an ensemble learning method that utilizes multiple experts to collaboratively solve a task. It contains a gate routing mechanism that selectively activates optimal experts based on inputs. Recently, MoE has been widely applied to various fields, e.g., large language models (Cai et al. 2025) and computer vision (Jain et al. 2023; Yang et al. 2025). Unlike prior methods, we design a new domain-aware MoE with prototypes to assist the model in identifying different detection domains for IMSTD.

Methodology

Overall Architecture

Problem Formulation. In our task, the cross-domain ($i.e.$, multi-dataset) data $\boldsymbol{\mathcal{D}} = \{\boldsymbol{\mathcal{D}_i} = (\boldsymbol{I^i}, \boldsymbol{y^i})\}_{i=1}^{m}$ with heterogeneous domain distributions is used for training. $\boldsymbol{I^i} = \{I_1^i, I_2^i, \cdots, I_t^i\}$ is a collection of consecutive frames, randomly sampled from an infrared video with a time window of t. Our primary goal is to train a unified detector $f_\theta(\cdot)$ across multiple domains to predict the bounding boxes of targets y_t^i in the keyframe I_t^i by using its adjacent frames. It can effectively utilize cross-domain training samples to optimize the overall loss, that is

$$\min_{\theta} \sum_{i=1}^{m} \mathbb{E}_{(\boldsymbol{I^i}, y_t^i) \sim \boldsymbol{\mathcal{D}_i}} [\mathcal{L}(f_\theta(\boldsymbol{I^i}), y_t^i)], \tag{1}$$

where $\mathcal{L}$ denotes a training loss, and $\mathbb{E}_{(\boldsymbol{I^i}, y_t^i) \sim \boldsymbol{\mathcal{D}_i}}$ is the expectation over all samples from $\boldsymbol{\mathcal{D}_i}$.

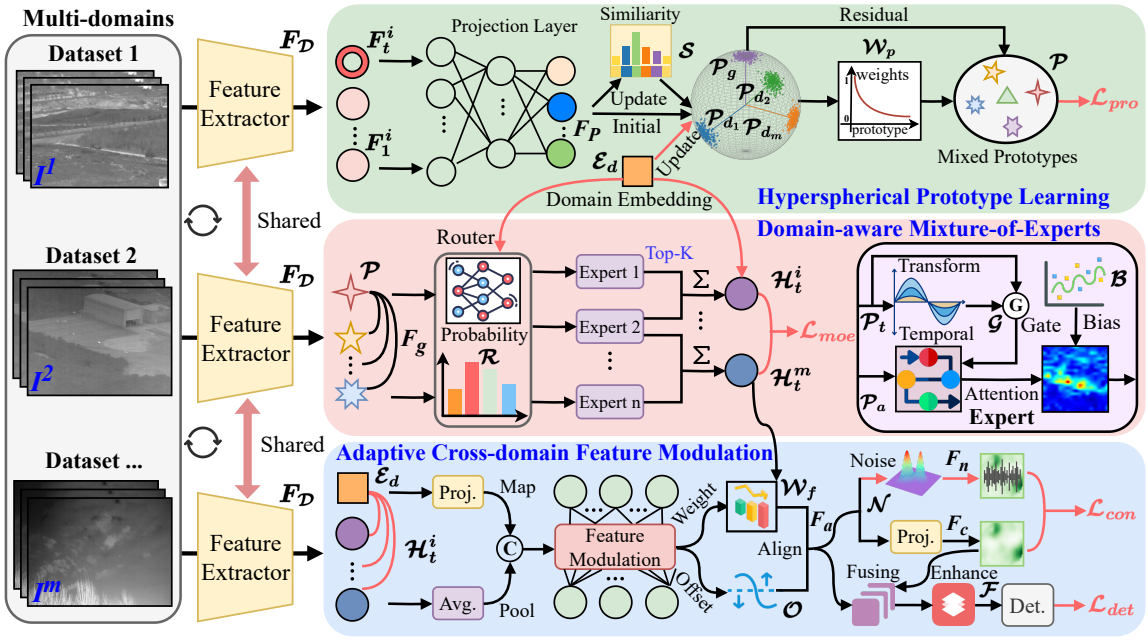

Figure 2: Our CoMoE framework with three parts. **Hyperspherical Prototype Learning** is proposed to capture domain-specific feature prototypes and public features. **Domain-aware Mixture-of-Experts** is designed to identify domains and assign optimal domain experts. **Adaptive Cross-domain Feature Modulation** is presented to address possible infrared target domain gaps, thereby enhancing cross-domain consistency and mitigating domain discrepancies. Red lines are only for training.

Overview. Our core insight is that we could train a cross-domain detector in the same way as training multiple single-domain detectors separately, as long as we effectively mitigate the domain gaps. Therefore, we propose a cross-domain joint learning task framework, *i.e.*, CoMoE, in Figure 2.

In detail, it takes video clips $\{\boldsymbol{I}^1, \boldsymbol{I}^2, \cdots, \boldsymbol{I}^m\}$ from m different domains as inputs. Each $\boldsymbol{I}^i$ contains t consecutive frames. The pretrained CSPDarknet (Ge et al. 2021) is used as the shared feature extractor. Following most video object detection methods (Zhou et al. 2022), we extract multi-frame features $\boldsymbol{F_D} = \{\boldsymbol{F_1^i}, \cdots, \boldsymbol{F_t^i}\}_{i=1}^{m} \in \mathbb{R}^{m\times t\times C\times H\times W}$ by iteratively feeding each frame into the feature extractor, where C, H and W are the channel, height and weight, respectively. Then, domain-specific prototypes $\boldsymbol{\mathcal{P}_d}$ and global prototypes $\boldsymbol{\mathcal{P}_g}$ are obtained by projecting $\boldsymbol{F_D}$ into a hyperspherical subspace. We can extract mixed prototypes $\boldsymbol{\mathcal{P}}$ by integrating domain embedding $\boldsymbol{\mathcal{E}_d}$ (*i.e.*, the encoded domain labels for each dataset) in **HPL**. Besides, mixed prototypes $\boldsymbol{\mathcal{P}}$ are further processed in **DMoE** to capture motion features $\boldsymbol{\mathcal{H}_t}$ by using the target motion consistency of different domains. Moreover, we propose **ACFM** to mitigate domain gaps. Finally, the aligned and enhanced features $\boldsymbol{\mathcal{F}}$ are used for detection to obtain final results.

Hyperspherical Prototype Learning

Unlike traditional feature spaces, we design a new prototype learning mechanism to constrain features within a normalized hypersphere, as shown in Figure 2.

First, multi-frame features $\boldsymbol{F_D}$ are mapped to a hypersphere space to initialize domain prototypes $\boldsymbol{\mathcal{P}_d}$ and global prototypes $\boldsymbol{\mathcal{P}_g}$, as follows:

$$\boldsymbol{\mathcal{P}_d}, \boldsymbol{\mathcal{P}_g} = f_{nor}(\boldsymbol{F_P}) = f_{nor}(f_{mlp}(\sum_{i=1}^{m}\sum_{j=1}^{t}\boldsymbol{F_j^i})), \tag{2}$$

where $f_{mlp}(\cdot)$ denotes two linear layers with GELU activation functions and $f_{nor}(\cdot)$ is the normalization that ensures features can be projected onto a unit hypersphere. m is the total number of domains, and t is the time window size of video clips. Then, in the next iteration, prototypes are updated by calculating the similarity $\boldsymbol{\mathcal{S}}$ between new prototypes $\boldsymbol{\mathcal{P}_n}$ and old prototypes $\boldsymbol{\mathcal{P}_d}$, that is

$$\begin{cases} \boldsymbol{\mathcal{S}} = f_{cos}(\boldsymbol{\mathcal{P}_n}, \boldsymbol{\mathcal{P}_{d_i}}) = \sum_{i=1}^{m} \dfrac{\boldsymbol{\mathcal{P}_n} \bullet \boldsymbol{\mathcal{P}_{d_i}}}{\|\boldsymbol{\mathcal{P}_n}\|_2 \, \|\boldsymbol{\mathcal{P}_{d_i}}\|_2}, \\ \boldsymbol{\mathcal{P}'_{d_i}} = \mu_i \cdot \boldsymbol{\mathcal{P}_{d_i}} + (1-\mu_i)\cdot \boldsymbol{\mathcal{P}_n} + \boldsymbol{\mathcal{E}_d}, \\ i = \arg\max_k \boldsymbol{\mathcal{S}}(\boldsymbol{\mathcal{P}_n}, k), \ \mu_i = \mu_0^{1+0.01\times C_{d_i}}, \end{cases} \tag{3}$$

where $f_{cos}(\cdot)$ denotes calculating cosine similarity, "$\bullet$" means inner product, $\|\cdot\|_2$ represents l_2 norm, $\mu_i \in (0,1)$ is a momentum coefficient that decays with update count C_{d_i} to ensure stability, $\mu_0 = 0.95$, i is the domain index most similar to $\boldsymbol{\mathcal{P}_n}$, and $\boldsymbol{\mathcal{P}'_{d_i}}$ is the updated domain prototypes.

Second, the global prototypes are updated, that is

$$\boldsymbol{\mathcal{P}'_g} = f_{nor}(\frac{1}{m}\sum_{i=1}^{m}(\mu \cdot \boldsymbol{\mathcal{P}_{d_i}})) + \boldsymbol{\mathcal{P}_g}, \tag{4}$$

where $\boldsymbol{\mathcal{P}'_g}$ is the updated global prototypes. Then, to bridge domain-specific and domain-irrelevant representations, we

integrate domain prototypes $\mathcal{P}'_{\boldsymbol{d}}$ with global prototypes $\mathcal{P}'_{\boldsymbol{g}}$ to obtain mixed prototypes $\mathcal{P}$, as follows:

$$\begin{cases} \mathcal{P}_{\boldsymbol{m}} = \alpha \cdot \mathcal{P}'_{\boldsymbol{d}} + \beta \cdot \mathcal{P}'_{\boldsymbol{g}}, \\ \mathcal{P} = \mathcal{P}_{\boldsymbol{m}} + \sum_{i=1}^{m} \mathcal{W}_{\boldsymbol{p_i}} \mathcal{P}_{\boldsymbol{m_i}}, \\ \mathcal{W}_{\boldsymbol{p}} = Softmax(f_{mlp}(\mathcal{P}_{\boldsymbol{m}})), \end{cases} \tag{5}$$

where $\mathcal{W}_{\boldsymbol{p}}$ is the weights calculated by a softmax function over the intermediate results $\mathcal{P}_{\boldsymbol{m}}$ to aggregate prototypes.

Finally, we develop a prototype-level supervision loss $\mathcal{L}_{pro}$ to constrain features to be close to the most relevant domain prototypes, thereby enhancing cross-domain semantic consistency. It can be formulated as follows:

$$\mathcal{L}_{pro} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{p=1}^{P} \tilde{y}_{ip} \log \left(\frac{\exp(\mathcal{S}_{ip})}{\sum_{j=1}^{N} \exp(\mathcal{S}_{ij})} \right), \tag{6}$$

where N is the total number of projected feature samples, P is the total number of domain prototypes, $\mathcal{S}_{ip}$ is the similarity between the i-th features and the p-th prototypes, and $\tilde{y}_{ip}$ is the probability belonging to the p-th domain prototypes.

Domain-aware Mixture-of-Experts

Developing a feature processing strategy that is adaptable to multiple domains is a key challenge in our task. To address this, we present DMoE to dynamically assign the optimal experts for different domains, enhancing the detection performance of each domain, as shown in Figure 2.

First, a domain-aware router is designed to dynamically identify different detection domains. For each mixed prototype, it can jointly consider the information of keyframe and adjacent frames as well as domain embedding $\mathcal{E}_{\boldsymbol{d}}$, that is

$$\begin{cases} \boldsymbol{F_g} = f_{cat}(f_{avg}(\sum_{i=1}^{m} \sum_{j=1}^{t} \mathcal{P}_{\boldsymbol{j}}^{\boldsymbol{i}}), \mathcal{E}_{\boldsymbol{d}}), \\ \mathcal{R} = Softmax(f_{mlp}(\boldsymbol{F_g})) \in \mathbb{R}^n, \end{cases} \tag{7}$$

where $f_{cat}(\cdot)$ denotes the concatenation operation, $f_{avg}(\cdot)$ means the average pool, $\boldsymbol{F_g}$ is the global features with domain embeddings, n is the total number of experts, and $\mathcal{R}$ is the probability that each expert is selected. Besides, a Top-K routing strategy is present to prioritize the optimal experts based on the computed probabilities $\mathcal{R}$ to obtain the motion features $\mathcal{H}_{\boldsymbol{t}}$ across domains, as follows:

$$\begin{cases} \mathcal{I} = \text{TopK}(\mathcal{R}, \text{K}) \in \mathbb{R}^{\text{K}}, \\ \tilde{r_i} = \begin{cases} r_i / \sum_{j \in \mathcal{I}} r_j, & \text{if } i \in \mathcal{I}, \\ 0, & \text{otherwise}, \end{cases} \\ \mathcal{H}_{\boldsymbol{t}} = \sum_{i \in \mathcal{I}} \tilde{r_i} \cdot f_{exp}(\mathcal{P}, \mathcal{E}_{\boldsymbol{d}}), \end{cases} \tag{8}$$

where $\mathcal{I}$ is the index of selected experts, $\tilde{r_i}$ is the weights by normalizing probabilities, and $f_{exp}(\cdot)$ is the expert network.

Second, for each expert, we perform feature transformation $f_{trans}(\cdot)$ to extract stable spatial semantics, and temporal fusion $f_{tem}(\cdot)$ to capture motion cues $\boldsymbol{h^i}$, as follows:

$$\begin{cases} \hat{\mathcal{P}^{\boldsymbol{i}}} = \mathcal{P}_{\boldsymbol{t}}^{\boldsymbol{i}} + \sigma(\mathcal{G}(f_{trans}(\mathcal{P}_{\boldsymbol{a}}^{\boldsymbol{i}}))) \odot f_{tem}(\mathcal{P}_{\boldsymbol{t}}^{\boldsymbol{i}}, \mathcal{P}_{\boldsymbol{a}}^{\boldsymbol{i}}) \\ \boldsymbol{h^i} = CrossAtt(\hat{\mathcal{P}^{\boldsymbol{i}}}) + \mathcal{B}, \ \mathcal{B} = f_{mlp}(\mathcal{E}_{\boldsymbol{d}}), \end{cases} \tag{9}$$

where $\mathcal{P}_{\boldsymbol{t}}^{\boldsymbol{i}}$ is the mixed prototypes of keyframe in i-th domain, $\mathcal{P}_{\boldsymbol{a}}$ represents the ones of adjacent frames, σ is the sigmoid function, $\mathcal{G}$ is a gate mechanism with an average pool and 1×1 convolutions, $f_{trans}(\cdot)$ is a residual block comprising two 3×3 convolutions, and $f_{tem}(\cdot)$ denotes a two-step temporal fusion method. It first computes attention-guided weights based on the sum of keyframe and adjacent prototypes, and then iteratively refines motion features. $\mathcal{B}$ is the domain bias, and $CrossAtt(\cdot)$ is a cross attention.

Finally, to promote balanced utilization of different experts and obtain diversified features, the load balancing loss $\mathcal{L}_{ban}$ and diversity loss $\mathcal{L}_{div}$ are designed as follows:

$$\begin{aligned} \mathcal{L}_{ban} &= \|\mathbb{E}(\mathcal{R}) - \frac{1}{n}\|_{mse} = \|\frac{1}{n} \sum_{i=1}^{n} \mathcal{R}_{\boldsymbol{i}} - \frac{1}{n}\|_{mse}, \\ \mathcal{L}_{div} &= \frac{\text{K}}{n(n-1)} \sum_{1 \le i \le j \le n} \frac{1}{B} \sum_{b=1}^{B} f_{cos}(\boldsymbol{h_b^i}, \boldsymbol{h_b^j}), \end{aligned} \tag{10}$$

where n and K is the total number of experts and selected ones, respectively. B is the batch size, and $\| \cdot \|_{mse}$ denotes the MSE loss. Therefore, the overall loss of DMoE could be denoted by $\mathcal{L}_{moe} = \mathcal{L}_{ban} + \mathcal{L}_{div}$.

Adaptive Cross-domain Feature Modulation

To achieve feature alignment and alleviate domain gaps, we propose ACFM to adaptively adjust target responses and enhance cross-domain consistency, as shown in Figure 2.

Specifically, the aligned features $\boldsymbol{F_a}$ are obtained by integrating the motion features $\mathcal{H}_{\boldsymbol{t}}^{\boldsymbol{i}}$ and domain embedding $\mathcal{E}_{\boldsymbol{d}}$, performing domain-aware feature modulating, as follows:

$$\begin{cases} \mathcal{W_f}, \mathcal{O} = f_{mlp}(f_{cat}(f_{avg}(\mathcal{H}_{\boldsymbol{t}}^{\boldsymbol{i}}), \mathcal{E}_{\boldsymbol{d}})), \\ \boldsymbol{F_m} = \mathcal{W_f} \odot \mathcal{H}_{\boldsymbol{t}}^{\boldsymbol{i}} + \mathcal{O}, \\ \boldsymbol{F_a} = \boldsymbol{F_m} + \mathcal{A_s}(\boldsymbol{F_m} \cdot \mathcal{A_c}(\boldsymbol{F_m})), \end{cases} \tag{11}$$

where $\mathcal{W_f}$ denotes modulation weights, $\mathcal{O}$ is domain offsets, $\boldsymbol{F_m}$ represents the modulated features. $\mathcal{A_s}$ and $\mathcal{A_c}$ denotes the spatial and channel attention, respectively. Then, we design the noise-guided contrastive learning to inject domain-specific noise $\epsilon_i \sim \mathcal{N}(0, \gamma_i^2)$ into aligned features $\boldsymbol{F_a}$, keeping the consistency between clean and noisy representations, $i.e.$, $\boldsymbol{F_{n_i}}$ and $\boldsymbol{F_{c_i}}$, as follows:

$$\begin{cases} \hat{\boldsymbol{F_{a_i}}} = \boldsymbol{F_{a_i}} + \epsilon_i, \ \epsilon_i \sim \mathcal{N}(0, \gamma_i^2) \\ \boldsymbol{F_{n_i}} = f_{avg}(f_{mlp}(\hat{\boldsymbol{F_{a_i}}})), \boldsymbol{F_{c_i}} = f_{avg}(f_{mlp}(\boldsymbol{F_{a_i}})) \\ \mathcal{L}_{con} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{f_{cos}(\boldsymbol{F_{n_i}}, \boldsymbol{F_{c_i}})}{\sum_{j=1}^{B} f_{cos}(\boldsymbol{F_{n_i}}, \boldsymbol{F_{c_j}})}, \end{cases} \tag{12}$$

where γ_i is a learnable parameter to control the noise distribution of i-th domain. $\mathcal{L}_{con}$ is a contrastive loss to facilitate preserving the cross-domain consistency under disturbance.

	Method	Pub	DAUB-H				ITSDT-15K				IRDST-R			
			mAP_{50}(%)	Pr (%)	Re (%)	F1 (%)	mAP_{50}(%)	Pr (%)	Re (%)	F1 (%)	mAP_{50}(%)	Pr (%)	Re (%)	F1 (%)
Single-domain	ACM	WACV'21	49.84	74.83	67.21	70.81	55.38	78.37	71.69	74.88	57.65	79.34	73.30	76.20
	ISNet	CVPR'22	44.73	56.35	80.99	66.46	62.29	83.46	75.32	79.18	59.27	73.75	81.98	77.65
	UIUNet	TIP'22	49.23	73.72	67.49	70.47	65.15	84.07	78.39	81.13	59.22	76.55	78.47	77.50
	AGPCNet	TAES'23	22.89	37.86	61.67	46.91	67.27	91.19	74.77	82.16	58.01	79.76	73.91	76.72
	DNANet	TIP'23	50.76	71.04	72.01	71.52	70.46	88.55	80.73	84.46	67.43	85.97	79.04	82.36
	RPCANet	WACV'24	52.57	70.74	75.16	72.88	62.28	81.46	77.10	79.22	65.63	81.62	81.16	81.39
	SIRST5K	TGRS'24	40.66	57.63	71.14	63.68	61.52	86.95	71.32	78.36	47.43	71.08	67.03	68.99
	SCTrans	TGRS'24	34.74	54.52	64.89	59.25	71.37	**91.74**	78.49	84.60	51.75	73.13	72.00	72.56
	ST-Trans	TGRS'24	44.93	**87.21**	52.54	65.57	76.02	89.96	85.18	87.50	64.50	**86.84**	75.16	80.58
	SSTNet	TGRS'24	52.25	83.49	63.25	71.98	76.96	91.05	85.29	88.07	68.21	75.70	**91.34**	82.79
	MSHNet	CVPR'24	26.47	48.56	55.39	51.75	60.82	89.69	68.44	77.64	60.86	78.54	78.52	75.53
	DTUM	TNNLS'25	50.32	81.53	63.00	71.08	67.97	77.95	88.28	82.79	64.72	83.32	78.81	81.00
	MLPNet	TGRS'25	49.17	68.00	73.13	70.47	53.76	74.06	73.19	73.63	50.20	75.10	67.34	71.01
	LSKNet	TGRS'25	46.34	56.97	**82.76**	67.48	66.07	90.75	73.72	81.35	65.84	82.66	80.88	81.76
	PConv	AAAI'25	48.48	70.74	69.14	69.93	59.85	87.11	69.43	77.27	58.45	79.50	74.74	77.04
Multi-domain	DAMEX	NeurIPS'23	30.15	52.34	60.21	56.00	46.78	58.47	65.19	61.65	43.82	60.73	56.33	58.45
	PlainDet	ECCV'24	25.37	48.72	55.79	52.02	38.39	52.38	48.27	50.24	32.45	49.18	47.94	48.55
	UniDet	IJCV'24	33.69	54.68	64.26	59.08	42.15	60.34	54.26	57.14	40.56	58.79	53.17	55.84
	SSTNet★	TGRS'24	48.39↓ 3.86	85.04↑ 1.55	56.01↓ 7.24	67.54↓ 4.44	67.44↓ 9.52	90.33↓ 0.72	75.33↓ 9.96	82.15↓ 5.92	56.38↓ 11.83	74.57↓ 1.13	76.41↓ 14.93	75.48↓ 7.31
	DTUM★	TNNLS'25	39.94↓ 10.38	54.46↓ 27.07	74.33↑ 11.33	62.86↓ 8.22	47.50↓ 20.47	67.38↓ 10.57	71.10↓ 17.18	69.19↓ 13.60	61.77↓ 2.95	74.13↓ 9.19	84.42↑ 5.61	78.94↓ 2.06
	CoMoE	-	**53.84**	71.15	75.22	**73.13**	**78.19**	85.77	**92.78**	**89.14**	**69.47**	80.15	88.28	**84.02**

Table 1: Quantitative comparisons. The best one is marked in **bold**, and the second-best one is underlined. All multi-domain methods utilize all samples from three datasets to train a universal model, and then evaluate it separately on each dataset. "★" means single-domain methods for multi-domain training. "↓" means a decline on single-domain, while "↑" means a gain.

After that, final features $\boldsymbol{\mathcal{F}}$ is obtained by integrating the aligned features $\boldsymbol{F_a}$ and clean representations $\boldsymbol{F_c}$, that is

$$\boldsymbol{\mathcal{F}} = f_{non}(\boldsymbol{\mathcal{A}_s}(\boldsymbol{\mathcal{A}_c}(Conv(f_{cat}(\boldsymbol{F_a}, \boldsymbol{F_c}))))) \tag{13}$$

where $f_{non}(\cdot)$ denotes the nonlocal attention (Zhang et al. 2023a), especially for small targets. Finally, $\boldsymbol{\mathcal{F}}$ is fed into the detection head to obtain final results. Therefore, the total training loss of our CoMoE could be defined as follows:

$$\mathcal{L} = \mathcal{L}_{pro} + \mathcal{L}_{moe} + \mathcal{L}_{con} + \mathcal{L}_{det}, \tag{14}$$

where $\mathcal{L}_{det}$ is the detection loss based on the decoupled detection head of YOLOX (Ge et al. 2021).

Experiments

Implement Details

We evaluate our CoMoE on a new benchmark comprising three datasets: DAUB-H (Hui et al. 2019), ITSDT-15K (Duan et al. 2024) and IRDST-R (Sun et al. 2023). Following previous works (Chen et al. 2024), we use standard evaluation metrics, *i.e.*, Precision (Pr), Recall (Re), F1 and mAP_{50} (the mean Average Precision with an IoU threshold 0.5). Moreover, the input frames of all compared methods are resized to 512×512. In detail, our CoMoE and compared methods are trained for 100 epochs with a batch size of 4. SGD is adopted as the optimizer with an initial learning rate of 0.01 and a weight decay of 5×10^{-4}. Hyper-parameters t, K, P, n, α, and β are set to 5, 2, 32, 6, 0.6, and 0.4, respectively.

Comparisons with SOTA Methods

Quantitative Comparisons Table 1 presents the quantitative comparisons on recent single-domain learning methods and multi-domain joint learning ones, revealing **two** obvious findings. **One** is that our CoMoE consistently achieves the highest performance, establishing new SOTA across most metrics. For example, on ITSDT-15K, CoMoE obtains the highest mAP_{50} 78.19%, Re 92.78% and F1 89.14%. Only in terms of Pr, the 85.77% by CoMoE is slightly lower than the SOTA 91.74% by SCTrans (Yuan et al. 2024). SCTrans achieves a higher Pr at the expense of Re, while our method achieves a more balanced performance and an higher F1.

The other is that general multi-domain joint learning is ineffective in highly challenging scenarios. For instance, on DAUB-H, the SOTA multi-domain one, *i.e.*, UniDet (Lin et al. 2024) only achieves mAP_{50} 33.69% and F1 59.08%, significantly lower than the mAP_{50} 53.84% and F1 73.13% by our CoMoE. Besides, using single-domain methods directly for multi-domain training is invalid. For example, DTUM★ decreases 10.38% on mAP_{50} and 8.22% on F1.

Methods	Frames	mAP_{50}	F1	Params ↓	GFlops ↓	FPS ↑
ACM	1	55.38	74.88	3.04M	**24.73**	**29.11**
ISNet	1	62.29	79.18	3.49M	265.73	11.20
UIUNet	1	65.15	81.13	53.06M	456.70	3.63
AGPCNet	1	67.27	82.16	14.88M	366.15	4.79
RDIAN	1	68.49	82.68	**2.74M**	50.44	20.52
DNANet	1	70.46	84.46	7.22M	135.24	4.82
SIRST5K	1	61.52	78.36	11.48M	182.61	7.37
MSHNet	1	60.82	77.64	6.59M	69.59	18.55
MLPNet	1	53.76	73.63	10.79M	34.72	5.93
LSKNet	1	66.07	81.35	3.42M	76.00	40.63
PConv	1	59.85	77.27	6.59M	69.29	10.01
ST-Trans	5	76.02	87.50	38.13M	145.16	3.90
SSTNet	5	76.96	88.07	11.95M	123.59	9.24
DTUM	5	67.97	82.79	9.64M	128.16	14.28
DAMEX	1	46.78	61.65	46.74M	368.73	5.54
UniDet	1	42.15	57.14	51.38M	417.82	3.32
CoMoE (Ours)	5	**78.19**	**89.14**	19.61M	322.39	12.73

Table 2: The inference cost comparisons on ITSDT-15K.

Inference Cost Comparisons The inference cost comparisons are presented in Table 2. From it, **two** notable obser-

Settings	HPL			DMoE			ACFM		DAUB-H				ITSDT-15K				IRDST-R			
	H1	H2	H3	D1	D2	D3	A1	A2	mAP_{50}	Pr	Re	F1	mAP_{50}	Pr	Re	F1	mAP_{50}	Pr	Re	F1
w/o All	-	-	-	-	-	-	-	-	15.49	52.14	30.34	38.36	36.85	78.00	47.83	59.30	30.67	63.88	58.39	61.01
w H	✓	-	-	-	-	-	-	-	24.39	38.27	50.93	43.70	58.06	72.11	80.84	76.23	49.68	65.78	74.36	69.81
	-	✓	-	-	-	-	-	-	28.94	44.63	49.27	46.84	57.38	70.23	81.13	75.29	48.43	63.64	75.46	69.05
	-	-	✓	-	-	-	-	-	36.87	54.38	60.24	57.16	62.89	76.39	80.18	78.24	57.40	69.70	83.78	76.10
w H & D	-	-	✓	✓	-	-	-	-	39.15	56.89	62.53	59.58	64.17	72.74	89.61	80.30	61.67	74.09	84.88	79.12
	-	-	✓	✓	✓	-	-	-	42.03	59.26	66.84	62.82	69.38	80.16	87.94	83.87	63.49	75.13	88.34	81.20
	-	-	✓	✓	✓	✓	-	-	45.86	63.57	70.92	67.05	74.29	81.51	92.00	86.44	66.85	75.39	**90.27**	82.16
w H & D & M	-	-	✓	✓	✓	✓	✓	-	50.72	67.83	74.65	71.08	77.30	84.78	92.70	88.56	68.27	78.36	88.49	83.12
	-	-	✓	✓	✓	✓	-	✓	50.35	67.31	74.88	70.90	76.44	83.85	92.33	87.88	67.64	77.14	89.47	82.85
w All	-	-	✓	✓	✓	✓	✓	✓	**53.84**	**71.15**	**75.22**	**73.13**	**78.19**	**85.77**	**92.78**	**89.14**	**69.47**	**80.15**	88.28	**84.02**

Table 3: Ablation study on CoMoE with different settings. **HPL**: three prototype learning schemes (**H1** only uses domain-specific prototypes, **H2** only uses global prototypes, **H3** uses mixed prototypes). **DMoE**: three components of domain-aware MoE (**D1** is temporal modeling, **D2** denotes domain-aware gating, **D3** is Top-K routing strategy). **ACFM**: two components of adaptive cross-domain feature modulation (**A1** is feature alignment, **A2** is the noise-guided contrast learning with $\mathcal{L}_{con}$).

vations emerge. **One** is that our Params and GFlops are increasing slightly due to domain-aware modeling. For example, our CoMoE has 19.61M parameters, higher than the SOTA method RDIAN (2.74M), but still lower than the 53.06M by UIUNet (Wu, Hong, and Chanussot 2022) and the 46.74M by DAMEX. Besides, its GFlops is 322.39, greatly lower than the 366.15 by AGPCNet (Zhang et al. 2023a). **The other** is that our CoMoE achieves an middle inference speed, despite the increased parameters. For instance, it has an FPS of 12.73, higher than those by many single-domain ones, $e.g.$, PConv and DTUM.

PR Curve Comparisons As usual, we employ precision-recall (PR) curves on DAUB and ITSDT-15K to visually assess the overall performance of various methods, as shown in Figure 3. From it, it is obvious that our curves outperform those of compared methods. Specifically, on DAUB-H, our curve consistently reaches the top-right positions. This pattern continues on ITSDT-15K. The closer a method is to the top-right corner, the higher its validity. Therefore, these PR curves highlight the superiority of our CoMoE in balancing precision and recall compared to other methods.

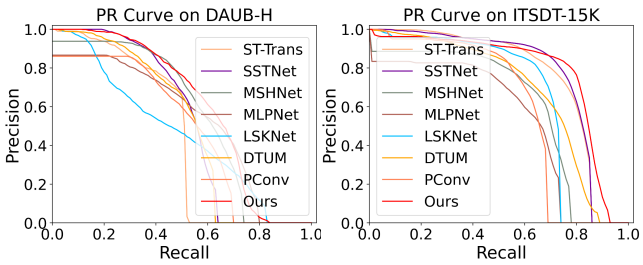

Figure 3: PR curve comparisons on two datasets.

Ablation and Analysis

Effects of Different Assemblies To investigate the impact of each component on CoMoE, we conduct a series of ablation studies on three datasets, as shown in Table 3. By comparison, we could have two apparent findings. One is that each component in CoMoE contributes to the performance improvement. For example, on IRDST-R, the baseline setting without any specialized components (w/o All) only achieves mAP_{50} 30.67% and F1 61.01%. After assembling HPL (w H3), these metrics rise to 57.40% for mAP_{50} and 76.10% for F1. Similarly, incrementally applying domain-aware gating (w H3 & D1 & D2) increases the mAP_{50} from 61.67% to 63.49%, and the F1 from 79.12% to 81.20%. The other is that when all components are fully combined (w All), performance improves remarkably, with an mAP_{50} of 69.47% and an F1 of 84.02%, achieving the highest level. It verifies that these components have collaborative effects, and each individual is effective.

Effects of Cross-domain Joint Learning To validate the effectiveness of our cross-domain joint learning, a group of experiments is conducted. As shown in Table 4, it is obvious that our cross-domain joint learning framework could simultaneously enhance detection performance across multiple training domains. For example, on IRDST-R, after integrating ITSDT-15K for joint training, mAP_{50} rises from 65.52% to 67.09%, and F1 from 81.97% to 82.96%. When all three datasets are combined, the performance peaks, with mAP_{50} 69.47% and F1 84.02%. Besides, the model often exhibits a noticeable decline on unseen domains, further verifying the significant domain gaps in IMSTD. These results indicate that our proposed cross-domain joint learning is a new paradigm worth exploring. It could simultaneously achieve superior performance across multiple domains.

Training	DAUB-H (D)		ITSDT-15K (T)		IRDST-R (R)	
	mAP_{50}	F1	mAP_{50}	F1	mAP_{50}	F1
D	50.39	67.01	0.77	6.53	0.13	2.08
D+T	51.45	64.44	74.93	85.34	3.45	20.56
D+T+R	**53.84**	**73.13**	**78.19**	**89.14**	**69.47**	**84.02**
T+R	1.15	3.29	76.08	86.58	67.89	82.96
R	0.02	0.41	3.84	11.92	65.52	81.97

Table 4: The performance of our CoMoE, as the number of datasets gradually increases. "D" → "D+T+R" is the forward order, and "R" → "D+T+R" is reverse.

Effects of Limited Data Setting To analyze the performance of our CoMoE under limited data, we randomly select different shots from each domain to conduct a group of

experiments, as shown in Figure 4. From this, it is obvious that our scheme consistently achieves peak performance under different settings. Besides, it requires only 1000 shots to perform comparably to the full dataset setting. This demonstrates the superiority of our CoMoE in alleviating the issue of limited training data in real-world applications.

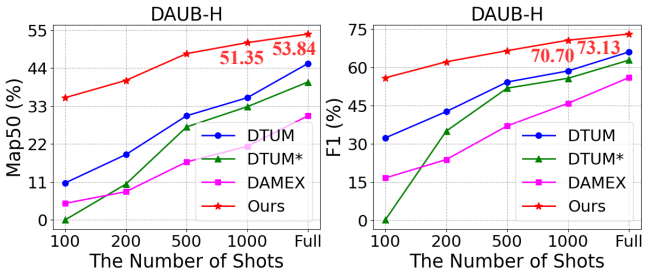

Figure 4: The ablation study of different limited data settings (100, 200, 500 and 1000 shots) on DAUB-H.

Effects of Hyperspherical Prototype Learning To visually verify the effectiveness of HPL, we compare the feature distributions before and after HPL, as shown in Figure 5. From this, it could be observed that our method efficiently captures domain-specific prototypes and the public domain-irrelevant knowledge ($i.e.$, global prototypes) across multiple domains. These comparisons further validate the quantitative results in Table 3 (w/o All and w H3).

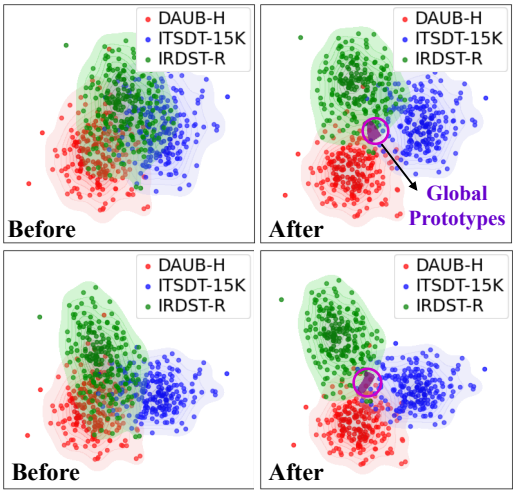

Figure 5: Feature distributions before and after HPL.

Effects of Domain-aware MoE To thoroughly analyze the impacts of DMoE, we select two samples from each dataset to visualize the expert utilization probabilities (with and without domain-aware routing), as shown in Figure 6. From it, we can clearly see that the samples from different domains are processed by different experts after integrating domain-aware routing. It indicates that DMoE could identify different detection domains and allocate the optimal experts for each domain. These results also prove the numerical results in Table 3 (w H3 and w H3 & D1 & D2 & D3).

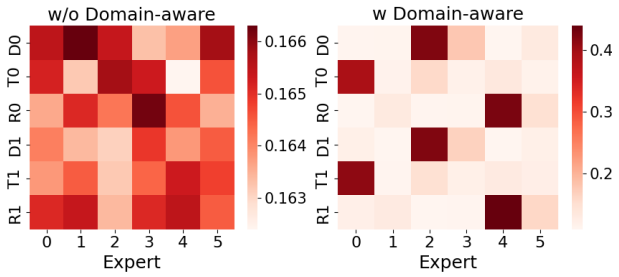

Figure 6: The visualizations of expert usage probabilities.

Effects of Cross-domain Feature Modulation To visually analyze the effects of ACFM, we present four groups of feature heatmaps before and after ACFM, as shown in Figure 7. From this, in all heatmap groups, it is evident that the focus positions before ACFM are unclear, resulting in targets being lost in complex backgrounds and domain disturbances. Conversely, after integrating ACFM, the feature response of small targets is notably enhanced, and the noisy background is clearer. It indicates that feature modulation and noise-guided contrastive learning could effectively mitigate domain gaps, thereby achieving feature alignment.

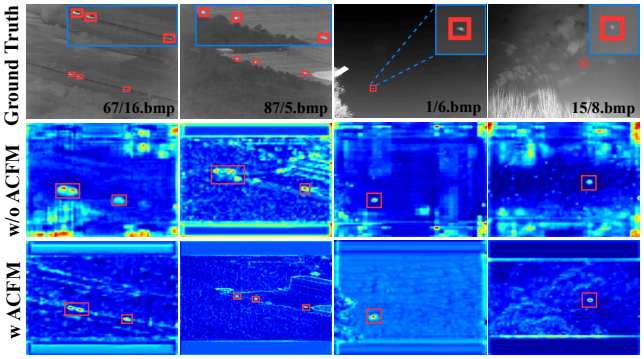

Figure 7: The feature heatmap comparisons before (w/o) and after (w) ACFM. The first two columns are on ITSDT-15K, and the last two are on IRDST-R.

Conclusions

To overcome the weakness of single-domain learning, this paper proposes the first cross-domain joint learning task framework with prototype-guided MoE for infrared small target detection ($i.e.$, CoMoE). Rather than traditional one detector one domain, it constructs a universal detector with domain-aware modeling by hyper-spherical domain prototype learning, MoE routing, and cross-domain feature modulating. The experiments on a new cross-domain benchmark verify the effectiveness and superiority of our CoMoE, even under limited data settings. On primary metrics, it could often evidently surpass current SOTA methods. One of its main drawbacks is the low generalization to the unseen detection domains of infrared small targets. In the future, an efficient cross-domain joint learning scheme with better domain generalization merits further exploration.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No.62476049 and 62276048.

References

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.

Cai, W.; Jiang, J.; Wang, F.; Tang, J.; Kim, S.; and Huang, J. 2025. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*.

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. 2025. 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. 2024. 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, Y.; Wang, M.; Mittal, A.; Xu, Z.; Favaro, P.; Tighe, J.; and Modolo, D. 2023. Scaledet: A scalable multi-dataset object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7288–7297.

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*.

Dai, Y.; Wu, Y.; Zhou, F.; and Barnard, K. 2021. Asymmetric Contextual Modulation for Infrared Small Target Detection. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 949–958.

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.

Duan, W.; Ji, L.; Chen, S.; Zhu, S.; Huang, J.; and Ye, M. 2025a. Weakly Supervised Contrastive Learning With Quantity Prompts for Moving Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–14.

Duan, W.; Ji, L.; Chen, S.; Zhu, S.; and Ye, M. 2024. Triple-Domain Feature Learning With Frequency-Aware Memory Enhancement for Moving Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–14.

Duan, W.; Ji, L.; Huang, J.; Chen, S.; Peng, S.; Zhu, S.; and Ye, M. 2025b. Semi-supervised Multi-view Prototype Learning with Motion Reconstruction for Moving Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*.

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

Hui, B.; Song, Z.; Fan, H.; Zhong, P.; Hu, W.; Zhang, X.; Lin, J.; Su, H.; Jin, W.; Zhang, Y.; and Bai, Y. 2019. A dataset for infrared image dim-small aircraft target detection and tracking under ground / air background.

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.

Jain, Y.; Behl, H.; Kira, Z.; and Vineet, V. 2023. Damex: Dataset-aware mixture-of-experts for visual understanding of mixture-of-datasets. *Advances in Neural Information Processing Systems*, 36: 69625–69637.

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.; 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.

Li, Y.; Li, X.; Li, Y.; Zhang, Y.; Dai, Y.; Hou, Q.; Cheng, M.-M.; and Yang, J. 2024. Sm3det: A unified model for multi-modal remote sensing object detection. *arXiv preprint arXiv:2412.20665*.

Lin, F.; Hu, W.; Wang, Y.; Tian, Y.; Lu, G.; Chen, F.; Xu, Y.; and Wang, X. 2024. Universal object detection with large vision model. *International Journal of Computer Vision*, 132(4): 1258–1276.

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*.

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.

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.

Shi, C.; Zhu, Y.; and Yang, S. 2024. Plain-Det: A Plain Multi-Dataset Object Detector. In *European Conference on Computer Vision*, 210–226. Springer.

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.

Tong, X.; Zuo, Z.; Su, S.; Wei, J.; Sun, X.; Wu, P.; and Zhao, Z. 2024. ST-Trans: Spatial-Temporal Transformer for Infrared Small Target Detection in Sequential Images. *IEEE Transactions on Geoscience and Remote Sensing*.

Wang, G.; Tao, B.; Kong, X.; and Peng, Z. 2021. Infrared small target detection using nonoverlapping patch spatial-temporal tensor factorization with capped nuclear norm regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–17.

Wang, Z.; Li, Y.; Chen, X.; Lim, S.-N.; Torralba, A.; Zhao, H.; and Wang, S. 2024. UniDetector: Towards Universal Object Detection with Heterogeneous Supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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, 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.

Yang, J.; Zhu, H.; Wang, Y.; Wu, G.; He, T.; and Wang, L. 2025. Tra-moe: Learning trajectory prediction model from multiple domains for adaptive policy conditioning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6960–6970.

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*.

Zhang, T.; Li, L.; Cao, S.; Pu, T.; and Peng, Z. 2023a. Attention-Guided Pyramid Context Networks for Detecting Infrared Small Target Under Complex Background. *IEEE Transactions on Aerospace and Electronic Systems*, 59(4): 4250–4261.

Zhang, Y.; Zhang, Y.; Shi, Z.; Fu, R.; Liu, D.; Zhang, Y.; and Du, J. 2023b. Enhanced cross-domain dim and small infrared target detection via content-decoupled feature alignment. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–16.

Zhou, Q.; Li, X.; He, L.; Yang, Y.; Cheng, G.; Tong, Y.; Ma, L.; and Tao, D. 2022. TransVOD: end-to-end video object detection with spatial-temporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 7853–7869.

Zhu, S.; Ji, L.; Chen, S.; and Duan, W. 2025. Spatial–temporal-channel collaborative feature learning with transformers for infrared small target detection. *Image and Vision Computing*, 154: 105435.

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*, 124731.