Motion Prior Knowledge Learning with Homogeneous Language Descriptions for Moving Infrared Small Target Detection

Shengjia Chen, Luping Ji*, Weiwei Duan, Shuang Peng, 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, {dww, shuangpeng}@std.uestc.edu.cn, maoye@uestc.edu.cn

Abstract

Different from traditional object detection, pure vision is not enough to infrared small target detection, due to small target size and weak background contrast. For promoting detection performance, more target representations are needed. Currently, motion representations have been proved to be one of the most potential feature kinds for infrared small target detection. Existing methods have an obvious weakness, that besides vision features, they could only capture coarse motion representations from temporal domain. With vision features, fine motion representations could be more effective to enhance detection performance. To overcome this weakness, inspired by prevalent vision-language models, we propose the first vision-language framework with motion prior knowledge learning (MoPKL). Breaking through traditional pure-vision modality, it utilizes homogeneous language descriptions, formatted for moving targets, to directionally guide vision channel learning motion prior knowledge. With the facilitation of motion-vision alignment and motion-relation mining, the motion of infrared small targets is further refined by graph attention, to generate more fine motion representations. The extensive experiments on datasets ITSDT-15K and IRDST show that our framework is effective. It could often obviously outperform other methods.

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

Introduction

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, including military use, autonomous vehicles and maritime monitoring (Liu et al. 2024b). It has garnered extensive research attention over the past decades (Zeng, Li, and Peng 2006; Dai et al. 2021b).

Compared to general RGB objects (Chen, Li, and Tang 2020; Chen, Yang, and Li 2023), infrared small targets possess two unique characteristics: *Small* and *Dim* (Chen et al. 2023), although their real-world sizes could be large (Deshpande et al. 1999), such as the vehicles in infrared images (Fu et al. 2022). Sometimes, impacted by target distances and imaging means, these targets in infrared images could typically have low contrast against backgrounds and may even lack distinct brightness, shape, or texture features (Zeng, Li, and Peng 2006; Zhang et al. 2022). Given these two characteristics above, it is often an extremely challenging issue for ISTD models to effectively learn and accurately detect small targets from highly complex infrared application scenarios (Hou et al. 2022; Zhang et al. 2024).

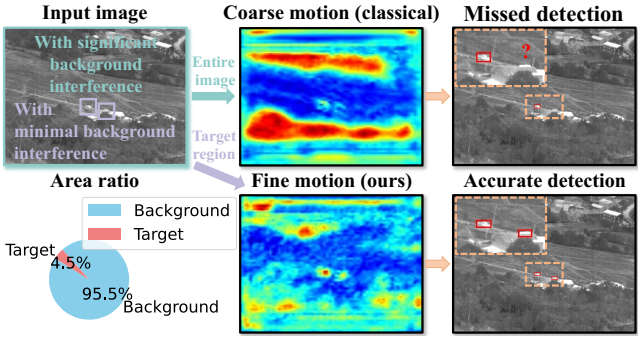

Figure 1: Typical limitation of coarse target motion capturing in existing classic schemes.

At present, single-frame schemes seem insufficient to the increasing demand for video signals (Yan et al. 2023). Some recent works have begun to explore ISTD on video signals, leading to the emergence of a new research branch: *moving infrared small target detection* (MISTD). Limited by weak visual features, auxiliary motion representations have shown impressive potential in moving infrared target detection.

Existing classical MISTD approaches mainly capture differences between consecutive frames of entire images to compute motion representations (Li et al. 2023b). However, despite their simplicity and efficiency, these approaches to computing motion features could be pregnant with a potential limitation: *coarse motion representations*, as shown in Figure 1. In infrared images, targets occupy only a small portion of images, with the vast majority consisting of background information. This leads to classical methods being highly susceptible to background changes or diverting significant attention to background motion, resulting in only coarse motion representations (Chen et al. 2024a,c).

Humans tend to understand new concepts by the vision information under the guiding of language prompts (Jack-

*Corresponding Author.

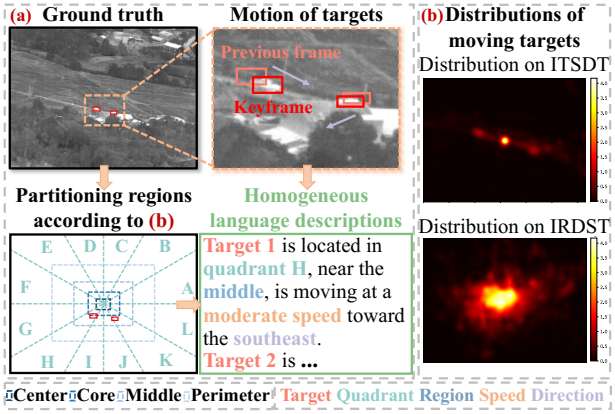

Figure 2: Homogeneous language description problem. From language perspective, target motion could be represented as a description involving multiple motion parameters. (a) Generation scheme of language descriptions. (b) Location distributions of moving targets.

endoff 1987; Smith and Gasser 2005). Language describes certain concepts with greater precision, making it easier to comprehend them. Therefore, we instinctively speculate that in most natural images, the concept of motion could be strengthened by the structured language, composing of key clues, *e.g*, relative position, motion speed and direction. For example, a car (*target*) is moving slowly (*speed*) from southeast corner (*position*) towards northwest (*direction*).

Inspired by the analysis above, we consider addressing this coarse motion representations, from a visual-language perspective. Specifically, we can transform this problem of obtaining fine motion features in vision channel, into the construction of homogeneous text descriptions in language channel, as presented in Figure 2. In scheme, we could partition an image into multiple regions, and determine the moving locations of targets. Then, motion speed and direction can be computed by the target locations in two adjacent frames. Therefore, a complete language description should include: *target*, *quadrant*, *region*, *speed* and *direction*.

In view of this, introducing language encoding, this paper explores the first visual-language framework, *i.e.,* MoPKL, suitable for ISTD. It aids detection models to learn motion prior knowledge, under the guiding of the homogeneous language descriptions to moving small targets. Fine motion feature representations are obtained by the further refining of graph attention. The primary contributions of this paper are summarized as follows. **(I)** We propose the first vision-language framework with motion prior knowledge learning. **(II)** We design motion-vision alignment and motion-relation mining to capture the motion prior knowledge of targets. **(III)** We explore a general method to generate homogeneous language descriptions for moving infrared small targets.

Related Works

Infrared Small Target Detection

Representative single-frame model-driven schemes encompass feature-based detection techniques such as the local contrast measure (LCM) (Chen et al. 2013), along with enhanced methods including ILCM (Han et al. 2014). Furthermore, several approaches separate targets utilizing the low-rank characteristics of background (Liu et al. 2023).

With the development of deep neural networks, data-driven schemes have advanced significantly and become a mainstream paradigm (Dai et al. 2021a). For instance, AGPCNet (Zhang et al. 2023) develops a context pyramid network with attention-guided context computation blocks. To address the issue of features being easily lost in deep layers, DNANet (Li et al. 2023a) proposes a dense nested convolutional network. RDIAN (Sun et al. 2023) introduces a new ISTD dataset called IRDST and proposes an efficient network with low complexity. RPCANet (Wu et al. 2024) presents a deep unfolding-based method that achieves high performance with only a minimal number of parameters. However, single-frame methods ignore motion features in multi-frame images, rendering them ineffective in challenging multi-frame cases (Chen et al. 2024b).

Moving Infrared Small Target Detection

In contrast to single-frame methods, multi-frame methods could extract additional motion features, providing them with greater potential (Du and Hamdulla 2019). For example, inter-frame difference method (Kim, Sun, and Kim 2014), detect moving targets by calculating the difference between two successive frames. Recently, some tensor-optimized methods, including the 4-D STT model (Wu et al. 2023), have demonstrated high performance in ISTD.

Even though these multi-frame model-driven approaches have made significant advancements in ISTD, their robustness is limited, and they may struggle to effectively handle real-world scenarios with complex noise (Zhu et al. 2023). Therefore, to overcome these limitations, data-driven methods based on multiple frames have begun to emerge continuously (Li et al. 2023b). For instance, state-of-the-art (SOTA) multi-frame method SSTNet (Chen et al. 2024a) introduces a spatio-temporal network that captures coarse motion in feature maps across multiple frames. Unlike existing multi-frame methods, we construct motion prior knowledge using homogeneous language descriptions to assist the model in learning fine target motion.

Vision-Language Models

Recently, the development of natural language processing (NLP) has entered a new era. For example, Bert (Devlin et al. 2018), has provided a universal solution that simplifies and enhances the efficiency of developing numerous NLP tasks. Inspired by advancements in NLP, VLMs are pre-trained using large-scale image-text pairs (Yao et al. 2021). For example, pre-trained CLIP (Radford et al. 2021) capture rich visual-language correspondence knowledge and can perform zero-shot prediction (Liu et al. 2024a) by matching the embedding of any given image and text. Current VLMs may only be suitable for the tasks with clearly defined classes. However, ISTD does not usually involve explicit semantic classes. Therefore, we propose a new framework specifically designed for ISTD, distinct from VLMs.

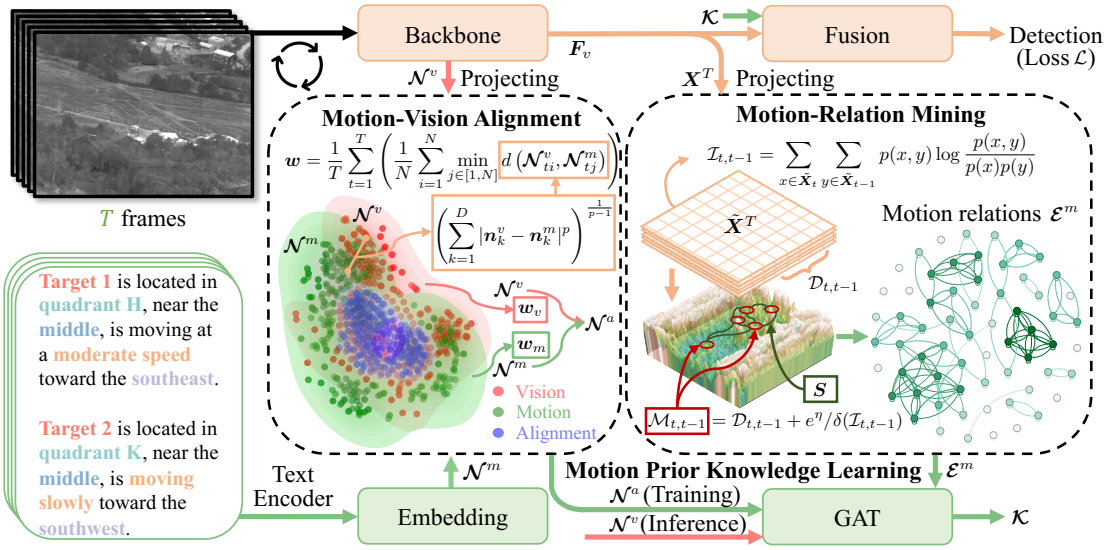

Figure 3: MoPKL framework. Frame set, $\boldsymbol{I}^T = [\boldsymbol{I}_1, \boldsymbol{I}_2, ..., \boldsymbol{I}_T]$, serves as original input. Output set $\boldsymbol{F}_v$, generated by backbone with shared weights, gains new features $\boldsymbol{\mathcal{N}}^v$ and $\boldsymbol{X}^T$ through subspace projection. Both $\boldsymbol{\mathcal{N}}^v$ and language description embeddings $\boldsymbol{\mathcal{N}}^m$, obtained via text Encoder, are fed into **Motion-Vision Alignment** to derive aligned motion presentations $\boldsymbol{\mathcal{N}}^a$ (node features). $\boldsymbol{X}^T$ is fed into **Motion-Relation Mining** to obtain the motion relation $\boldsymbol{\mathcal{E}}^m$ of potential target regions. Then, $\boldsymbol{\mathcal{N}}^a$ ($\boldsymbol{\mathcal{N}}^v$ for inference) and $\boldsymbol{\mathcal{E}}^m$ are fed into GAT for motion prior knowledge learning, obtaining knowledge features $\boldsymbol{\mathcal{K}}$. Finally, $\boldsymbol{\mathcal{K}}$ is utilized to enhance features $\boldsymbol{F}_v$, before detection loss $\mathcal{L}$ is computed.

Proposed Method

Overall MoPKL Pipeline

Definition 1 *(Motion Prior Knowledge):* Motion prior knowledge is represented as a relation graph $\boldsymbol{\mathcal{G}} = (\boldsymbol{\mathcal{N}}, \boldsymbol{\mathcal{E}})$, where $\boldsymbol{\mathcal{N}} \in \mathbb{R}^{N \times D}$ and $\boldsymbol{\mathcal{E}} \in \mathbb{R}^{N \times N}$ are sets of nodes (motion presentations) and relations (between potential moving target regions), respectively. Here, N and D denote the number of nodes and feature dimensions. Each edge $e_{ij} \in \boldsymbol{\mathcal{E}}$ specifies a type of motion relation between regions. By mapping nodes and relations to feature vectors using a graph model, knowledge features $\boldsymbol{\mathcal{K}} \in \mathbb{R}^{N \times D}$ could be obtained.

To efficiently reducing semantic gap for enhancing the degree of cooperation between two modalities to acquire motion representations $\boldsymbol{\mathcal{N}}$, we propose motion-vision alignment. To capture the motion relations $\boldsymbol{\mathcal{E}}$ of potential target regions, we design motion-relation mining. Motion representations and motion relations constitute a relation graph $\boldsymbol{\mathcal{G}}$. Furthermore, MoPKL framework is proposed for motion prior knowledge learning, as illustrated in Figure 3.

First, the visual input to entire network is a T-frame image set, denoted as $\boldsymbol{I}^T \in \mathbb{R}^{3 \times T \times W \times H}$, where H and W represent the height and width of input images, respectively. We use CSPDarknet (Ge et al. 2021) as a backbone network. We extract multi-frame primary visual features by feeding each frame image iteratively into backbone, similar to most video object detection methods (Gong et al. 2021). This process results in a visual feature set $\boldsymbol{F}_v \in \mathbb{R}^{T \times c \times w \times h}$, where c, h, and w indicate the channel, height, and width of the features, respectively. Then, new visual representations $\boldsymbol{\mathcal{N}}^v \in \mathbb{R}^{T \times N \times M}$ and $\boldsymbol{X}^T \in \mathbb{R}^{T \times c \times W/2 \times H/2}$ are obtained by projecting $\boldsymbol{F}_v$ into a semantic subspace. In addition, we transform the generated language descriptions into motion representations $\boldsymbol{\mathcal{N}}^m \in \mathbb{R}^{T \times N \times M}$ by utilizing pre-trained GloVe (Pennington, Socher, and Manning 2014) word embedding vectors, where E represents embedding dimension.

Second, both $\boldsymbol{\mathcal{N}}^v$ and $\boldsymbol{\mathcal{N}}^m$ are processed in *Motion-Vision Alignment* to derive aligned cross-modal motion presentations $\boldsymbol{\mathcal{N}}^a \in \mathbb{R}^{(T-1) \times N \times M}$ (*i.e.,* node features). $\boldsymbol{X}^T$ is input into *Motion-Relation Mining* to obtain the motion relations $\boldsymbol{\mathcal{E}}^m \in \{0,1\}^{(T-1) \times N \times N}$ (*i.e.,* an adjacency matrix) for potential target regions. Next, graph $\boldsymbol{\mathcal{G}} = (\boldsymbol{\mathcal{N}}^a, \boldsymbol{\mathcal{E}}^m)$ (training) is input into a graph attention network (GAT) (Veličković et al. 2018) to learn and obtain motion knowledge features $\boldsymbol{\mathcal{K}} \in \mathbb{R}^{(T-1) \times N \times M}$, as follows:

$$\boldsymbol{\mathcal{K}}_i^{l+1} = \frac{1}{K} \sum_{k=1}^{K} \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^k \boldsymbol{\mathcal{K}}_j^{lk} \right), \tag{1}$$

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{\mathcal{K}}_i^l \,\|\, \boldsymbol{\mathcal{K}}_j^l]))}{\sum_{k \in \mathcal{N}(i)} \exp(\text{LeakyReLU}(\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{\mathcal{K}}_i^l \,\|\, \boldsymbol{\mathcal{K}}_k^l]))}, \tag{2}$$

where $\boldsymbol{\mathcal{K}}_i^{l+1}$ is the i-th node feature at layer $l+1$ obtained through attention mechanism-weighted aggregation. This vector represents the feature representation after integrating the information from the neighbors $\mathcal{N}(i)$ of node i. α_{ij}^k denotes the attention coefficient for the k-th head, with K being the total number of attention heads. σ is a nonlinear activation function. $\boldsymbol{a}$ is a learnable attention weight vector, $[\cdot\|\cdot]$ denotes concatenation and $(\cdot)^{\mathrm{T}}$ indicates transpose.

Finally, fusing $\boldsymbol{F}_v$ with reshaped $\boldsymbol{\mathcal{K}}$ to obtain output features. These features are then processed through decoupled detection heads to compute detection loss $\mathcal{L}$ (Ge et al. 2021).

Text-guided Motion-Vision Alignment

To reduce the semantic gap between motion (*i.e.,* language modality) and vision construct cross-modal motion representations (*i.e.,* nodes), we propose a Motion-Vision Alignment (MVA), as shown in Figure 3. First, we calculate the spatial distance between distributions in motion and vision modality nodes, described by following equations:

$$d(\boldsymbol{n}^v, \boldsymbol{n}^m) = \left(\sum_{k=1}^{D} |\boldsymbol{n}_k^v - \boldsymbol{n}_k^m|^p \right)^{\frac{1}{p-1}}, \tag{3}$$

where $\boldsymbol{n}^v$ and $\boldsymbol{n}^m$ denote the vision and motion representations of a single frame, respectively. D is the dimension of node representations and p is a hyper-parameter with $p \geq 2$.

Second, in this way, we could calculate the average minimum distance between the feature distributions $\boldsymbol{\mathcal{N}}^v$ and $\boldsymbol{\mathcal{N}}^m$ across all frames and nodes, as follows:

$$w = \frac{1}{T} \sum_{t=1}^{T} \left(\frac{1}{N} \sum_{i=1}^{N} \min_{j \in [1,N]} d\left(\boldsymbol{\mathcal{N}}_{ti}^v, \boldsymbol{\mathcal{N}}_{tj}^m\right) \right), \tag{4}$$

where w is is a weight that reflects the degree of alignment between feature distributions $\boldsymbol{\mathcal{N}}^v$ and $\boldsymbol{\mathcal{N}}^m$. Here, T is the number of frames and N is the number of nodes, respectively. This means that the closer vision and motion feature distributions are, the smaller w becomes, and vice versa.

Finally, take the reciprocals of calculated w_v and w_m to obtain vision and motion weights, ensuring that feature distributions with larger distances have smaller weights. These weights are then multiplied by $\boldsymbol{\mathcal{N}}^v$ and $\boldsymbol{\mathcal{N}}^m$, respectively, to achieve a cross-modal mapping with significant semantic (*i.e.,* feature distribution) correlations, resulting in distribution-aligned node features $\boldsymbol{\mathcal{N}}^a$.

Motion-Relation Mining

To construct motion relations $\boldsymbol{\mathcal{E}}^m$ through potential fine-moving target regions, we propose a Motion-Relation Mining method, as illustrated in Figure 3.

First, we divide the input visual feature map $\boldsymbol{X}^T \in \mathbb{R}^{T \times c \times W/2 \times H/2}$ into q sub-regions of size $s \times s$, where the size is close to that of target bounding boxes, obtaining $\tilde{\boldsymbol{X}}^T \in \mathbb{R}^{T \times c \times q \times s \times s}$ ($q = \frac{W}{2s} \times \frac{H}{2s}$). Here, T is the number of frames and c is the number of channels. H, W represent the height and width of input images, respectively.

Second, to identify motion-salient regions, we consider two sources of information: one is the frame difference motion information (specific differences), and the other is mutual information (focusing on statistical dependence rather than specific differences) (Belghazi et al. 2018). To achieve this, we calculate the inter-frame difference $\mathcal{D}_{t,t-1}$ and mutual information $\mathcal{I}_{t,t-1}$ between adjacent frames (t and $t-1$) for each region to obtain motion information $\mathcal{M}_{t,t-1}$, where t is key frame, and $t-1$ is previous frame, as follows:

$$\mathcal{M}_{t,t-1} = \mathcal{D}_{t,t-1} + e^{\eta}/\delta(\mathcal{I}_{t,t-1}), \tag{5}$$

where η is a hyper-parameter and δ is a *Tanh* function. The inter-frame difference $\mathcal{D}_{t,t-1}$ is calculated by calculating the feature difference between each region of key frame $\tilde{\boldsymbol{X}}_t$ and previous frame $\tilde{\boldsymbol{X}}_{t-1}$ as follows:

$$\mathcal{D}_{t,t-1} = \sum_{k \in q} \sum_{i \in s} \sum_{j \in s} |\tilde{\boldsymbol{X}}_t(k,i,j) - \tilde{\boldsymbol{X}}_{t-1}(k,i,j)|. \tag{6}$$

The features of a region are mutually dependent between adjacent frames and share a large amount of information, indicating that no significant motion could have occurred in this region. Conversely, less shared information suggests noticeable feature changes, indicating the high possibility of motion. Thus, a smaller value of mutual information indicates more significant motion, and it is described as follows:

$$\mathcal{I}_{t,t-1} = \mathcal{H}_t + \mathcal{H}_{t-1} - \mathcal{H}_{t,t-1}, \tag{7}$$

where $\mathcal{I}_{t,t-1}$ represents the amount of shared information or mutual dependency between t and $t-1$ frames. $\mathcal{H}_t$ represents the entropy of key frame, $\mathcal{H}_{t-1}$ represents the entropy of previous frame, and $\mathcal{H}_{t,t-1}$ represents joint entropy. They are calculated as following:

$$\mathcal{H}_t = -\sum_{x \in \tilde{\boldsymbol{X}}_t} p(x) \log p(x), \tag{8}$$

$$\mathcal{H}_{t,t-1} = -\sum_{x \in \tilde{\boldsymbol{X}}_t} \sum_{y \in \tilde{\boldsymbol{X}}_{t-1}} p(x,y) \log p(x,y), \tag{9}$$

where entropy $\mathcal{H}_t$ represents the information content of $\tilde{\boldsymbol{X}}_t$. Joint entropy $\mathcal{H}_{t,t-1}$ represents overall information content when considering two frames $\tilde{\boldsymbol{X}}_t$ and $\tilde{\boldsymbol{X}}_{t-1}$ together. $p(x)$ is the marginal probability distribution of $\tilde{\boldsymbol{X}}_t$. We then obtain a detailed description of $\mathcal{I}_{t,t-1}$ as follows:

$$\mathcal{I}_{t,t-1} = \sum_{x \in \tilde{\boldsymbol{X}}_t} \sum_{y \in \tilde{\boldsymbol{X}}_{t-1}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{10}$$

Finally, select the N regions with the highest motion information values. These selected regions have a high likelihood of containing motion targets. To mine their mutual motion relations, we calculate the Mahalanobis distance $\boldsymbol{S}$ between these regions, as shown in following equations:

$$\boldsymbol{S} = \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} \sqrt{(\boldsymbol{c}_{t,i} - \boldsymbol{c}_{t,j})^{\mathrm{T}} \Sigma_t^{-1} (\boldsymbol{c}_{t,i} - \boldsymbol{c}_{t,j})} \tag{11}$$

where Σ^{-1} is the inverse of covariance matrix and $\boldsymbol{c}$ is the center point of regions. Next, construct the motion relations between regions based on calculated distances. To construct adjacency matrix $\boldsymbol{\mathcal{E}}^m$, assigning the top τ relations a value of 1, while the others are assigned a value of 0.

Experiments

Implementation Details

We evaluate our MoPKL on two datasets: ITSDT-15K (Fu et al. 2022) and IRDST (Sun et al. 2023). To ensure a fair comparison with other methods, we adhere to the standard evaluation paradigm (Chen et al. 2024a), using metrics such as Precision (*Pr*), Recall (*Re*), F1 score, and Average Precision (mAP_{50}, indicating the mean Average Precision at an

Scheme	Method	Publication	ITSDT-15K				IRDST			
			mAP_{50} (%)	Pr (%)	Re (%)	F1 (%)	mAP_{50} (%)	Pr (%)	Re (%)	F1 (%)
Model-driven	MaxMean	SPIE 1999	0.87	10.85	8.74	9.68	0.01	0.28	1.48	0.47
	TopHat	IPT 2006	11.61	27.21	43.07	33.35	1.81	18.22	10.60	13.40
	RLCM	IEEE TGRS 2018	4.62	15.38	30.76	20.50	1.58	16.28	9.70	12.16
	HBMLCM	IEEE GRSL 2019	0.72	7.97	9.37	8.61	1.16	29.14	4.66	8.03
	PSTNN	RS 2019	7.99	22.98	35.21	27.81	1.45	16.28	9.70	12.16
	WSLCM	SP 2020	2.36	16.78	14.53	15.58	1.69	20.87	8.70	12.28
Data-driven	ACM	WACV 2021	55.38	78.37	71.69	74.88	52.40	76.33	69.32	72.66
	RISTD	IEEE GRSL 2022	60.47	85.49	71.60	77.93	66.57	84.70	79.63	82.08
	ISTDUNet	IEEE GRSL 2022	64.74	82.73	80.02	81.35	63.01	83.40	76.73	79.93
	ISNet	CVPR 2022	62.29	83.46	75.32	79.18	59.78	80.24	75.08	77.58
	UIUNet	IEEE TIP 2022	65.15	84.07	78.39	81.13	56.38	80.95	70.29	75.25
	SANet	ICASSP 2023	62.17	87.78	71.23	78.64	64.54	84.29	77.02	80.49
	AGPCNet	IEEE TAES 2023	67.27	<u>91.19</u>	74.77	82.16	59.21	79.47	75.51	77.44
	RDIAN	IEEE TGRS 2023	68.49	90.56	76.06	82.68	59.08	77.99	76.35	77.16
	DNANet	IEEE TIP 2023	70.46	88.55	80.73	84.46	63.61	82.92	77.48	80.11
	DTUM★	IEEE TNNLS 2023	67.97	77.95	**88.28**	82.79	71.48	82.87	**87.79**	<u>85.26</u>
	SIRST5K	IEEE TGRS 2024	61.52	86.95	71.32	78.36	52.28	76.12	69.07	72.42
	MSHNet	CVPR 2024	60.82	89.69	68.44	77.64	63.21	82.31	77.64	79.91
	RPCANet	WACV 2024	62.28	81.46	77.10	79.22	56.50	77.77	73.80	75.73
	SSTNet★	IEEE TGRS 2024	<u>76.96</u>	91.05	85.29	<u>88.07</u>	<u>71.55</u>	<u>88.56</u>	81.92	85.11
	MoPKL (ours)★	AAAI 2025	**79.78**	**93.29**	<u>86.80</u>	**89.92**	**74.54**	**89.04**	<u>84.74</u>	**86.84**

Table 1: The quantitative detection performance comparisons on two datasets. The best metric is marked in bold, and the second-best one is underlined. Symbol ★ indicates that this method is a data-driven multi-frame scheme.

Intersection over Union (IoU) threshold of 0.5). Since most comparison methods rely on pixel-level segmentation and evaluation datasets are annotated with bounding boxes, we add a detection head (Ge et al. 2021) to these methods for generating bounding boxes. For all comparison methods, input images are resized to a resolution of 512×512 (without data augmentation). Furthermore, our MoPKL and the other comparison methods are trained for 100 epochs with a batch size of 4. Initial learning rate is set to 0.01, using SGD as the optimizer, with a momentum of 0.937, a weight decay factor of 5×10^{-4}. Hyper-parameters τ, s, η, N, T and K are set to 13, 8, 0.8, 64, 5 and 8, respectively.

Comparisons with SOTA Ones

Numerical Comparisons Table 1 presents numerical comparisons across fifteen recent methods, revealing two notable observations. One is that our MoPKL establishes new SOTA benchmarks across numerous metrics, consistently securing the top performance indicators on both datasets. For example, on ITSDT-15K, MoPKL could achieve the highest mAP_{50} 79.78%,*Pr* 93.29% and F1 score 89.92%. Only in terms of *Re*, the 86.80% by MoPKL is slightly lower than the SOTA 88.28% by DTUM (Li et al. 2023b). DTUM achieves higher *Re* by sacrificing *Pr* performance, whereas our method is more balanced, achieving a higher F1 score. Moreover, on IRDST, MoPKL could also obtain the highest mAP_{50} 74.54%, *Pr* 89.04% and F1 score 86.84%, surpassing most other methods.

Inference Cost Comparisons. The cost comparisons are presented in Table 2. From these comparisons, two clear findings emerge. First, although our method utilizes a sequence of five frames, the number of parameters remains at a moderate level. For instance, our model has 9.46M parameters, higher than the SOTA method RDIAN (2.74M), but most compared methods exceed 10M parameters, such as SIRST5K with 11.48M (Lu et al. 2024) and UIUNet with 53.06M (Wu, Hong, and Chanussot 2022). Second, the use of multi-frame images leads to a large number of FLOPs, but modeling multi-frame features enhances model performance. For example, our MoPKL has 119.64G FLOPs, higher than most methods. However, the mAP_{50} and F1

Methods	Frames	mAP_{50}↑	F1↑	FLOPs↓	Params↓	FPS↑
ACM	1	55.38	74.88	**24.66G**	<u>3.04M</u>	**22.57**
RISTD	1	60.47	77.93	76.28G	3.28M	13.64
ISTDUNet	1	64.74	81.35	394.32G	5.27M	7.01
ISNet	1	62.29	79.18	265.74G	3.48M	8.87
UIUNet	1	65.15	81.13	456.70G	53.06M	2.89
SANet	1	62.17	78.64	<u>42.04G</u>	12.40M	8.62
AGPCNet	1	67.27	82.16	366.15G	14.88M	4.37
RDIAN	1	68.49	82.68	50.44G	**2.74M**	<u>19.15</u>
DNANet	1	70.46	84.46	135.24G	7.22M	4.78
SIRST5K	1	61.25	78.36	182.61G	11.48M	7.37
MSHNet	1	60.82	77.64	69.49G	6.59M	18.55
RPCANet	1	62.28	79.22	382.69G	3.21M	15.89
DTUM	5	67.97	82.79	128.16G	9.64M	14.28
SSTNet	5	76.96	88.07	123.59G	11.95M	9.24
MoPKL (ours)	5	**79.78**	**89.92**	119.64G	9.46M	10.03

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

Settings	B	D_A	D_B	E_A	E_B	N_A	N_B	N_C	ITSDT-15K				IRDST			
									mAP_{50} (%)	Pr (%)	Re (%)	F1 (%)	mAP_{50} (%)	Pr (%)	Re (%)	F1 (%)
w/o All	-	-	-	-	-	-	-	-	71.95	83.43	**87.35**	85.34	64.44	84.59	76.52	80.35
w/ B	✓	-	-	-	-	-	-	-	76.43	91.46	84.20	87.68	68.71	85.18	81.18	83.13
w/ B & D_A	✓	✓	-	-	-	-	-	-	75.42	89.28	84.36	86.75	66.51	83.47	78.74	81.04
w/ B & D_B & E_A & N_A	✓	-	✓	✓	-	✓	-	-	76.81	93.03	83.09	87.78	69.35	85.93	81.40	83.61
w/ B & D_B & E_B & N_A	✓	-	✓	-	✓	✓	-	-	77.62	90.79	86.13	88.40	71.02	86.84	82.87	84.81
w/ B & D_B & E_B & N_B	✓	-	✓	-	✓	-	✓	-	78.87	**93.35**	85.16	89.07	72.27	87.38	83.44	85.36
w/ all	✓	-	✓	-	✓	-	-	✓	**79.78**	93.29	86.80	**89.92**	**74.54**	**89.04**	**84.74**	**86.84**

Table 3: The ablation study on MoPKL with different assemblies and settings. **B:** improved baseline with backbone optimization, **D:** homogeneous language descriptions ($\mathbf{D_A}$ simple embedding and integration, $\mathbf{D_B}$ graph learning (require nodes and edges)), **E:** capture graph edges ($\mathbf{E_A}$ calculating the cosine similarity of visual features, $\mathbf{E_B}$ motion-relation mining), **N:** capture graph node features ($\mathbf{N_A}$ only motion features, $\mathbf{N_B}$ concatenate motion and visual features, $\mathbf{N_C}$ motion-vision alignment).

scores are higher than those of all other methods.

PR Curve Comparisons As usual, we utilize precision-recall (PR) curves to evaluate the comprehensive performance of various methods on ITSDT-15K and IRDST, as shown in Fig. 4. These figures clearly demonstrate that our curves outperform those of competing methods. Specifically, on ITSDT-15K, our curve consistently occupies higher positions, particularly near top-right. This pattern continues on IRDST, where our curve often exceeds others toward top-right. The proximity of a method to top-right corner directly indicates its effectiveness. Therefore, PR curves highlight the superior balance of MoPKL in precision and recall compared to other approaches.

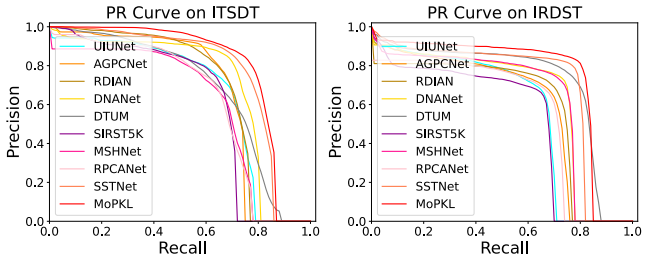

Figure 4: PR curve comparisons on ITSDT-15K and IRDST.

Ablation Study

Effects of Different Assemblies We undertake a series of ablation studies to evaluate the impact of various configurations within our MoPKL, as detailed in Table 3.

Through comparisons, it is apparent that the individual components are consistently effective in enhancing detection capabilities. For instance, on ITSDT-15K, the baseline configuration, devoid of any specialized components, achieves mAP_{50} and F1 score of 71.95% and 85.34%, respectively. The integration of components (w/ B) improves these scores to 76.43% for mAP_{50} and 87.68% for F1. The implementation of settings (w/ B & D_B & E_B & N_A) further increases metrics to 77.62% for mAP_{50} and 88.40% for F1. Ultimately, when all components are fully integrated, performance improves significantly, with an mAP_{50} of 79.78% and an F1 scores of 89.92%, reaching their best levels.

Impacts of Homogeneous Language Descriptions To explore the impacts of homogeneous language descriptions on the performance of MoPKL, a group of experiments is conducted. According to the results presented in Table 4, the performance of general descriptions is the lowest, with a mAP_{50} of 77.48% and an F1 score of 88.57%. One possible reason is that general descriptions are insufficient to clearly describe the fine motion of targets. In contrast, under the settings of **S**, **D** and **Q**, the mAP_{50} and F1 scores of homogeneous descriptions achieve optimal results. This indicates that homogeneous descriptions we constructed could provide a more refined motion representation for targets.

Descriptions	S	D	Q	mAP_{50}	Pr	Re	F1
General	-	-	-	77.48	91.62	85.71	88.57
Homogeneous	✓	-	-	78.05	91.65	85.99	88.73
	✓	✓	-	78.65	92.99	85.75	89.22
	✓	✓	✓	**79.78**	**93.29**	**86.80**	**89.92**

Table 4: The impacts of homogeneous language descriptions on ITSDT-15K. **S:** speed estimation, **D:** direction calculation, **Q:** quadrant division (spatial regions labeled $A \sim L$).

Impacts of Word Embedding Methods To effectively project language descriptions into a semantic feature subspace, we investigate different word embedding methods, as detailed in Table 5. From experimental results, we could observe two obvious findings. One is that pre-trained GloVe embeddings (Pennington, Socher, and Manning 2014) generally achieves better performance compared to pre-trained Bert (Devlin et al. 2018) and neural network (NN) embedding. A possible reason is that static word embeddings are more effective for constructing graph nodes. GloVe generates static word vectors by statistically analyzing word co-occurrence information in large corpora. In contrast, Bert produces context-dependent word vectors, meaning the same word may have different vector representations in different sentences. Another is that as the dimensionality of GloVe word embeddings increases, performance improves accordingly, indicating that higher dimensions encode more fine-grained information compared to lower dimensions.

Word embedding	w/o training	mAP_{50}	Pr	Re	F1
Pre-trained Bert	✓	77.90	93.23	84.57	88.69
NN embedding	×	77.45	91.61	85.98	88.71
GloVe (50 dim.)	✓	77.96	92.27	85.16	88.57
GloVe (100 dim.)	✓	78.23	92.54	85.72	89.00
GloVe (200 dim.)	✓	78.77	92.10	86.23	89.07
GloVe (300 dim.)	✓	**79.78**	**93.29**	**86.80**	**89.92**

Table 5: The comparisons of different word embedding methods on ITSDT-15K. **Bert:** Bert encoder, **NN embedding:** a lookup table that stores embeddings of a fixed dictionary and size, **GloVe:** pre-trained embeddings.

Effects of Motion Prior Knowledge To visually demonstrate the effects of motion prior knowledge, we present four cases of visualization comparisons (feature heatmaps) in Figure 5 corresponding to four moving small target scenes. In this figure, compared with the ground truth, on all heatmap groups, it is evident that the focus positions of the fine motion captured by the MoPKL are more precise than those by the coarse motion across all heatmap groups. This indicates that motion prior knowledge could effectively help model focus on fine motion, thereby enabling precise attention to moving small targets.

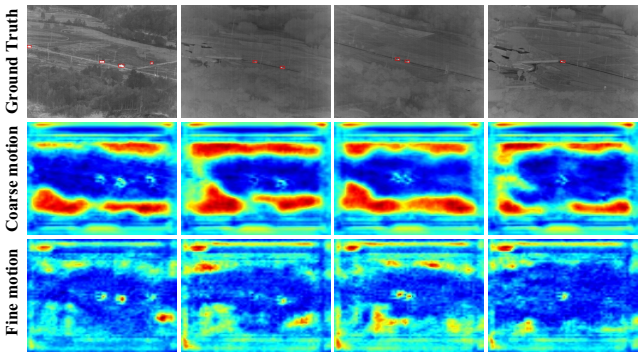

Figure 5: Feature heatmaps of coarse motion (baseline) and our fine motion (MoPKL) on ITSDT-15K.

Effects of Motion-Vision Alignment To visually demonstrate the effectiveness of Motion-Vision Alignment, we construct a group of experiments comparing it with simple concatenation method, as shown in Figure 6. From this, it could be observed that our method aligns motion and vision modal features more effectively on both datasets, whereas concatenation method fails to effectively integrate the features of two modalities. These visualization results further validate the numerical results shown in Table 3 (w/ B & D_B & E_B & N_B and w/ all).

Effects of Mutual Information We further explore the effectiveness of mutual information in Motion-Relation Mining method by conducting a set of visualization experiments, as shown in Figure 7. From it, we could clearly see that incorporating mutual information strengthens strongly correlated motions and weakens weakly correlated motions,

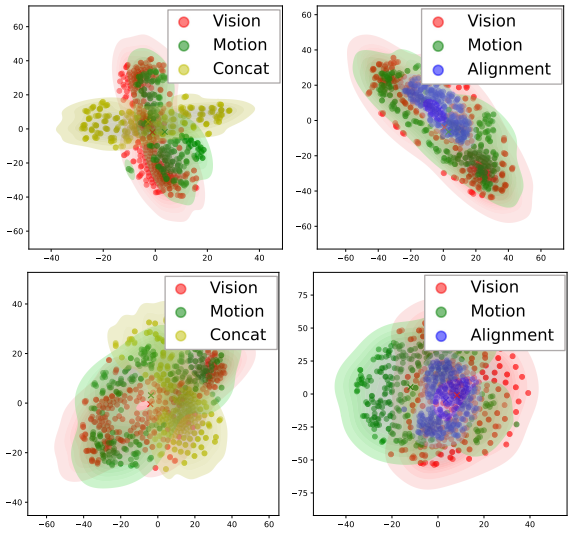

Figure 6: Feature distribution comparisons of Motion-Vision Alignment and simple concatenation on ITSDT-15K (top row) and IRDST (bottom row) datasets.

thereby constructing distinct motion relations for potential motion regions. Without incorporating mutual information, captured motion relations are relatively dispersed and chaotic, lacking distinctiveness. These visualization results also prove the numerical results shown in Table 3.

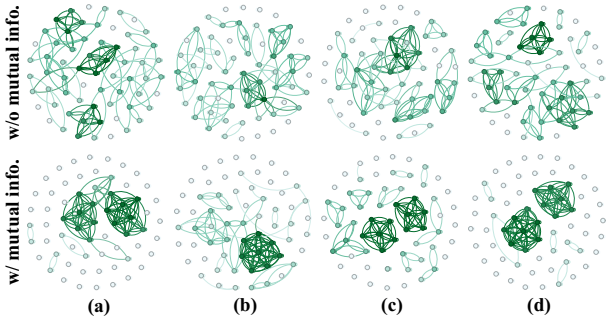

Figure 7: Comparisons of motion relations on ITSDT-15K.

Conclusions

To learn fine motion representation for moving small target detection, this paper proposes the first vision-language framework with motion prior knowledge learning, i.e., MoPKL. Under the guiding of homogeneous description texts, motion-vision alignment and motion-relation mining work together to learn generating fine motion representations. Extensive experiments verify that our MoPKL is effective in capturing the fine motion of moving small targets. On primary metrics, it could often obviously outperform current SOTA methods. Its weakness is low FPS, due to large parameter quantity. In the future, an optimized light detection scheme with more efficient motion prior knowledge learning is worthy of further exploration.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62476049, in part by the Aeronautical Science Foundation of China (ASFC) under Grant 2022Z071080006.

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.; 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. 2024b. MICPL: Motion-Inspired Cross-Pattern Learning for Small-Object Detection in Satellite Videos. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.

Chen, S.; Ji, L.; Zhu, S.; Ye, M.; Ren, H.; and Sang, Y. 2024c. 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.; Yang, X.; and Li, Z. 2023. Improving semantic segmentation with knowledge reasoning network. *Journal of Visual Communication and Image Representation*, 96: 103923.

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. 2021a. Asymmetric Contextual Modulation for Infrared Small Target Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 950–959.

Dai, Y.; Wu, Y.; Zhou, F.; and Barnard, K. 2021b. 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.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Fu, R.; Fan, H.; Zhu, Y.; Hui, B.; Zhang, Z.; Zhong, P.; Li, D.; Zhang, S.; Chen, G.; and Wang, L. 2022. A dataset for infrared time-sensitive target detection and tracking for air-ground application. *Science Data Bank*.

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

Gong, T.; Chen, K.; Wang, X.; Chu, Q.; Zhu, F.; Lin, D.; Yu, N.; and Feng, H. 2021. Temporal ROI Align for Video Object Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1442–1450.

Han, J.; Ma, Y.; Zhou, B.; Fan, F.; Liang, K.; et al. 2014. A robust infrared small target detection algorithm based on human visual system. *IEEE Geoscience and Remote Sensing Letters*, 11(12): 2168–2172.

Hou, Q.; Wang, Z.; Tan, F.; Zhao, Y.; Zheng, H.; and Zhang, W. 2022. RISTDnet: Robust Infrared Small Target Detection Network. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.

Jackendoff, R. 1987. On beyond zebra: The relation of linguistic and visual information. *Cognition*, 26(2): 89–114.

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. 2023a. 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. 2023b. Direction-coded temporal U-shape module for multiframe infrared small target detection. *IEEE Transactions on Neural Networks and Learning Systems*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, Q.; Liu, R.; Zheng, B.; Wang, H.; and Fu, Y. 2024b. 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*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Smith, L.; and Gasser, M. 2005. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2): 13–29.

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.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

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.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

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, 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.

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.