# SeViL: Semi-supervised Vision-Language Learning with Text Prompt Guiding for Moving Infrared Small Target Detection

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

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

## Abstract

Unlike traditional object detection, moving infrared small target detection is highly challenging due to tiny target size and limited labeled samples. Currently, most existing methods mainly focus on the pure-vision features usually by fully-supervised learning, heavily relying on extensive high-cost manual annotations. Moreover, they almost have not concerned the potentials of multi-modal (*e.g.*, vision and text) learning yet. To address these issues, inspired by prevalent vision-language models, we propose the first semi-supervised vision-language (SeViL) framework with adaptive text prompt guiding. Breaking through traditional pure-vision modality, it takes text prompts as prior knowledge to adaptively enhance target regions and then filter the low-quality pseudo-labels generated on unlabeled data. In the meanwhile, we employ an adaptive cross-modal masking strategy to align text and vision features, promoting cross-modal deep interactions. Remarkably, our extensive experiments on three public datasets (DAUB, ITSDT-15K and IRDST) verify that our new scheme could outperform other semi-supervised ones, and even achieve comparable performance to fully-supervised state-of-the-art (SOTA) methods, with only 10% labeled training samples.

**Code** — https://github.com/UESTC-nnLab/SeViL

## Introduction

Infrared small target detection (ISTD) has become a critical task, due to its independence from external lighting and all-weather working capability (Duan et al. 2025a). It has been widely applied in various areas, including remote sensing, monitor system and rescue missions (Dai et al. 2021). As an important research branch of object detection, it is currently attracting more and more attention (Zhu et al. 2024).

Compared with conventional vision objects, due to long imaging distance, infrared small targets usually show two special vision characteristics: *small* (image size) and *dim* (background contrast), often lacking distinct shapes and texture features. Besides, infrared small targets are usually with low *signal-to-noise ratios* in complex backgrounds. In view of this, it is often extremely challenging to accurately locate, detect and recognize moving infrared small targets in infrared background images and videos (Zhu et al. 2025).

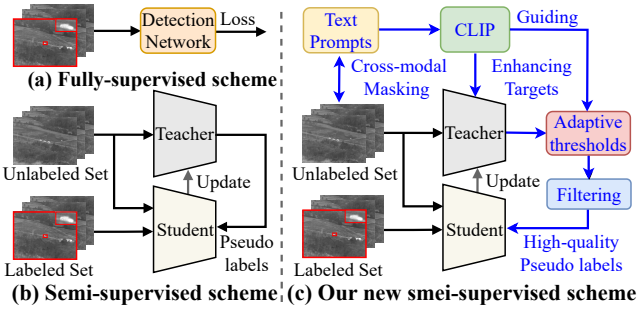


Figure 1: The comparisons of typical fully-supervised, semi-supervised, and our new vision-language learning schemes.

In early stages, to address ISTD, many model-driven methods were firstly proposed, *e.g.*, MaxMean (Deshpande et al. 1999) and TopHat (Bai and Zhou 2010). This category of methods heavily depends on handcrafted features, no adaptive learning ability. To overcome the obvious shortcomings of model-driven ones, in the past decade, many effective data-driven detection schemes have been proposed. This type of schemes could intelligently learn target features, achieving significant performance promotion.

According to the number of input frames, data-driven methods could be grouped into two categories: *Single-frame* and *Multi-frame* (Peng et al. 2025). Single-frame schemes often rely on the vision features of only an individual image, ignoring the relations between two neighboring frames, *e.g.*, DNANet (Li et al. 2022a), MSHNet (Liu et al. 2024) and PConv (Yang et al. 2025). They are often unsuitable for video scenes. For detecting dim-small targets, it is essential to mine more available information from multiple consecutive frames. To implement this purpose, in the past years, some multi-frame methods have been proposed, *e.g.*, DTUM (Li et al. 2025), MoPKL (Chen et al. 2025). In contrast, this type of schemes often uses the spatio-temporal features (including motion information) of infrared targets, outperforming the pure vision features of single-frame ones.

Totally observing, almost all currently-existing data-driven methods depend on fully-supervised learning, requir-

*Corresponding Author

ing a large number of high-cost annotated samples, as shown in Figure 1 (a). Once there are not enough annotations, this category of methods will be with serious performance degradation. Therefore, exploring new detection schemes ($e.g.$, semi-supervised), not requiring too many sample annotations, seems very meaningful in ISTD research field.

Currently, in traditional object detection, semi-supervised schemes have been widely explored, $e.g.$, (Liu et al. 2023b). They mainly utilize the strong vision features of objects. However, for infrared small targets, no strong vision features could be reliably captured. As a result, these methods are often unsuitable for infrared moving small target detection. To address a special semi-supervised scheme for infrared small targets, the S2MVP (Duan et al. 2025b), a pioneer work breaking through fully-supervised framework, as shown in Figure 1 (b), is firstly proposed, obtaining the comparable performance to fully-supervised ones. Nevertheless, like traditional semi-supervised ones, this one also only focuses on classic vision modality. For further improvement, utilizing more modalities could be a potential choice.

For humans, in visual perception tasks, they could often high-effectively recognize and understand objects, under the help of text prompts (Jackendoff 1987). It implies that the text prompts could be used to guide visual learning. This phenomena could exactly explain the main reason that many Vision-Language Models (VLMs), $e.g.$, CLIP (Radford et al. 2021), are flourishing and developing.

Aiming to break through pure vision modality, and motivated by VLMs, we propose a new semi-supervised ISTD scheme with target prior knowledge ($i.e.$, text prompts), as shown in Figure 1 (c). It adopts both classic vision modality and text modality to construct detectors, and utilizes the text prompts designed for targets to guide the generation and filtering of high-quality pseudo-labels, as well as the feature learning of infrared targets. Unlike fully-supervised schemes, it doesn't need many sample annotations, by adaptively generating high-quality pseudo-labels. The experiments on three public infrared datasets show that our scheme could even achieve comparable results to fully-supervised SOTA ones, with only about 10% labeled samples.

In summary, the primary contributions of our work include: **(i)** breaking through traditional pure-vision modality, the first semi-supervised vision-language learning framework with adaptive text prompt guiding is proposed for infrared small targets; **(ii)** a text-guided adaptive enhancing scheme to strengthen target regions is designed, with the cross-modal masking for further optimizing the interaction and alignment of vision-text modalities; **(iii)** a new filtering strategy with VLMs is developed to obtain high-quality pseudo-labels by adaptive dynamic thresholds.

## Related Work

### Moving Infrared Small Target Detection

Moving ISTD (MISTD) could often obtain extra motion features, enhancing their potential. For instance, motion estimation-based method (Zhao et al. 2020) uses optical flow to estimate the motion vector of each pixel. Moreover, some tensor-optimized methods (Wu et al. 2023; Luo, Li, and Chen 2024) construct spatio-temporal tensors by low-rank and sparse theories, achieving promising performance.

With the development of deep neural networks, multi-frame data-driven methods have advanced significantly (Duan et al. 2024). For example, ST-Trans (Tong et al. 2024) designs a spatio-temporal transformer to employ the motion information of consecutive frames. Recently, DTUM (Li et al. 2025) uses a direction-coded temporal U-shape module to encode the motion patterns. MoPKL (Chen et al. 2025) proposes a motion prior learning method with language descriptions. Unlike existing schemes, we break through the traditional fully-supervised learning and construct a multi-modal semi-supervised learning framework with the text description prompts of small targets.

### Semi-supervised Object Detection

Semi-supervised object detection (SSOD) aims to train detectors using limited labeled data and numerous unlabeled data. In early stages, STAC (Sohn et al. 2020) introduces weak and strong augmentations and generates pseudo-labels via a pre-trained teacher. After that, inspired by the Exponential Moving Average (EMA) from MeanTeacher (Tarvainen and Valpola 2017), many end-to-end schemes (Wang et al. 2023; Liu et al. 2023a,b) are proposed to streamline the complicated multi-stage training process. Recently, SSVOD (Mahmud et al. 2024) introduces a semi-supervised video object detection framework to exploit the temporal motion of videos with sparse annotations. However, existing SSOD methods often concern general objects in still images, rendering them ineffective in challenging MISTD cases.

### Vision-Language Models

With advancements in multi-modal learning, coupled with the support of web-scale image-text pairs, since the inception of CLIP (Radford et al. 2021), VLMs have achieved great success and are applied to a wide range of fields (Yi et al. 2024). For example, S-CLIP (Mo et al. 2023) proposes to train CLIP using additional unpaired images in the remote sensing field. Moreover, Flair (Xiao et al. 2025) proposes a fine-grained language-informed image learning with VLMs to capture detailed visual features. Given the good foundation and strong potential of VLMs, it is strongly anticipated that VLMs will be applied to MISTD.

## Proposed Method

### Overall Architecture

Our primary objective is to address the challenge of limited labeled data in MISTD, using text description prompts as target prior knowledge. Therefore, we propose a new semi-supervised vision-language framework, $i.e.$, SeViL, as shown in Figure 2. It follows a general paradigm of SSOD, $i.e.$, student-teacher network (Tarvainen and Valpola 2017), almost with the same architecture. Due to no labels for unlabeled data, in training the teacher has to generate the pseudo-labels, used as the supervision for the student. Besides, the text prompts are generated by the GPT4o and then adjusted to have the same pattern. Details are in **Appendix**.

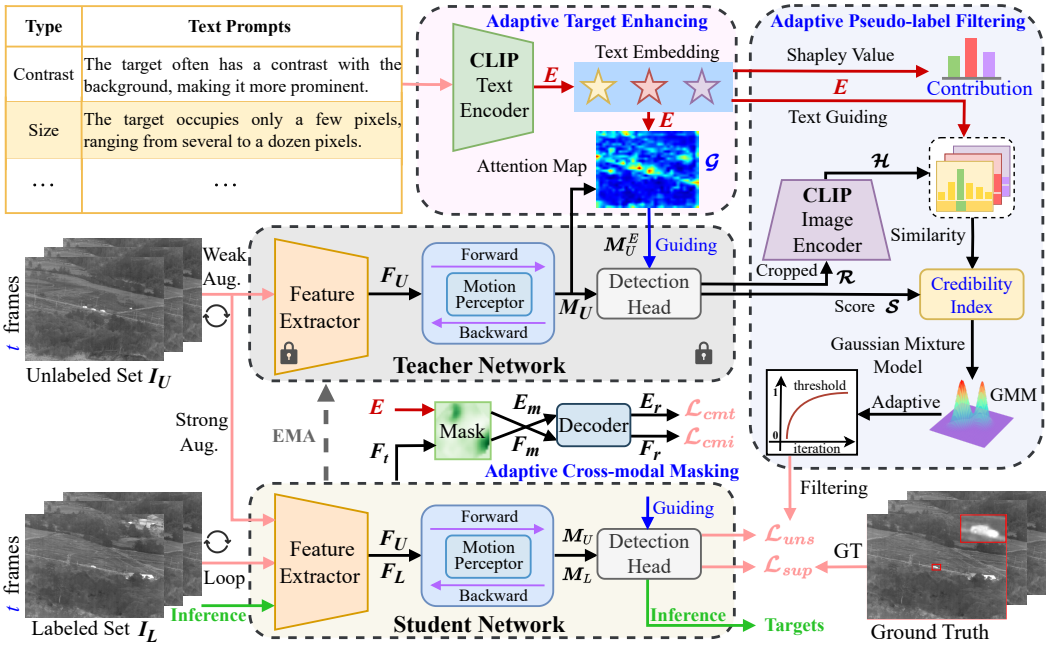


Figure 2: Our SeViL framework. It first obtains the text embedding $\boldsymbol{E}$ by a CLIP text encoder. New features $\boldsymbol{M_U^E}$ is obtained by **Adaptive Target Enhancing** to strengthen the target regions in $\boldsymbol{M_U}$. Then, $\boldsymbol{E}$ and cropped pseudo-label regions $\boldsymbol{\mathcal{R}}$ are fed into **Adaptive Pseudo-labels Filtering** to depress the noisy pseudo-labels generated by Teacher. After that, Student is trained on both the unlabeled set with high-quality pseudo-labels and labeled set. Finally, Teacher is updated by Student through EMA.

In vision branch, the $t$ frames respectively from labeled set $\boldsymbol{I_L}$ and unlabeled set $\boldsymbol{I_U}$ are used as training samples, with the ResNet50 (Lin et al. 2017) as visual feature extractors. Following typical video object detection methods (Zhou et al. 2022), we extract multi-frame features $\boldsymbol{F_L}, \boldsymbol{F_U} \in \mathbb{R}^{T\times C\times H\times W}$ by iteratively feeding each frame into extractors, where $C$, $H$ and $W$ are the channel, height and weight of features, respectively. Then, motion features $\boldsymbol{M_L}$ and $\boldsymbol{M_U}$ are obtained by a motion perceptor with bi-temporal modeling. Moreover, we transform designed text prompts into text embedding $\boldsymbol{E} \in \mathbb{R}^{N\times D}$ by a pretrained CLIP text encoder (Radford et al. 2021), where $N$ is text prompt number and $D$ denotes embedding dimension. Besides, $\boldsymbol{M_L}$, $\boldsymbol{M_U}$ and $\boldsymbol{E}$ are processed in *Adaptive Target Enhancing* to strengthen potential target regions by using text prompts as guidance signals to obtain an attention map $\boldsymbol{\mathcal{G}}$. This process produces the enhanced feature sets: $\boldsymbol{M_L^E}$ and $\boldsymbol{M_U^E}$. Furthermore, *Adaptive Pseudo-labels Filtering* is designed to improve pseudo-label quality by calculating the similarity between the cropped pseudo-label regions $\boldsymbol{\mathcal{R}}$ and text embedding $\boldsymbol{E}$. These pseudo-labels are predicted by the Teacher on unlabeled data. Moreover, we present *Adaptive Cross-modal Masking* to align text and vision modalities, facilitating the deep interaction of cross-modal features. After semi-supervised vision-language training with text-guiding, only the trained Student Network is used for inferring.

## Text-guided Adaptive Target Enhancing

To guide the detector to focus on potential target regions by text prompts, we propose a **Text-guided Adaptive Target Enhancing (TATE)** to optimize target features, as shown in Figure 2. First, the text embedding $\boldsymbol{E}$ is projected into the same semantic subspace as motion features $\boldsymbol{M}$ (including $\boldsymbol{M_L}$ and $\boldsymbol{M_U}$) via simple linear layers. Then, global text prior is generated by averaging all text embedding and expanded to obtain the final text embedding $\boldsymbol{E_p}$. This calculation process can be formulated as follows:

$$\boldsymbol{E_p} = f_{exp}(\frac{1}{N}\sum_{i=1}^{N} f_{mlp}(\boldsymbol{E_i})), \tag{1}$$

where $f_{mlp}(\cdot)$ denotes two linear layers and $f_{exp}(\cdot)$ is an expansion operation to make features have same dimensions. Here, $N$ indicates the total number of text prompts.

Second, in this way, we could calculate the attention map $\boldsymbol{\mathcal{G}}$ between text embedding $\boldsymbol{E_p}$ and motion features $\boldsymbol{M}$ pixel-by-pixel across all frames, as follows:

$$\boldsymbol{\mathcal{G}} = Softmax(f_c(\boldsymbol{M} \odot \boldsymbol{E_p})), \tag{2}$$

where "$\odot$" is an element-wise product and $f_c(\cdot)$ denotes a convolution operation. "$Softmax(\cdot)$" is used to transform original attention scores into normalized weights to adaptively adjust the importance to different regions.

Finally, the attention map $\boldsymbol{\mathcal{G}}$ is used for the pixel-by-pixel weighting on motion features $\boldsymbol{M}$. Then, it is further fused by a residual mechanism, as follows:

$$\boldsymbol{M^E} = f_c(\boldsymbol{\mathcal{G}} \odot \boldsymbol{M} + \boldsymbol{M}), \tag{3}$$

where $\boldsymbol{M^E}$ exactly represents the enhanced features guided through the semantic target prior knowledge $\boldsymbol{E}$.

## Text-guided Adaptive Pseudo-label Filtering

In SSOD, the quality of pseudo-labels would often critically impact the feature learning on unlabeled data. Therefore, we present a **Text-guided Adaptive Pseudo-label Filtering (TAPF)** to depress possible noisy pseudo-labels by using text prompts as the prior knowledge, adapting to target properties, as shown in Figure 2.

First, for each pseudo-label $p_i = (x_l, y_l, x_r, y_r)$, we crop the corresponding region $\mathcal{R}_i$ from the unlabeled keyframe $U_t$. Then, we feed cropped regions into the CLIP image encoder $f_{ie}(\cdot)$, described by following equations:

$$\boldsymbol{\mathcal{H}} = \sum_{i=1}^{m} f_{ie}(Crop(p_i, U_t)) = \sum_{i=1}^{m} f_{ie}(\mathcal{R}_i), \tag{4}$$

where $\boldsymbol{\mathcal{H}}$ is the visual feature set of all cropped regions and $m$ is the number of pseudo-labels.

Second, to enhance semantic understanding by text prompts, the similarity $\boldsymbol{\mathcal{K}}$ between vision features $\boldsymbol{\mathcal{H}}$ and text embedding $\boldsymbol{E}$ is calculated, as follows:

$$\boldsymbol{\mathcal{K}} = f_{cos}(\boldsymbol{\mathcal{H}}, \boldsymbol{E}) = \sum_{i=1}^{m}\sum_{j=1}^{N} \frac{\mathcal{H}_i \bullet E_j}{\|\mathcal{H}_i\|_2 \, \|E_j\|_2}, \tag{5}$$

where $f_{cos}(\cdot)$ denotes calculating cosine similarity, "$\bullet$" means inner product and $\|\cdot\|_2$ represents $l_2$ norm. Due to that confidence scores $\boldsymbol{\mathcal{S}}$ provides local credibility and text embedding $\boldsymbol{E}$ contains global semantic constraints, combining them could obtain more reliable and semantically consistent pseudo-labels. As such, we propose a new filtering metric, **Credibility Index** ($\mathcal{CI}$), as follows:

$$\mathcal{CI} = \frac{2 \times \boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{K}}}{\boldsymbol{\mathcal{S}} + \boldsymbol{\mathcal{K}}}, \tag{6}$$

Considering that the distribution of $\mathcal{CI}$ is multi-modal and a static threshold is difficult to accurately separate noise, we expect finding a way to automatically distinguish the positive ones from all pseudo-labels. In detail, we adopt a Gaussian Mixture Model (GMM) to model positive and negative pseudo-labels. It could be formulated as follows:

$$\mathcal{D}(\mathcal{CI}) = \pi_n \boldsymbol{\mathcal{N}}(\mathcal{CI} \mid \mu_n, \sigma_n^2) + \pi_p \boldsymbol{\mathcal{N}}(\mathcal{CI} \mid \mu_p, \sigma_p^2), \tag{7}$$

where $\boldsymbol{\mathcal{N}}(\mu, \sigma^2)$ represents a Gaussian distribution. Besides, $\pi_p, \mu_p, \sigma_p^2$ and $\pi_n, \mu_n, \sigma_n^2$ denote the weight, mean and variance of positive and negative pseudo-labels, respectively. After that, we employ an Expectation Maximization (EM) algorithm to infer the posterior $\mathcal{D}(pos \mid \mathcal{CI}, \mu_p, \sigma^2)$ and update GMM parameters. The final adaptive dynamic threshold $\tau$ is obtained by maximizing the posterior, as follows:

$$\tau = \underset{\mathcal{CI}}{\operatorname{argmax}}\, \mathcal{D}(pos \mid \mathcal{CI}, \mu_p, \sigma_p^2), \tag{8}$$

## Adaptive Cross-modal Masking

To enhance the interaction between vision and text modalities, we design an **Adaptive Cross-modal Masking (ACM)** strategy. It reconstructs the inherent signals of one modality by a randomly masked input, while keeping the other modality unmasked. In detail, it contains Cross-modal Masked Image Reconstructing (CMIR) and Cross-modal Masked Text Reconstructing (CMTR) to facilitate vision-text interaction.

Taking CMIR as an example, inspired by MAE (He et al. 2022), we randomly mask the key-frame features $\boldsymbol{F_t}$ and text embedding $\boldsymbol{E}$ to obtain masked visual features $\boldsymbol{F_m}$ and masked text embedding $\boldsymbol{E_m}$, as follows:

$$\begin{aligned} \boldsymbol{\Theta}(x,y) &= \begin{cases} 1, & \text{if } p < \eta, \\ 0, & \text{otherwise}, \end{cases} \quad p \sim U(0,1), \\ \boldsymbol{F_m} &= \boldsymbol{F_t} \odot (1 - \boldsymbol{\Theta}), \end{aligned} \tag{9}$$

where $\boldsymbol{\Theta}$ is mask matrix, $p$ denotes a random variable sampled from a uniform distribution $U(0,1)$ to determine whether the position $(x, y)$ is masked, and $\eta$ is used to control the masking ratio. Then, a cross-modal image decoder $f_{id}(\cdot)$, consisting of two-layer convolution blocks, is designed to reconstruct the original image features $\boldsymbol{F_r}$ on masked image $\boldsymbol{F_m}$ and original text embedding $\boldsymbol{E}$. Finally, we boost the complementary learning of cross-modal features by computing the discrepancy between $\boldsymbol{F_r}$ and original visual features $\boldsymbol{F_t}$, as follows:

$$\mathcal{L}_{cmi} = \frac{1}{\Psi(\boldsymbol{F_t})} \left\| \boldsymbol{F_t} - f_{id}(\boldsymbol{F_m}, \boldsymbol{E}) \right\|_{mse}, \tag{10}$$

where $\Psi(\cdot)$ represents the number of feature points and the loss function $\mathcal{L}_{cmi}$ is based on the mean squared loss.

Similar to CMIR, given the representation of masked text embedding $\boldsymbol{E_m}$ and original visual features $\boldsymbol{F_t}$, we employ the $l_2$ loss to measure the distance between reconstructed and original text embedding, $i.e.$, $\boldsymbol{E_r}$ and $\boldsymbol{E}$. Therefore, the objective of CMTR could be denoted by

$$\mathcal{L}_{cmt} = \frac{1}{\Psi(\boldsymbol{E})} \left\| \boldsymbol{E} - f_{td}(\boldsymbol{E_m}, \boldsymbol{F_t}) \right\|_{mse}, \tag{11}$$

where $f_{td}(\cdot)$ is a cross-modal text embedding decoder, including a self-attention layer and residual connections. The total loss of ACM could be denoted by $\mathcal{L}_{cmm} = \mathcal{L}_{cmi} + \mathcal{L}_{cmt}$. By minimizing $\mathcal{L}_{cmm}$, the model is guided to reconstruct the original features through cross-modal interaction. This process effectively facilitates the exploration of deeper relations that exist between vision and text modalities.

## Loss Function

The total training loss of our proposed SeViL could be described as follows:

$$\mathcal{L} = \mathcal{L}_{sup} + \alpha \mathcal{L}_{uns} + \beta \mathcal{L}_{cmm}, \tag{12}$$

where $\alpha$ and $\beta$ are two weight coefficients to balance loss terms, $\mathcal{L}_{sup}$ and $\mathcal{L}_{uns}$ are the detection loss based on labeled and unlabeled data. Specifically, we use the task aligned detection head of TOOD (Feng et al. 2021). As such, $\mathcal{L}_{sup}$ and $\mathcal{L}_{uns}$ could be denoted as follows:

$$\mathcal{L}_{sup} = \mathcal{L}_{uns} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{reg}, \tag{13}$$

where $\lambda_1$ and $\lambda_2$ are two hyper-parameters. Following the default settings of TOOD, we employ Focal loss as the classification loss $\mathcal{L}_{cls}$, GIoU as the localization loss $\mathcal{L}_{reg}$.

| Scheme | Methods | Publication | DAUB | | | | ITSDT-15K | | | | IRDST | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $mAP_{50}$ | Pr | Re | F1 | $mAP_{50}$ | Pr | Re | F1 | $mAP_{50}$ | Pr | Re | F1 |
| Model-driven | RLCM | TGRS 2013 | 0.02 | 0.27 | 5.21 | 0.51 | 4.62 | 15.38 | 30.76 | 20.50 | 1.58 | 16.28 | 9.70 | 12.16 |
| | HBMLCM | GRSL 2019 | 3.90 | 23.96 | 16.52 | 19.56 | 0.72 | 7.97 | 9.37 | 8.61 | 1.16 | 29.14 | 4.66 | 8.03 |
| | PSTNN | RS 2019 | 17.31 | 25.56 | 68.86 | 37.28 | 7.99 | 22.98 | 35.21 | 27.81 | 1.45 | 16.28 | 9.70 | 12.16 |
| | WSLCM | SP 2020 | 1.37 | 11.88 | 11.57 | 11.73 | 2.36 | 16.78 | 14.53 | 15.58 | 1.69 | 20.87 | 8.70 | 12.28 |
| Fully-supervised | ACM | WACV 2021 | 64.02 | 70.96 | 91.30 | 79.86 | 55.38 | 78.37 | 71.69 | 74.88 | 52.40 | 76.33 | 69.32 | 72.66 |
| | ISNet | CVPR 2022 | 83.43 | 89.36 | 94.99 | 92.09 | 62.29 | 83.46 | 75.32 | 79.18 | 59.78 | 80.24 | 75.08 | 77.58 |
| | UIUNet | TIP 2022 | 86.41 | 94.46 | 92.03 | 93.23 | 65.15 | 84.07 | 78.39 | 81.13 | 56.38 | 80.95 | 70.29 | 75.25 |
| | DNANet | TIP 2023 | 89.93 | 92.49 | 98.27 | 95.29 | 70.46 | 88.55 | 80.73 | 84.46 | 63.61 | 82.92 | 77.48 | 80.11 |
| | SIRST5K | TGRS 2024 | 93.31 | 97.78 | 96.93 | 97.35 | 61.52 | 86.95 | 71.32 | 78.36 | 52.28 | 76.12 | 69.07 | 72.42 |
| | MSHNet | CVPR 2024 | 85.97 | 93.13 | 93.12 | 93.13 | 60.82 | 89.69 | 68.44 | 77.64 | 63.21 | 82.31 | 77.64 | 79.91 |
| | RPCANet | WACV 2024 | 85.98 | 89.38 | 97.56 | 93.29 | 62.28 | 81.46 | 77.10 | 79.22 | 56.50 | 77.77 | 73.80 | 75.73 |
| | ST-Trans | TGRS 2024 | 92.73 | 97.75 | 95.52 | 96.62 | 76.02 | 89.96 | 85.18 | 87.50 | 70.04 | 88.21 | 80.01 | 83.91 |
| | MLPNet | TGRS 2025 | 93.58 | 97.08 | 97.89 | 97.49 | 53.76 | 74.06 | 73.19 | 73.63 | 57.48 | 80.36 | 72.14 | 76.03 |
| | LSKNet | TGRS 2025 | 84.66 | 88.84 | 96.04 | 92.30 | 66.07 | 90.75 | 73.72 | 81.35 | 61.19 | 77.85 | 79.98 | 78.90 |
| | PConv | AAAI 2025 | 90.84 | 95.32 | 96.54 | 95.93 | 76.02 | 89.96 | 85.18 | 87.50 | 66.43 | 88.93 | 75.54 | 81.69 |
| | DTUM | TNNLS 2025 | 85.86 | 87.54 | **99.79** | 93.26 | 67.97 | 77.95 | 88.28 | 82.79 | 71.48 | 82.87 | 87.79 | 85.26 |
| | MoPKL | AAAI 2025 | 94.85 | 97.83 | 98.79 | 98.31 | 79.78 | **93.29** | 86.80 | 89.92 | 74.54 | **89.04** | 84.74 | 86.84 |
| Semi-supervised | SoftTeacher | ICCV 2021 | 73.17 | **98.76** | 74.79 | 85.12 | 60.95 | 69.26 | 84.94 | 76.30 | 59.14 | 77.45 | 81.71 | 75.66 |
| | PseCo | ECCV 2022 | 83.97 | 89.44 | 94.83 | 92.06 | 66.76 | 86.95 | 77.79 | 82.12 | Training Without Convergence | | | |
| | LabelMatch | CVPR 2022 | 66.78 | 70.12 | 96.1 | 81.08 | 38.61 | 62.74 | 62.23 | 62.48 | Training Without Convergence | | | |
| | MixTeacher | CVPR 2023 | 79.54 | 81.90 | 98.79 | 89.55 | 62.67 | 68.69 | **92.41** | 78.80 | 61.92 | 71.60 | 84.15 | 76.19 |
| | ConsistentTeacher | CVPR 2023 | 85.71 | 93.57 | 92.95 | 93.26 | 66.02 | 85.79 | 77.23 | 81.29 | 63.77 | 74.95 | 85.04 | 79.67 |
| | Semi-DETR | CVPR 2023 | 87.39 | 92.53 | 95.50 | 93.99 | 70.01 | 82.13 | 85.35 | 83.71 | 64.79 | 80.66 | 81.13 | 80.89 |
| | SSVOD | WACV 2024 | 89.31 | 93.79 | 95.37 | 94.57 | 72.46 | 83.09 | 86.12 | 84.58 | 65.55 | 80.50 | 82.80 | 81.63 |
| | S2MVP | TGRS 2025 | 93.84 | 97.08 | 98.42 | 97.75 | 78.17 | 88.49 | 88.88 | 88.69 | 72.15 | 85.36 | 86.87 | 86.11 |
| | **SeViL (Ours)** | - | **95.13** | 98.14 | 99.19 | **98.66** | **80.18** | 88.96 | 91.46 | **90.20** | **74.62** | 86.53 | **87.82** | **87.17** |

Table 1: Quantitative comparisons. The best and second-best results are highlighted in bold and underlined, respectively. All semi-supervised methods **only use 10%** labeled training samples, while fully-supervised ones use 100%.

# Experiments

## Implementation Details

Our SeViL is evaluated on three datasets, DAUB (Hui et al. 2019), ITSDT-15K (Fu et al. 2022) and IRDST (Sun et al. 2023). For a fair comparison, we use the standard evaluation metrics (Chen et al. 2024), $i.e.$, Precision ($Pr$), Recall ($Re$), F1 and $mAP_{50}$ (the mean Average Precision with an IoU threshold 0.5). For all compared methods, their input frames are resized to $512 \times 512$. In detail, our SeViL is trained for 50K iterations with a batch size 4 (3 unlabeled frames and 1 labeled one). The SGD optimizer with a momentum of 0.9 is adopted, and initial learning rate is 0.001, with a weight decay of $1 \times 10^{-4}$. The Teacher is updated through the EMA with a momentum of 0.9995. Hyper-parameters $t$, $\eta$, $\alpha$, $\beta$, $\lambda_1$ and $\lambda_2$ are set to 5, 0.25, 2, 1, 2 and 1, respectively.

## Comparisons with SOTA Methods

**Quantitative Comparisons** Table 1 shows the quantitative comparisons on 27 representative methods, covering fully-supervised and semi-supervised ones. In table, we could observe that our SeViL significantly outperforms all other semi-supervised ones, and achieves nearly the equivalent performance to fully-supervised ones, on most metrics.

For example, on DAUB, our SeViL achieves the highest $mAP_{50}$ 95.13% and F1 98.66%. In terms of $Pr$ and $Re$, the 98.14% and 99.19% by SeViL is slightly lower than the SOTA 98.76% by SoftTeacher (Xu et al. 2021) and 99.79% by DTUM. However, SoftTeacher only obtains the $Re$ 74.79% and DTUM only acquires the $Pr$ 87.54%. They achieve higher $Pr$ or $Re$ by sacrificing the other one, while our method is more balanced with a higher F1. Moreover, on ITSDT-15K, our SeViL reaches the peak $mAP_{50}$ 80.18% and F1 90.20%. Besides, on IRDST, due to some semi-supervised schemes primarily focusing on static objects, they cannot often converge in training, such as PseCo (Li et al. 2022b) and LabelMatch (Chen et al. 2022).

| 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 |
| 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 |
| RPCANet | 1 | 62.28 | 79.22 | 3.21M | 382.69 | 15.89 |
| MLPNet | 1 | 53.76 | 73.63 | 10.79M | 34.72 | 5.93 |
| ST-Trans | 5 | 76.02 | 87.50 | 38.13M | 145.16 | 3.90 |
| DTUM | 5 | 67.97 | 82.79 | 9.64M | 128.16 | 14.28 |
| MoPKL | 5 | 79.78 | 89.92 | 9.46M | 119.64 | 10.03 |
| MixTeacher | 1 | 62.67 | 78.80 | 41.10M | 202.57 | 33.10 |
| ConsistentTeacher | 1 | 66.02 | 81.29 | 32.04M | 52.35 | **35.79** |
| SSVOD | 5 | 72.46 | 84.58 | 47.65M | 127.38 | 30.25 |
| S2MVP | 5 | 78.17 | 88.69 | 52.74M | 140.29 | 28.76 |
| **SeViL (Ours)** | 5 | **80.18** | **90.20** | 40.48M | 140.54 | 21.70 |

Table 2: Complexity comparisons on ITSDT-15K.

**Model Complexity Comparisons** The complexity comparisons on 16 methods are presented in Table 2, revealing **two** notable findings. **One** is that although our SeViL uses a sequence of five frames, the Params and GFlops only increase slightly. For example, it has 40.48M parameters, higher than that of most single-frame methods, but still lower than the 47.65M by SSVOD. Besides, its GFlops is 140.54, greatly lower than the 202.57 by MixTeacher. **The other** is that semi-supervised ones depend less on labeled data, and infer only by Student, leading to a high FPS. For example, SeViL has an FPS of 21.70, higher than those of many fully-supervised ones, $e.g.$, DTUM and MoPKL.

| Settings | Text | | Mask | | | Filtering | | | DAUB | | | | ITSDT-15K | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T1 | T2 | M1 | M2 | M3 | F1 | F2 | F3 | $mAP_{50}$ | Pr | Re | F1 | $mAP_{50}$ | Pr | Re | F1 |
| w/o All | - | - | - | - | - | - | - | - | 85.35 | 91.50 | 93.66 | 92.57 | 66.96 | 83.32 | 81.14 | 82.22 |
| w T | ✓ | - | - | - | - | - | - | - | 88.43 | 93.89 | 95.26 | 94.57 | 70.34 | 82.54 | 86.03 | 84.25 |
| | - | ✓ | - | - | - | - | - | - | 90.45 | 94.61 | 96.53 | 95.56 | 72.48 | 84.22 | 86.41 | 85.30 |
| w T & M | - | ✓ | ✓ | - | - | - | - | - | 91.21 | 94.52 | 97.33 | 95.91 | 73.32 | 85.56 | 86.72 | 86.14 |
| | - | ✓ | - | ✓ | - | - | - | - | 91.36 | 95.34 | 96.87 | 96.10 | 73.97 | 84.77 | 88.97 | 86.82 |
| | - | ✓ | - | - | ✓ | - | - | - | 92.41 | 94.95 | 98.16 | 96.53 | 75.56 | 86.93 | 88.75 | 87.83 |
| w T & M & F | - | ✓ | - | - | ✓ | ✓ | - | - | 92.90 | 96.55 | 97.02 | 96.78 | 76.79 | 87.02 | 89.14 | 88.07 |
| | - | ✓ | - | - | ✓ | - | ✓ | - | 93.46 | 96.50 | 97.63 | 97.06 | 77.32 | 87.49 | 88.87 | 88.17 |
| w All | - | ✓ | - | - | ✓ | - | - | ✓ | **95.13** | **98.14** | **99.19** | **98.66** | **80.18** | **88.96** | **91.46** | **90.20** |

Table 3: The ablation study of our SeViL with different settings on DAUB and ITSDT-15K. **Text**: two text prompt aligning schemes (**T1** aligns vision into text space, **T2** aligns text into vision space). **Mask**: three cross-modal masking schemes (**M1** only masks vision features, **M2** only masks text embedding, **M3** masks the both). **Filtering**: three pseudo-label filtering schemes (**F1** adopts text-vision similarity, **F2** uses ours Credibility Index, and **F3** uses our text-guided adaptive pseudo-label filtering).

**Visual Comparisons** For intuitiveness, we select 2 representative methods to visually compare our SeViL on three datasets, as depicted in Figure 3. It is evident that our method could usually precisely detect targets, while other two could often produce missed detections and false detections. For example, on ITSDT-15K, our SeViL could accurately detect all six targets, while MSHNet (Liu et al. 2024) only correctly detect three targets. On challenging IRDST, other two could not even detect the targets occluded by a building.

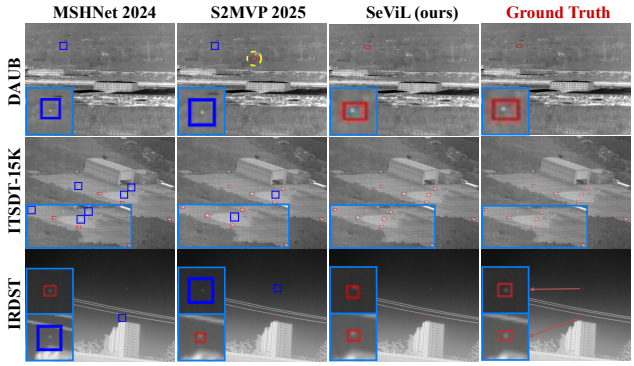


Figure 3: Visual comparisons to two representative methods. Blue boxes and yellow circles are the missed and false detections. More comparisons and PR curves are in **Appendix**.

## Ablation Study

**Effects of Different Components** We perform a series of ablation studies to investigate the impact of each component on our SeViL, as shown in Table 3. Through comparisons, we could have **two** obvious observations. **First**, each individual component is consistently effective. For instance, on DAUB, the baseline without any specialized components (w/o All), obtains the $mAP_{50}$ 85.35% and F1 92.57%. After integrating text prompt guiding (w T2), these metrics increase to 90.45% for $mAP_{50}$ and 95.56% for F1. The setting (w T2 & M3) further raises $mAP_{50}$ to 92.41% and F1 to 96.53%. Besides, credibility index (w T2 & M3 & F2) could also have an obvious gain with $mAP_{50}$ 93.46% and F1 97.06%. **The other** is that these components have synergistic effects. For example, on DAUB, when all components are fully assembled (w All), detection performance can reach a peak, with an $mAP_{50}$ 95.13% and an F1 98.66%.

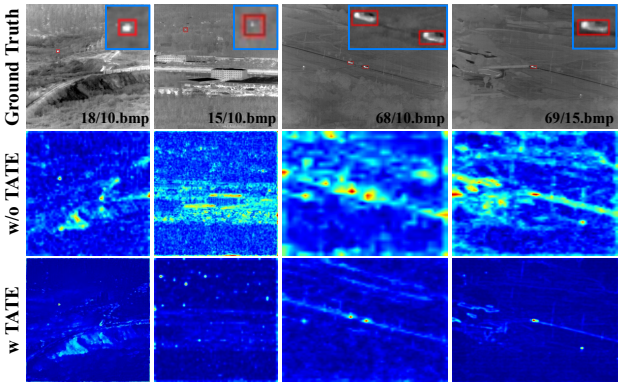


Figure 4: Feature heatmap comparisons. The first two columns are on DAUB, and the last two are on ITSDT-15K.

**Impacts of Adaptive Target Enhancing** To demonstrate the effectiveness of Text-guided Adaptive Target Enhancing (*i.e.*, TATE), we visualize the feature heatmaps with and without TATE, as illustrated in Figure 4. In this figure, it is obvious that the focus positions of feature heatmaps by "w/o TATE" are obscure and targets are even lost in complex backgrounds. In contrast, after employing TATE, the feature response of infrared small targets is significantly enhanced, in the meanwhile, the noisy background becomes clearer. This indicates that the text prompts could effectively match target regions, enhancing moving small targets.

**Impacts of Text Applying Methods** To effectively utilize target prior knowledge, we investigate different text applying methods, as shown in Table 4. From results, we could observe that the performance with "Learnable weights" is the lowest, the $mAP_{50}$ only 87.79% and F1 94.34% on DAUB. One possible reason is that "Learnable weights" does not

use any prior knowledge to focus on target features. In contrast, our text prompts play an important role in enhancing targets, achieving the optimal detection performance. This reveals that our text prompts could provide useful semantic guidance to help networks focus on target regions.

| Settings | DAUB | | ITSDT-15K | | Params |
|---|---|---|---|---|---|
| | $mAP_{50}$ | F1 | $mAP_{50}$ | F1 | |
| Concat (Fuse) | 92.49 | 96.65 | 76.02 | 87.77 | 40.61M |
| Self-attention | 92.86 | 96.93 | 76.74 | 87.99 | 42.75M |
| Learnable map $\mathcal{G}$ | 87.79 | 94.34 | 74.75 | 87.29 | 40.54M |
| Text-guiding | **95.13** | **98.66** | **80.18** | **90.20** | **40.48M** |

Table 4: Ablation study on different text applying schemes. **Concat:** concatenating text and vision features. **Self-attention:** getting target enhancing weights by self-attention on vision features. **Learnable map $\mathcal{G}$:** generating the attention map $\mathcal{G}$ without text guiding.

**Impacts of Cross-modal Masking** To visualize the effect of Adaptive Cross-modal Masking, we perform a group of feature distribution comparisons before and after ACM, as shown in Figure 5. From this, it could be observed that ACM could effectively align text and vision features on both datasets. These comparisons further confirm those quantitative results in Table 3 ($i.e.$, w T2 and w T2 & M3).

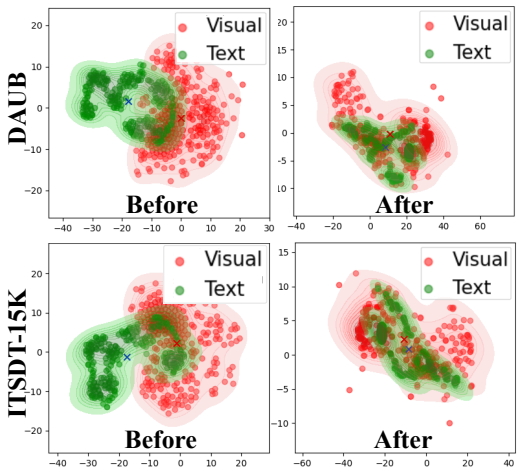


Figure 5: Feature distributions before and after ACM.

**Impacts of Pseudo-label Filtering** Figure 6 (a) shows the number of pseudo-labels per image in unlabeled set under different $\tau$. Notably, it reveals a crucial issue that the number of pseudo-labels keeps increasing in training for traditional static thresholds, $e.g.$, $\tau = 0.4$ and 0.6. In contrast, our proposed TAPF could adaptively adjust the optimal threshold according to the detector capability. Figure 6 (b) plots the threshold curves obtained by our TAPF on three datasets. We see these values will steadily increase with training iterations. Besides, on relatively simple DAUB, TAPF inclines to use a higher $\tau$ to avoid overfitting. It indicates that TAPF could provide a proper solution to filter low-quality pseudo-labels, while typical static thresholds could not.

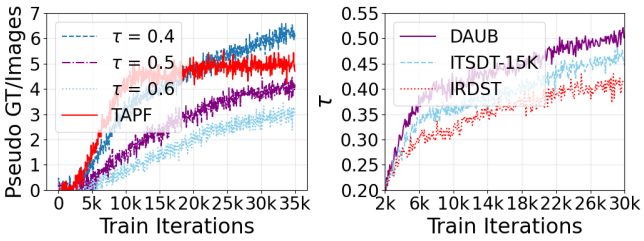


Figure 6: (a) Number of pseudo-labels/ image with different thresholds on DAUB, (b) Average thresholds by our TAPF.

**Impacts of Each Text Prompt** We explore the contribution of each text prompt (totally 13) based on a generalized model interpretation framework, $i.e.$, Shapley value (Shapley 1953) (its theory supports are given in **Appendix**). As shown in Figure 7, we can find that the contribution of each text prompt to the detection performance varies widely across datasets. For example, on DAUB, the text prompts of brightness (T1) and occlusion (T9) are the most influential, while on ITSDT-15K, trajectory (T4) is the most powerful. This highlights the scene dependence of text prompts, underscoring the importance of generating suitable text prompts.

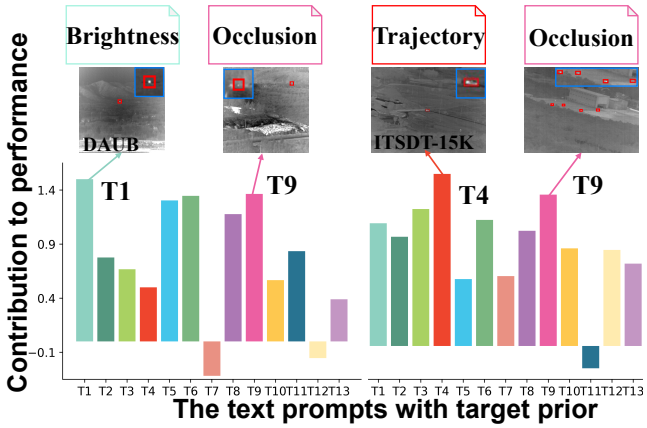


Figure 7: The contribution of each text prompt to detection.

## Conclusions

In this paper, we propose the first semi-supervised vision-language learning framework with adaptive text prompt guiding for MISTD, $i.e.$, SeViL. It could explore the potential of extensive unlabeled data by adaptively enhancing target regions and filtering low-quality pseudo-labels. Besides, cross-modal masking could further enhance the alignment learning of image-text, and focus on cross-modal relations. The experiments on three benchmarks verify the effectiveness and superiority of our SeViL. It surpasses other semi-supervised SOTA ones, even reaching fully-supervised ones. Its weakness is the large parameter number, leading to high model complexity. In the future, an optimized lightweight semi-supervised scheme with more efficient vision-language learning is worthy of further exploration.

## Acknowledgments

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

## 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.

Chen, B.; Chen, W.; Yang, S.; Xuan, Y.; Song, J.; Xie, D.; Pu, S.; Song, M.; and Zhuang, Y. 2022. Label matching semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14381–14390.

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.

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*.

Feng, C.; Zhong, Y.; Gao, Y.; Scott, M. R.; and Huang, W. 2021. Tood: Task-aligned one-stage object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3490–3499. IEEE Computer Society.

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.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

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.

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

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

Li, G.; Li, X.; Wang, Y.; Wu, Y.; Liang, D.; and Zhang, S. 2022b. Pseco: Pseudo labeling and consistency training for semi-supervised object detection. In *European Conference on Computer Vision*, 457–472. Springer.

Li, R.; An, W.; Xiao, C.; Li, B.; Wang, Y.; Li, M.; and Guo, Y. 2025. Direction-Coded Temporal U-Shape Module for Multiframe Infrared Small Target Detection. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1): 555–568.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117–2125.

Liu, C.; Zhang, W.; Lin, X.; Zhang, W.; Tan, X.; Han, J.; Li, X.; Ding, E.; and Wang, J. 2023a. Ambiguity-resistant semi-supervised learning for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15579–15588.

Liu, L.; Zhang, B.; Zhang, J.; Zhang, W.; Gan, Z.; Tian, G.; Zhu, W.; Wang, Y.; and Wang, C. 2023b. Mixteacher: Mining promising labels with mixed scale teacher for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7370–7379.

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*.

Luo, Y.; Li, X.; and Chen, S. 2024. 5-D spatial-temporal information-based infrared small target detection in complex environments. *Pattern Recognition*, 111003.

Mahmud, T.; Liu, C.-H.; Yaman, B.; and Marculescu, D. 2024. SSVOD: Semi-supervised video object detection with sparse annotations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6773–6782.

Mo, S.; Kim, M.; Lee, K.; and Shin, J. 2023. S-Clip: Semi-supervised vision-language learning using few specialist captions. *Advances in Neural Information Processing Systems*, 36: 61187–61212.

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.

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.

Shapley, L. S. 1953. A value for n-person games. *Contribution to the Theory of Games*, 2.

Sohn, K.; Zhang, Z.; Li, C.-L.; Zhang, H.; Lee, C.-Y.; and Pfister, T. 2020. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*.

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.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30.

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, X.; Yang, X.; Zhang, S.; Li, Y.; Feng, L.; Fang, S.; Lyu, C.; Chen, K.; and Zhang, W. 2023. Consistent-teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3240–3249.

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*.

Xiao, R.; Kim, S.; Georgescu, M.-I.; Akata, Z.; and Alaniz, S. 2025. Flair: Vlm with fine-grained language-informed image representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24884–24894.

Xu, M.; Zhang, Z.; Hu, H.; Wang, J.; Wang, L.; Wei, F.; Bai, X.; and Liu, Z. 2021. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3060–3069.

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

Yi, X.; Xu, H.; Zhang, H.; Tang, L.; and Ma, J. 2024. Text-IF: Leveraging Semantic Text Guidance for Degradation-Aware and Interactive Image Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27026–27035.

Zhao, F.; Wang, T.; Shao, S.; Zhang, E.; and Lin, G. 2020. Infrared Moving Small-Target Detection via Spatiotemporal Consistency of Trajectory Points. *IEEE Geoscience and Remote Sensing Letters*, 17(1): 122–126.

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.