

CHAL: Causal-guided Hierarchical Anomaly-aware Learning for Moving Infrared Small Target Detection

Weiwei Duan, Luping Ji*, Shipeng Lei, Sicheng Zhu, Jianghong Huang, and Mao Ye
 School of Computer Science and Engineering, University of Electronic Science and Technology of China
 dwuestc@163.com, jiluping@uestc.edu.cn, cvlab.uestc@gmail.com

Abstract

Infrared small target detection is one highly special category of object detection, faced with tiny target imaging size and cluttered backgrounds. Currently, almost all existing methods are target-centered, directly learning the target features from backgrounds. However, due to weak target signals, they are often difficult in effectively capturing stable features. Sometimes, they cannot even distinguish real targets from background confounders. To overcome these problems, from an opposite perspective, we propose the first Causal-guided Hierarchical Anomaly-aware Learning (CHAL) framework. Breaking through target-centered paradigm, it focuses on background learning, while the targets are handled as the anomalies in backgrounds. In detail, to fulfill the goal, a spatio-temporal neural field is designed to model the background evolution patterns from generative perspective. Meanwhile, a hierarchical anomaly-aware learning is proposed to decompose anomaly discovery. Furthermore, to block the spurious correlations often caused by background confounders, and enhance true target causality, a causal-guiding mechanism is designed. The experiments on three infrared datasets verify the effectiveness and superiority of our CHAL. Even in visible-light scenarios, it still possesses obvious adaptivity. Code is open at <https://github.com/UESTC-nnLab/CHAL>.

1. Introduction

Infrared small target detection (ISTD) has become a crucial task in computer vision due to its independence from external light sources and its all-weather visibility capability. It has been increasingly applied to a wide range of critical applications [8], including early warning systems, satellite remote sensing, and rescue missions [1, 4]. As an important research branch of object detection, it has garnered significant research attention over the past decades [13, 22].

Compared with conventional visible-light objects [10],

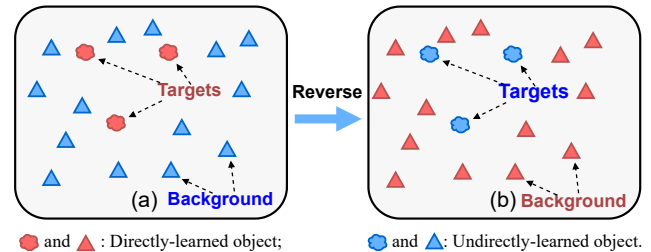


Figure 1. Paradigm comparisons: (a) target-centered and (b) ours.

due to the long-range imaging mechanism, infrared small targets usually show two special vision characteristics: *Small* (image size) and *Dim* (background contrast), often lacking distinct shapes and texture features.

In early stages, many traditional model-driven methods were first proposed to address ISTD, *e.g.*, TopHat [1] and PSTNN [33]. This type of methods often relies heavily on the prior knowledge of infrared images, lacking learning capabilities. To overcome their obvious shortcomings, with the progress of deep learning, many effective data-driven methods have been proposed. Depending on the frame number to be processed, they can be further grouped into *Single-frame* and *Multi-frame* [9]. The former usually rely on the vision features of only an individual image, no motion information is available, *e.g.*, ISNet [43], MSHNet [22] and PConv [39]. By contrast, multi-frame schemes could often extract the visual and motion patterns of small targets from consecutive frames, which have become mainstream, *e.g.*, SSTNet [2], Tridos [7] and DTUM [19].

Despite advances, almost all existing data-driven methods are primarily based on target-centered feature learning, as shown in Figure 1 (a). This paradigm faces a fundamental contradiction, *i.e.*, it often depends on rich target features (target-centered), which are precisely lacking in infrared small targets. As such, the detector is highly susceptible to confounding effects, *e.g.*, when a weak target is near bright cloud edges, the model can easily learn stronger background confounders rather than dim targets, becoming a confounder detector and resulting in false alarms.

*Corresponding Author

This critical challenge motivates us to reverse the existing paradigm from weak targets to dominant backgrounds, *i.e.*, background-centered feature learning, as shown in Figure 1 (b). Specifically, we reformulate moving ISTD (MISTD) as an anomaly-discovery task, treating small targets as the anomalies in backgrounds. However, one key challenge in this task is distinguishing true anomalies (targets) from spurious anomalies (background confounders) that also deviate from the learned normality (backgrounds). Traditional background model methods, *e.g.*, PSTNN [33] and FGLR-MCP [23], are insufficient due to their heavy reliance on rigid infrared image priors.

Currently, causal learning [21] has emerged as a powerful tool for addressing confounding effects. Prior works [24, 25] often use interventions to decouple high-level semantics. However, the confounders of MISTD are non-semantic (*e.g.*, textures, edges) and dynamically evolving, making existing causal designs unsuitable for direct application. Inspired by these analyses, we propose a *Causal-guided Hierarchical Anomaly-aware Learning (CHAL)* framework. It first models the background normality and then causally perceives the true source of deviations.

In our framework, it mainly comprises three core parts. First, the *Spatio-Temporal Neural Fields* serve as a normality estimator to learn the continuous spatio-temporal evolution of background patterns. Second, the *Hierarchical Anomaly-Aware Learning* breaks down anomaly discovery into a hierarchical process: mining potential appearance anomalies first and then verifying them via motion consistency. Finally, the *Causal Relation Guiding* is proposed to address the interference of confounders, blocking spurious correlations and enhancing true causality. The extensive experiments on three public infrared datasets verify the effectiveness and superiority of our CHAL. Even in visible-light scenarios, it still possesses obvious adaptivity.

In summary, our primary contributions include: (i) breaking through existing target-centered feature learning, the first causal-guided hierarchical anomaly-aware framework is proposed. (ii) rather than focusing on weak targets, a spatio-temporal neural field is proposed to perform background-centered feature learning, establishing a solid foundation for anomaly discovery. (iii) a hierarchical anomaly-aware learning strategy is developed to disentangle appearance and motion anomalies, achieving from discovery to verification. (iv) a causal relation guiding mechanism is presented to perform feature intervention, blocking spurious correlations and enhancing true causality.

2. Related Work

Infrared Small Target Detection. According to the differences in research strategies, current ISTD methods can be categorized into two types: *Single-frame* and *Multi-frame* [46]. Single-frame schemes mainly focus on ana-

lyzing individual frames, as they have no available motion information, *e.g.*, ACM [4], MSHNet [22], MTML-Net [38] and PConv [39]. In contrast, multi-frame schemes [7, 19, 32, 42] can simultaneously capture the visual features and motion patterns of targets, providing them with greater detection performance [9]. For example, Tridos [7] expands the pattern of feature learning from both spatial-temporal and frequency domains. Furthermore, DTUM [19] uses a direction-coded temporal U-shape module to encode the motion direction of targets.

Anomaly Detection. Anomaly Detection (AD) aims to detect abnormal samples that deviate from the normal patterns [15]. Recently, it has gained significant attention in various applications [28], *e.g.*, industrial manufacturing [26, 41] and video surveillance [44]. For example, DiffusionAD [41] designs a norm-guided denoising diffusion network, comprising reconstruction and segmentation. Moreover, MA-PDM [44] develops a patch diffusion model with motion and appearance guiding for video AD. However, existing methods often concern the visible images, making them ineffective in challenging infrared scenes.

Causal Reasoning. Causal learning [14] is a theoretical framework for studying causal relationships in data. It aims to remove spurious correlations by explicitly modeling confounding factors and designing adjustment strategies, and has been applied to various fields [3, 24, 25]. For example, UFR [24] designs an unbiased Faster R-CNN with generalizable feature learning to construct a structural causal model. Furthermore, CRCL [25] introduces the causal representation consistency learning for AD in surveillance videos, mining robust causal variables. Given the solid foundation and potential of causal learning, it is highly anticipated that it could be effectively applied to MISTD.

3. Methodology

3.1. Overall Architecture

Problem Formulation. Unlike prior works, we reformulate MISTD as an anomaly discovery task, rather than focusing on target features. Instead of directly classifying pixels as targets or backgrounds, we aim to analyze regions that deviate from the background’s spatio-temporal evolution patterns. Formally, given infrared video sequences $\mathcal{D}$, a set of consecutive frames within a time window t , *i.e.*, $\mathbf{I} = \{\mathbf{I}_1, \mathbf{I}_2, \dots, \mathbf{I}_t\}$, is randomly sampled as inputs. Our primary goal is to train an anomaly detector $f_\theta(\cdot)$ to locate the anomalies $\mathbf{Y}$ in the keyframe $\mathbf{I}_t$, using the temporal context of adjacent frames, that is

$$\min_{\theta} \mathbb{E}_{(\mathbf{I}, \mathbf{Y}) \sim \mathcal{D}} [\mathcal{L}_{total}(f_\theta(\mathbf{I}), \mathbf{Y})], \quad (1)$$

where $\mathcal{L}_{total}$ is the total training loss, and $\mathbb{E}_{(\mathbf{I}, \mathbf{Y}) \sim \mathcal{D}}$ denotes computing the expectation over all samples.

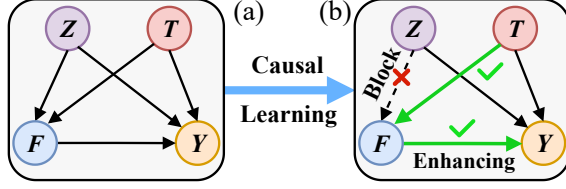


Figure 2. (a) Structural causal model; (b) Causal learning results.

Core Challenges. As illustrated in Figure 2, we construct a causal graph to explore the relations between different variables in the anomaly discovery process. In figure, the path $T \rightarrow Y$ denotes that the target features T directly influence the detection results Y . As such, the core challenges of our task primarily lie in the confounding path $F \leftarrow Z \rightarrow Y$. In detail, Z often represents complex background clutter, bright cloud edges, or sensor noise. It can directly corrupt the feature learning process ($F \leftarrow Z$) while simultaneously generating false alarms ($Z \rightarrow Y$). Therefore, our **core insight** is to distinguish spurious anomalies by causal learning, blocking the confounding path and enhancing true causality ($T \rightarrow F \rightarrow Y$).

Overview. The proposed causal-guided anomaly-aware learning framework, *i.e.*, **CHAL**, is shown in Figure 3, comprising three core components: SNF, HAL and CRG. Specifically, it first takes an infrared video clip I as input. A CSPDarknet [12] is utilized as the shared feature extractor to capture multiscale features by FPN. Following most video object detection methods [45], the multi-frame features $F_C \in \mathbb{R}^{b \times t \times c \times h \times w}$ are extracted by iteratively feeding each frame into the feature extractor, where b, c, h and w are the batch size, the channel, height and width of features, respectively. Second, scene encoding features Z and position encoding features Q are obtained by projecting F_C into a semantic subspace and constructing a 3D spatio-temporal grid $\mathcal{G}$. These two features, Z and Q , represent the scene context and high-frequency position details, processed by the background neural field $\mathcal{F}$ to generate predicted backgrounds $\mathcal{B}$. Then, $\mathcal{B}$ are fed into HAL. It performs anomaly perception by mining appearance anomaly candidates $\mathcal{A}_a$ and verifying real anomalies $\mathcal{A}_f$ by motion consistency. Moreover, CRG performs causal intervention and suppression for confounders in the feature space to obtain features F_f . Finally, F_f is fed into a detection head to compute the total training loss $\mathcal{L}_{total}$, thereby providing implicit supervision by backpropagation for the overall framework.

3.2. Spatio-Temporal Neural Fields

In our task, a key challenge is modeling complex and evolving infrared backgrounds. Previous methods [29, 32] often struggle to capture the continuous spatio-temporal properties of backgrounds. To address this, we develop SNF to implicitly learn spatio-temporal background evolution patterns via neural field modeling, as shown in Figure 3.

First, the multiscale features F are processed by early feature fusion to obtain single-scale multi-frame feature representations F_C . Then, the spatio-temporal volume V is obtained by stacking F_C , as follows:

$$V = f_{sta}(F_C) = f_{att}(f_{up}(\sum_{i=1}^t F_C^i)), \quad (2)$$

where $f_{sta}(\cdot)$ is stacking operation, $f_{att}(\cdot)$ denotes the spatial and channel attention, and $f_{up}(\cdot)$ means upsampling.

Second, the spatial scene representation Z_s , which contains invariant appearance features, and the dynamic motion encoding Z_d , which explicitly models motion evolution patterns, are extracted by a scene encoder with two disentangled branches, described by following equations:

$$\begin{cases} F_a = \Gamma(V) = \sigma_r(f_B(f_c(V))), \\ Z_s = \psi_s(F_a) = \zeta(f_{mlp}(f_F(f_p(F_a)))), \\ Z_d = \psi_d(V) = \zeta(f'_{mlp}(f_F(f'_c(V)))), \end{cases} \quad (3)$$

where F_a is the smoothed features obtained by an aggregator $\Gamma(\cdot)$, comprising a 3D convolution block $f_c(\cdot)$ with batch normalization $f_B(\cdot)$ and ReLU activation σ_r . $\psi_s(\cdot)$ is the scene encoder with the 3D adaptive pooling $f_p(\cdot)$, flatten operation $f_F(\cdot)$, and multi-layer perceptron $f_{mlp}(\cdot)$. $\psi_d(\cdot)$ is the motion encoder and ζ is the tanh function. Then, Z_s and Z_d are concatenated to obtain the overall scene encoding Z with background prior knowledge.

Third, to construct continuous neural scene representations, we construct a 3D coordinate grid $\mathcal{G}$ to generate spatio-temporal coordinates $O \in \mathbb{R}^{hw \times 3}$ for each point (i, j) of keyframe features F_t . Position encoding is used to map coordinates O to a high-dimensional Fourier feature space to obtain refined query signals Q , as follows:

$$\begin{cases} O = \{\mathcal{G}(i, j, t) \mid \forall (i, j) \in F_t\}, \\ \mathcal{G}(i, j, t) = (\frac{2j}{w-1} - 1, \frac{2i}{h-1} - 1, \alpha \cdot t), \\ Q = \{\mathcal{P}(\mu) \mid \forall \mu \in O\} = f_{cat}(\left[\begin{smallmatrix} \sin(2^l \pi \mu) \\ \cos(2^l \pi \mu) \end{smallmatrix} \right]_{l=1}^L), \end{cases} \quad (4)$$

where α is a learnable time scale parameter, $f_{cat}(\cdot)$ is concatenation, and L is the number of frequency bands in Fourier feature mapping to control the representation ability of the position encoder $\mathcal{P}$ for high-frequency details.

Finally, the enhanced global scene encoding Z and refined high-frequency query signals are fed into the background neural fields $\mathcal{F}_\theta(\cdot)$ to perform counterfactual prediction, obtaining normal backgrounds $\mathcal{B}$, as follows:

$$\mathcal{B} = \mathcal{F}_\theta(Z, Q) = \mathcal{R}(f_\theta(f_{cat}(\mathcal{T}(Z), Q))), \quad (5)$$

where $\mathcal{T}(\cdot)$ denotes expanding scene context to each query point, $f_\theta(\cdot)$ represents two linear layers with residual connections (128 hidden dimension), achieving high-level detail reconstruction. $\mathcal{R}$ means restoring the tensor shape.

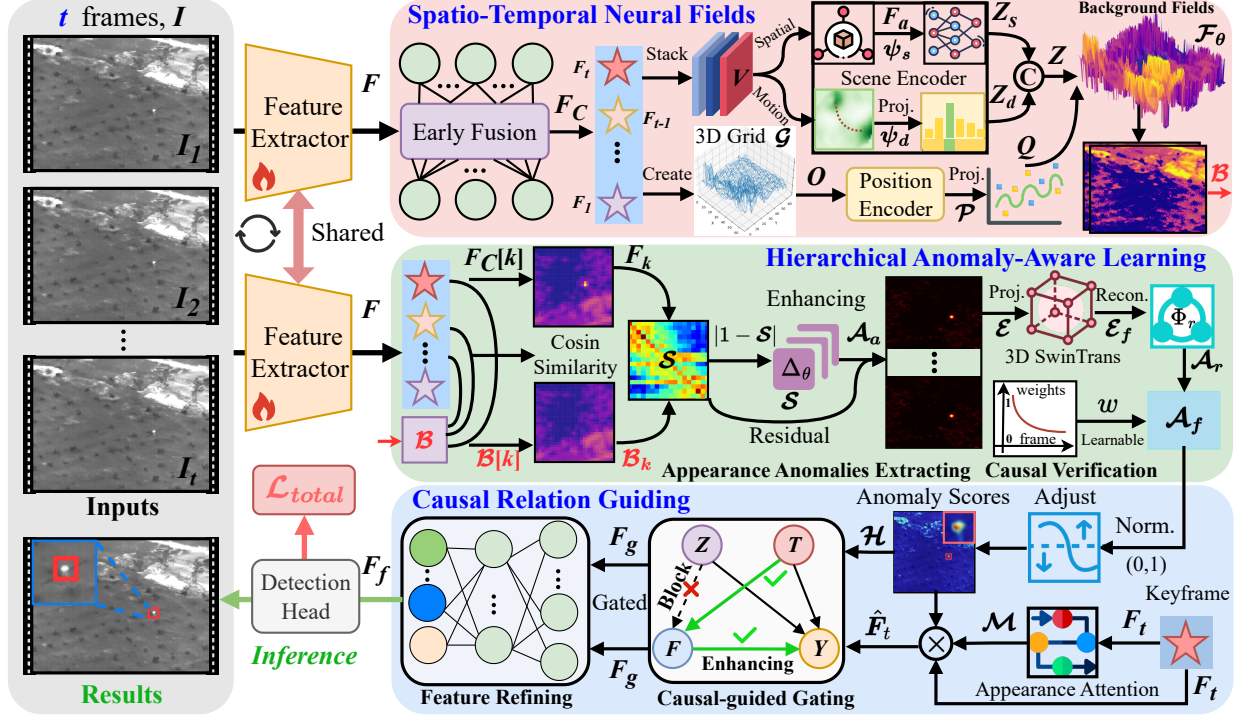


Figure 3. Our proposed CHAL framework. **Spatio-Temporal Neural Fields** is proposed to implicitly model the spatio-temporal evolution patterns of backgrounds, obtaining predicted backgrounds $\mathcal{B}$. **Hierarchical Anomaly-Aware Learning** is designed to decompose anomaly discovery from a simple single-step into a hierarchical one, mining potential appearance anomalies $\mathcal{A}_a$ and then performing causal verification based on motion consistency. **Causal Relation Guiding** is presented to address causal confounding, blocking spurious correlations from backgrounds (e.g., $F \leftarrow Z \rightarrow Y$), and enhancing true causality from real targets (e.g., $T \rightarrow F \rightarrow Y$).

3.3. Hierarchical Anomaly-Aware Learning

Unlike general anomaly detection methods [28, 41], our primary motivation is to decompose this process into a hierarchical one from discovery to validation. Therefore, we propose the HAL, as shown in Figure 3.

First, the appearance anomalies $\mathcal{A}_a$ are obtained by calculating the similarity $\mathcal{S}$ between every frame features F_k and predicted counterfactual backgrounds $\mathcal{B}_k$. It could be further described as follows:

$$\begin{cases} \mathcal{A}_a^k = \Delta_\theta(1 - f_{\cos}(F_k, \mathcal{B}_k)), k \in [1, t], \\ f_{\cos}(F_k, \mathcal{B}_k)(i, j) = \frac{F_k(i, j) \bullet \mathcal{B}_k(i, j)}{\|F_k(i, j)\|_2 \|\mathcal{B}_k(i, j)\|_2}, \end{cases} \quad (6)$$

where $\mathcal{A}_a^k \in \mathbb{R}^{b \times c \times h \times w}$ is the appearance anomalies of k -th frame, b is batch size, Δ_θ represents the anomaly amplifier with three convolution layers with residual connections. $f_{\cos}(\cdot)$ represents calculating cosine similarity, “ $\bullet$ ” means inner product, and $\|\cdot\|_2$ denotes l_2 norm.

Second, for the consecutive appearance anomaly sequence $\mathcal{A}_a$, it contains false anomalies caused by confounding factors. As such, we employ a N layers spatio-temporal swin transformer [30] to perform joint spatio-temporal modeling within local 3D windows using shifted

self-attention, achieving high-level anomaly perception and enhancing the motion consistency, that is

$$\begin{cases} \mathcal{E} = \{f_{pro}(\mathcal{A}_a^k) \mid \forall \mathcal{A}_a^k \in \mathcal{A}_a\} \in \mathbb{R}^{b \times t \times c \times h \times w} \\ f_{pro}(\mathcal{A}_a^k) = f_{sta}(\mathcal{W}_k \cdot \mathcal{A}_a^k + b_k) + \rho, k \in [1, t] \\ \mathcal{A}_c = \Upsilon[\mathcal{E} + \mathcal{F}_a^{(i)}(\mathcal{E}) + \mathcal{F}_{mlp}^{(i)}(\mathcal{E} + \mathcal{F}_a^{(i)}(\mathcal{E}))] \end{cases} \quad (7)$$

where $f_{pro}(\cdot)$ is projecting anomaly maps to a high-dimensional feature subspace, $\mathcal{W}_k$ and b_k are the weights and bias term, respectively. ρ is the position encoding information, $f_{sta}(\cdot)$ is stacking operation, $\mathcal{F}_a(\cdot)$ means the shifted window multi-head self-attention, $\mathcal{F}_{mlp}(\cdot)$ is the multi-layer perceptron with the DropPath and LayerNorm, and $\Upsilon[\cdot]$ denotes function composition ($i \in [1, N]$).

Finally, the anomaly maps $\mathcal{A}_c$ are reconstructed by a decoder Φ_r . Moreover, temporal-weight learning is designed to assign distinct weights ω_k to each frame, as follows:

$$\begin{cases} \mathcal{A}_r = \Phi_r(\mathcal{A}_c) = f_r(\pi_{-1}(\mathcal{A}_c)) \in \mathbb{R}^{b \times t \times c \times h \times w} \\ \mathcal{A}_f = \sum_{k=1}^t \omega_k \cdot \mathcal{A}_r^k, \omega_k = \frac{\exp(\beta_k)}{\sum_k \exp(\beta_k)}, k \in [1, t], \end{cases} \quad (8)$$

where $f_r(\cdot)$ represents a reconstructing decoder with two convolution layers, $\pi_{-1}(\cdot)$ means selecting the corresponding anomaly maps, and $\mathcal{A}_f$ is the final causal anomalies.

3.4. Causal Relation Guiding

As shown in Figure 2, the core challenge is the confounding path $F \leftarrow Z \rightarrow Y$. To block this path and estimate the true causal effect, we present CRG in Figure 3, to eliminate the confounding bias introduced by Z , **inspired** by the backdoor adjustment mechanism of causal theory [14].

The core idea is to stratify the data by the confounder Z and then adaptively compute the weights average of the predictions across all strata. Formally, the causal effect of F on Y could be estimated as follows:

$$P(Y|\text{do}(F)) = \sum_z P(Y|F, Z=z)P(Z=z), \quad (9)$$

where $\text{do}(\cdot)$ signifies an intervention to remove the influence of Z , $P(Y|F, Z=z)$ denotes the prediction conditioned on both features F and a specific z , and $P(Z=z)$ is the prior distribution of confounders. However, in MISTD, Z is continuous and dynamic, which makes direct computation difficult. Therefore, causal-guided gating is proposed as a differentiable substitution in feature subspace.

Specifically, the causally verified anomalies $\mathcal{A}_f$ are first normalized to $(0, 1)$, acting as a learned proxy for Z . We then apply a nonlinear filter to obtain anomaly scores $\mathcal{H}$ by

$$\mathcal{H} = \begin{cases} \eta_e \cdot f_{nor}(\mathcal{A}_f), & \text{if } f_{nor}(\mathcal{A}_f) > \tau, \\ \eta_s \cdot f_{nor}(\mathcal{A}_f), & \text{otherwise,} \end{cases} \quad (10)$$

where $\eta_e > 1$ and $\eta_s < 1$ are two learnable scaling factors to control the strength of intervention, stratifying $\mathcal{A}_f$ to two strata with anomaly threshold τ . $f_{nor}(\cdot)$ is normalization.

Meanwhile, $\hat{F}_t$ is computed by appearance attention $\mathcal{M}$, denoting the observed feature F with residual confounders. Then, the final features F_f are obtained by performing backdoor adjustment on confounders Z . By scaling $\hat{F}_t$ with anomaly scores $\mathcal{H}$, it could effectively isolate the confounding path $F \leftarrow Z \rightarrow Y$, thereby forcing the feature flow to approximate the true causal chain $T \rightarrow F \rightarrow Y$, that is

$$\begin{cases} \hat{F}_t = F_t \odot \mathcal{M} = F_t \odot [\sigma_s(f_e(F_t))], \\ F_f = f_{ref}(F_g) = f_{ref}(\hat{F}_t \odot (1 + \mathcal{H})), \end{cases} \quad (11)$$

where $f_{ref}(\cdot)$ is the feature refining with two-layer convolutions, guaranteeing the unconfounded feature F_f reflects the real causal effect $P(Y|\text{do}(F))$. Finally, F_f is fed into a detection head based on YOLOX [12] to obtain final results. The total training objective is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{obj} + \lambda_1 \mathcal{L}_{reg} + \lambda_2 \mathcal{L}_{cls}, \quad (12)$$

where λ_1 and λ_2 are two loss weights. We use sigmoid focal loss [20] as classification loss $\mathcal{L}_{cls}$ and probability loss $\mathcal{L}_{obj}$, NWD loss [37] as localization loss $\mathcal{L}_{reg}$. Crucially, $\mathcal{L}_{total}$ is designed to implicitly supervise all components (SNF, HAL and CRG), minimizing the total loss on F_f to drive the gradient to penalize deviations at every upstream stage, correctly identifying the true anomaly.

4. Experiments

4.1. Implement Details

All comparison methods are evaluated on three public datasets, DAUB-H [16], NUDT-MIRSDT [19], and IRDST-R [31]. For a fair comparison, following previous works [2, 7, 8], we use the standard detection evaluation metrics, *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 512×512 . In detail, our CHAL and compared methods are all trained for 100 epochs with a batch size of 4. AdamW is adopted as the optimizer with an initial learning rate of 8×10^{-5} and a weight decay of 5×10^{-2} . Hyper-parameters t , L , N , τ , λ_1 , and λ_2 are set to 5, 6, 2, 0.3, 4.0, and 1.2, respectively, with detailed experiments in Appendix.

4.2. Comparison with SOTA Methods

Quantitative Comparisons. Table 1 presents the quantitative comparisons across 17 representative methods, including single-frame and multi-frame ones. From table, we could have two obvious observations. One is that our CHAL establishes new SOTA benchmarks across most metrics, consistently securing the top performance on all three datasets. For example, on NUDT-MIRSDT, our CHAL obtains the highest mAP_{50} 75.25%, Re 88.49% and F1 87.41%. Only in terms of Pr , the 86.35% by our CHAL is slightly lower than the SOTA 92.50% by ADSUNet. It achieves a higher Pr by sacrificing Re , whereas our method is more balanced, achieving a higher F1. The other is that general anomaly detection methods are ineffective in highly challenging infrared scenarios. For instance, on DAUB-H, the SOTA AD one, *i.e.*, DiffusionAD, only obtains mAP_{50} 19.50% and F1 42.13%, notably lower than the mAP_{50} 54.28% and F1 74.15% by CHAL. Moreover, because these AD schemes primarily focus on visible-light images, on IRDST-R, they even fail to converge during training.

PR Curve Comparisons. As usual, we further employ precision-recall (PR) curves to visually assess the comprehensive performance of various methods, as illustrated in Figure 4. The proximity of a method to the top-right corner directly indicates its effectiveness. These figures clearly

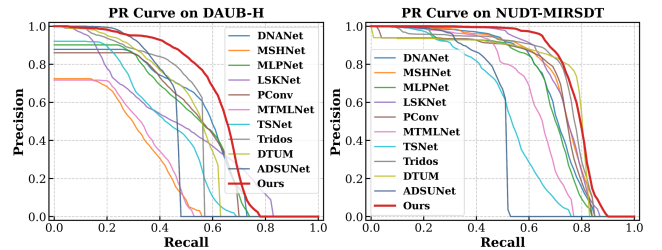


Figure 4. PR curve comparisons on two datasets (partly).

Table 1. Quantitative comparisons (**partly**). The best and second-best results are marked in **red** and **blue**, respectively. All multi-frame methods utilize five frames for fair comparison. “★” means the method is based on anomaly detection. FPS is tested on RTX4090.

Method	Pub	DAUB-H				NUDT-MIRSdT				IRDST-R				Param↓	GFlops↓	FPS↑	
		mAP ₅₀ (%)	Pr (%)	Re (%)	F1 (%)	mAP ₅₀ (%)	Pr (%)	Re (%)	F1 (%)	mAP ₅₀ (%)	Pr (%)	Re (%)	F1 (%)				
Single-frame	ISNet [43]	CVPR’22	44.73	56.35	80.99	66.46	68.72	81.62	85.50	83.51	59.27	73.75	81.98	77.65	3.49M	265.73	11.20
	UIUNet [35]	TIP’22	49.23	73.72	67.49	70.47	66.63	83.00	81.83	82.41	59.22	76.55	78.47	77.50	53.06M	456.70	3.63
	DNANet [18]	TIGR’23	50.76	71.04	72.01	71.52	68.97	82.87	84.61	83.73	67.43	85.97	79.04	82.36	7.22M	135.24	4.82
	SIRST5K [27]	TGRS’24	40.66	57.63	71.14	63.68	68.91	86.88	80.76	83.71	47.43	71.08	67.03	68.99	11.48M	182.61	7.37
	MSHNet [22]	CVPR’24	26.47	48.56	55.39	51.75	71.77	86.18	84.74	85.45	60.86	78.54	78.52	75.53	6.59M	69.59	18.55
	LSKNet [34]	TGRS’25	46.34	56.97	82.76	67.48	72.73	85.36	86.07	85.72	65.84	82.66	80.88	81.76	3.42M	76.00	40.63
	PConv [39]	AAAI’25	48.48	70.74	69.14	69.93	68.91	83.74	83.03	83.39	58.45	79.50	74.74	77.04	6.59M	69.29	10.01
	MTMLNet [38]	TIP’25	27.56	53.46	52.48	52.96	61.66	81.82	76.48	79.06	58.53	80.72	73.55	76.97	16.79M	117.97	15.21
	DiffusionAD ★ [41]	TPAMI’25	19.50	31.72	62.71	42.13	Training Without Convergence				Training Without Convergence				39.32M	551.78	5.05
	INPFormer ★ [28]	CVPR’25	1.31	5.73	23.84	9.24	0.28	2.00	14.02	3.50	Training Without Convergence				142.37M	362.92	0.67
TSNet [17]	ESWA’26	40.34	59.88	68.69	63.99	51.76	69.66	75.90	72.64	39.22	58.98	67.75	63.06	3.18M	197.81	12.81	
Multi-frame	ST-Trans [32]	TGRS’24	44.93	87.21	52.54	65.57	67.28	93.41	72.66	81.74	64.50	86.84	75.16	80.58	38.13M	145.16	3.90
	SSTNet [2]	TGRS’24	52.25	83.49	63.25	71.98	64.87	87.17	75.09	80.68	68.21	75.70	91.34	82.79	11.95M	123.59	9.24
	Tridos [7]	TGRS’24	49.72	88.55	56.69	69.13	73.01	87.66	84.15	85.87	61.69	83.16	74.11	78.37	14.13M	130.72	11.23
	DTUM [19]	TNNLS’25	50.32	81.53	63.00	71.08	72.45	88.99	82.77	85.77	64.72	83.32	78.81	81.00	9.64M	128.16	14.28
	ADSUNet [42]	PR’26	43.43	92.54	47.96	63.18	47.66	92.50	52.12	66.67	66.68	88.76	75.24	81.44	26.06M	112.84	10.52
	CHAL (Ours)	-	54.28	73.08	75.26	74.15	75.25	86.35	88.49	87.41	71.37	86.58	83.02	84.76	15.69M	137.04	12.96

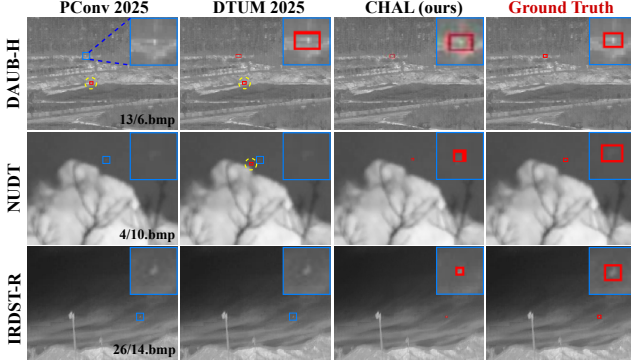


Figure 5. Visual comparisons. Blue boxes and yellow circles are the missed and false detections. More results are in **Appendix**.

show that our curves outperform those of competing methods. For example, on DAUB-H, our PR curve consistently occupies the higher positions. This pattern continues on NUDT-MIRSdT, where our curve exceeds others in the top-right. Therefore, these PR curves validate the superiority of our CHAL in balancing Pr and Re compared to others.

Inference Cost Comparisons. Table 1 also shows the inference Cost comparisons, revealing two notable findings. **First**, although our CHAL utilizes a sequence of five consecutive frames, its parameters remain at a moderate level. For example, our model has 15.69M parameters, higher than the SOTA method ISNet (3.49M), but still lower than many methods, *e.g.*, the 16,79M by MTMLNet, and the 26.06M by ADSUNet. **Second**, processing multi-frame images causes a high GFLOPs, but it often promotes detection performance. For instance, our CHAL has 137.04 GFLOPs, higher than some methods. However, compared to all other methods, its mAP₅₀ and F1 score are the highest.

Visual Comparisons. For intuitiveness, we select 2 representative methods to visually compare our CHAL on three datasets, as depicted in Figure 5. It is obvious that our proposed CHAL typically detects targets accurately, whereas the other two often result in missed or false detections. For example, on NUDT-MIRSdT, our CHAL could accurately detect the target near a tree, while PConv could not find the real target, and DTUM mistakenly treats a bright spot as a target. On challenging IRDST-R, the other two could not even detect the target occluded by clouds.

4.3. Ablation and Analysis

Effects of Different Components. To evaluate the impact of various configurations within our CHAL, we conduct a series of ablation studies, as shown in Table 2. Through comparisons, we could have **two** obvious findings. **First**, it is apparent that the individual components are consistently effective in enhancing detection capabilities. For example, on DAUB-H, the baseline configuration (w/o All), devoid of any specialized components, only achieves mAP₅₀ 23.48% and F1 46.36%. The integration of SNF (w S) improves these scores to 43.76% for mAP₅₀ and 60.90% for F1. Similarly, incrementally applying HAL (w S & H) further increases metrics to 48.63% for mAP₅₀ and 66.68% for F1. **Second**, these components have synergistic effects. For instance, on NUDT-MIRSdT, when all components are fully assembled (w All), performance reaches peaks, with an mAP₅₀ of 54.28% and an F1 of 74.15%.

Effects of Spatio-Temporal Neural Fields. Figure 6 presents the 3D visualizations before and after SNF to show the background modeling process. From this, it is obvious that the features before SNF are confounded, containing both a low-frequency background plane and a peak repre-

Table 2. The ablation study of our CHAL with different settings on DAUB-H and NUDT-MIRSDDT. **SNF**: three components of spatio-temporal neural fields (**S1** is background fields $\mathcal{F}_\theta$, **S2** denotes scene encoder, **S3** means position encoder). **HAL**: two anomaly-aware learning schemes (**H1** is appearance anomaly learning, **H2** denotes causal anomaly verification based on motion consistency). **CRG**: three components of causal relation guiding (**C1** is anomaly enhancing, **C2** denotes causal-guided gating, **C3** is causal feature refining).

Settings	SNF			HAL			CRG			DAUB-H				NUDT-MIRSDDT				Param↓	GFlops↓
	S1	S2	S3	H1	H2		C1	C2	C3	mAP ₅₀	Pr	Re	F1	mAP ₅₀	Pr	Re	F1		
w/o All	-	-	-	-	-	-	-	-	-	23.48	35.38	67.21	46.36	50.30	60.47	84.69	70.56	9.99M	52.53
w S	✓	-	-	-	-	-	-	-	-	37.72	53.47	59.62	56.38	58.08	69.03	74.74	71.77	10.61M	62.89
	✓	✓	-	-	-	-	-	-	-	41.21	54.12	63.66	58.50	61.68	73.95	75.81	74.87	11.93M	78.22
	✓	✓	✓	-	-	-	-	-	-	43.76	56.91	65.50	60.90	63.50	75.30	77.90	76.58	12.48M	89.47
w S & H	✓	✓	✓	✓	-	-	-	-	-	45.28	59.99	66.84	63.23	65.55	78.12	81.52	79.78	13.39M	102.94
	✓	✓	✓	✓	✓	-	-	-	-	48.63	66.20	67.17	66.68	68.49	80.64	84.93	82.73	14.17M	120.33
w S & H & C	✓	✓	✓	✓	✓	✓	-	-	-	50.45	62.92	79.20	70.13	71.17	83.22	86.13	84.65	14.86M	124.27
	✓	✓	✓	✓	✓	✓	✓	-	-	53.19	67.57	79.68	73.12	73.82	85.30	87.90	86.58	14.98M	128.39
	✓	✓	✓	✓	✓	✓	✓	✓	✓	52.54	63.08	83.00	71.68	72.73	83.55	88.86	86.12	15.32M	135.22
w All	✓	✓	✓	✓	✓	✓	✓	✓	✓	54.28	73.08	75.26	74.15	75.25	86.35	88.49	87.41	15.69M	137.04

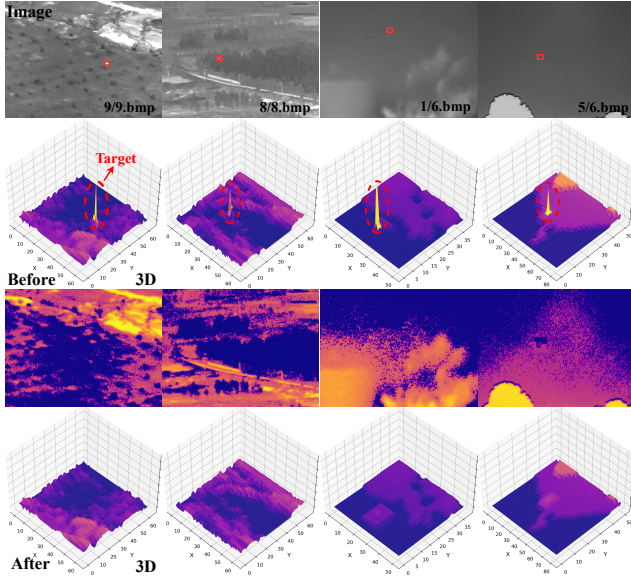


Figure 6. Visualization of the background modeling. The first and third rows show the original infrared images and predicted backgrounds $\mathcal{B}$. The second and fourth rows denote the 3D visualizations before and after SNF. More results are in **Appendix**.

sending anomaly signals. After SNF, the peak corresponding to the target appears to be removed, and it is not included in the predicted backgrounds. It indicates that our SNF does not simply provide a background average. Instead, as a critical foundation for discovering anomalies, it attempts to precisely reconstruct complex background confounders, *e.g.*, the cloud edges in the third column (1/6.bmp).

Effects of Hierarchical Anomaly-Aware Learning. To thoroughly analyze the impacts of HAL, we select two samples from DAUB-H and NUDT-MIRSDDT to visualize the feature distributions of targets (red) and backgrounds (blue) at different stages, as shown in Figure 7. From this, it could be easily observed that when only using the appearance

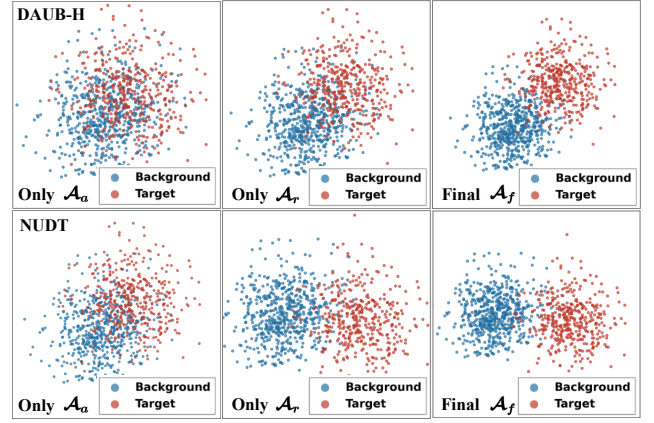


Figure 7. Feature distributions of the $\mathcal{A}_a$, $\mathcal{A}_r$ and $\mathcal{A}_f$ in HAL.

anomaly $\mathcal{A}_a$, targets and backgrounds overlap significantly. When employing the reconstructed temporal anomaly $\mathcal{A}_r$, they begin to separate gradually. After applying the final causal anomaly $\mathcal{A}_f$, clear boundaries emerge. It validates our initial motivation that relying solely on appearance anomaly is insufficient to distinguish true targets from background confounders. Moreover, this clearly demonstrates the effectiveness of our HAL strategy, successfully learning a robust feature representation.

Effects of Causal Relation Guiding. First, to visually demonstrate the effectiveness of CRG, we present four groups of feature heatmaps before and after CRG, as illustrated in Figure 8. From figure, it is evident that the focus positions of feature heatmaps by “w/o CRG” are obscure, and targets are even lost in complex backgrounds. In contrast, after employing CRG, the feature response of small targets is notably enhanced. Meanwhile, the noisy background becomes cleaner. This indicates that causal-guided gating and feature refinement could effectively suppress spurious correlations and enhance true causality, thereby improving the detection performance of small targets.

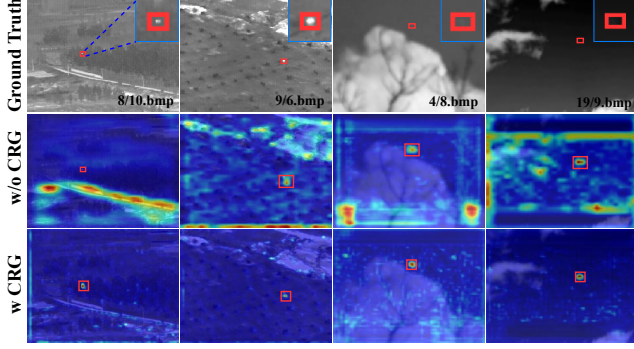


Figure 8. Feature heatmap comparisons before (w/o) and after (w) CRG. The first two columns are on DAUB-H, and the last two are on NUDT-MIRSdT. More results are in **Appendix**.

Second, to verify the motivation of CRG, we compare it with general attentions, as shown in Table 3. From it, we observe that “w CRG” exhibits overwhelming superiority. For example, its mAP_{50} surpasses the SOTA CBAM by 3.83% on DAUB-H and 5.07% on NUDT-MIRSdT. Notably, its 15.69M parameters are even lower than the 16.28M by self-attention. It indicates that the significant performance gain stems from our causal-guided design, which is more effective than target-centered attention mechanisms.

Table 3. Ablation study on general attention and our CRG.

Settings	DAUB-H		NUDT-MIRSdT		Param↓
	mAP_{50}	F1	mAP_{50}	F1	
w/o CRG	48.63	66.68	68.49	82.73	14.17M
Self-Attention	49.85	67.74	69.27	82.93	16.28M
CBAM	50.45	68.85	70.18	83.42	15.13M
w CRG	54.28	74.15	75.25	87.41	15.69M

Effect of Anomaly Strength and Type. To further explore the perception ability of our CHAL for different anomaly strengths and types, we perform a group of ablation experiments on several synthetic test sets based on NUDT-MIRSdT, as shown in Figure 9. Specifically, we control the anomaly strength by adjusting the contrast between synthetic targets and local backgrounds. Three representative anomaly types are designed, *i.e.*, brightness, texture, and motion (the motion trajectory deviates from the background). From figure, we could clearly find that our CHAL notably outperforms all ablation variants. For example, when the strength is less than 0.4, the performance of “w/o HAL” and “only $\mathcal{A}_a$ ” tends to be 0. Conversely, our CHAL still maintains a high F1. Moreover, this trend continues for all anomaly types. It shows that our CHAL could simultaneously perceive both low-level appearance anomalies and high-level motion ones, while others cannot.

Performance in Visible-Light Scenarios. Table 4 presents the comparisons on a challenging visible-light re-

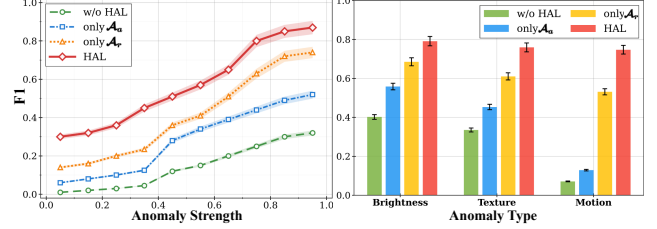


Figure 9. The ablation study of different anomaly strengths and types on our CHAL. More experimental results are in **Appendix**.

mote sensing dataset, *i.e.* RsCarData [36], to validate the adaptability of our CHAL. Despite differences in imaging modalities, the vehicle detection on RsCarData faces similar challenge as MISTD. From it, it is obvious that our CHAL significantly outperforms all compared methods across most metrics. For instance, our method achieves the highest mAP_{50} of 57.37% and F1 of 76.14%, surpassing the second-best method STF by 4.96% in mAP_{50} and 4.41% in F1. It indicates that our CHAL is not merely tailored for infrared domains, but a generalized framework can be successfully extended to other small-object detection tasks.

Table 4. Performance comparisons on visible-light scenarios. “▲” denotes a multi-frame method using five consecutive frames.

Methods	RsCarData				Param↓
	mAP_{50} (%)	Pr (%)	Re (%)	F1 (%)	
CenterNet [6]	48.90	63.17	69.42	66.15	14.44M
VFNet [40]	36.80	48.12	56.35	51.91	32.67M
TOOD [10]	42.00	52.09	63.23	57.12	31.79M
YOLOX [12]	43.51	54.26	64.35	58.88	8.97M
DSFNet ▲ [36]	51.75	69.98	73.46	70.07	19.84M
STF ▲ [47]	52.41	68.53	75.25	71.73	16.13M
HyperYOLO [11]	16.51	17.91	26.51	21.38	33.37M
CFPT [5]	20.42	41.63	49.99	45.43	37.08M
CHAL (Ours)	57.37	81.08	71.78	76.14	15.69M

5. Conclusion

To address the shortcomings of current target-centered detection paradigm, this paper proposes the first causal-guided hierarchical anomaly-aware learning framework (*i.e.* CHAL) for moving infrared small targets. Just opposite to existing works, it emphasizes the background feature learning, rather than small target features, with all targets treated as the anomalies appearing in backgrounds. In the CHAL, Spatio-Temporal Neural Fields, Hierarchical Anomaly-Aware Learning and Causal Relation Guiding strategy are designed to implement our new detection framework. The experiments on three infrared benchmarks verify the effectiveness and the superiority of our CHAL, even in visible-light scenarios. Its weakness is the large parameter scale, causing relatively high computational cost. In the future, a lighter-weight anomaly-aware scheme with efficient causal learning merits further exploration.

Acknowledgments This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No.62476049 and 62276048. The authors would like to thank the anonymous reviewers for their constructive comments that helped improve this work.

References

- [1] Xiangzhi Bai and Fugen Zhou. Analysis of new top-hat transformation and the application for infrared dim small target detection. *Pattern Recognition*, 43(6):2145–2156, 2010. 1
- [2] Shengjia Chen, Luping Ji, Jiewen Zhu, Mao Ye, and Xiaoyong Yao. 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, 2024. 1, 5, 6
- [3] Tiejuan Chen, Huabin Liu, Tianyao He, Yihang Chen, Chaofan Gan, Xiao Ma, Cheng Zhong, Yang Zhang, Yingxue Wang, Hui Lin, et al. MECD: Unlocking multi-event causal discovery in video reasoning. *Advances in Neural Information Processing Systems*, 37:92554–92580, 2024. 2
- [4] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. Asymmetric Contextual Modulation for Infrared Small Target Detection. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 949–958, 2021. 1, 2
- [5] Zewen Du, Zhenjiang Hu, Guiyu Zhao, Ying Jin, and Hongbin Ma. Cross-layer feature pyramid transformer for small object detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 8
- [6] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019. 8
- [7] Weiwei Duan, Luping Ji, Shengjia Chen, Sicheng Zhu, and Mao Ye. 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, 2024. 1, 2, 5, 6
- [8] Weiwei Duan, Luping Ji, Shengjia Chen, Sicheng Zhu, Jianghong Huang, and Mao Ye. Weakly supervised contrastive learning with quantity prompts for moving infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025. 1, 5
- [9] Weiwei Duan, Luping Ji, Jianghong Huang, Shengjia Chen, Shuang Peng, Sicheng Zhu, and Mao Ye. Semi-supervised multi-view prototype learning with motion reconstruction for moving infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 1, 2
- [10] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3490–3499. IEEE Computer Society, 2021. 1, 8
- [11] Yifan Feng, Jiangang Huang, Shaoyi Du, Shihui Ying, Jun-Hai Yong, Yipeng Li, Guiguang Ding, Rongrong Ji, and Yue Gao. Hyper-yolo: When visual object detection meets hyper-graph computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 8
- [12] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 3, 5, 8
- [13] Jinhui Han, Saed Moradi, Iman Faramarzi, Honghui Zhang, Qian Zhao, Xiaojian Zhang, and Nan Li. Infrared small target detection based on the weighted strengthened local contrast measure. *IEEE Geoscience and Remote Sensing Letters*, 18(9):1670–1674, 2020. 1
- [14] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986. 2, 5
- [15] Haoqi Huang, Ping Wang, Jianhua Pei, Jiacheng Wang, Shaheen Alexanian, and Dusit Niyato. Deep learning advancements in anomaly detection: A comprehensive survey. *IEEE Internet of Things Journal*, pages 1–1, 2025. 2
- [16] Bingwei Hui, Zhiyong Song, Hongqi Fan, Ping Zhong, Weidong Hu, Xiaofeng Zhang, Jianguo Lin, Hongyan Su, Wei Jin, Yongjie Zhang, and Yaxi Bai. A dataset for infrared image dim-small aircraft target detection and tracking under ground / air background, 2019. 5
- [17] Sixiang Ji, Haofei Zhang, Jingmin Zhang, Chun Fei, Xiaoyang Wang, Juanxiu Liu, and Ping Zhang. A three-stage model for infrared small target detection with spatial and semantic feature fusion. *Expert Systems with Applications*, 295:128776, 2026. 6
- [18] Boyang Li, Chao Xiao, Longguang Wang, Yingqian Wang, Zaiping Lin, Miao Li, Wei An, and Yulan Guo. Dense nested attention network for infrared small target detection. *IEEE Transactions on Image Processing*, 32:1745–1758, 2022. 6
- [19] Ruoqing Li, Wei An, Chao Xiao, Boyang Li, Yingqian Wang, Miao Li, and Yulan Guo. Direction-coded temporal u-shape module for multiframe infrared small target detection. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):555–568, 2025. 1, 2, 5, 6
- [20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 5
- [21] Xiangru Lin, Ziyi Wu, Guanqi Chen, Guanbin Li, and Yizhou Yu. A causal debiasing framework for unsupervised salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1610–1619, 2022. 2
- [22] Qiankun Liu, Rui Liu, Bolun Zheng, Hongkui Wang, and Ying Fu. Infrared small target detection with scale and location sensitivity. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition*, 2024. 1, 2, 6
- [23] Ting Liu, Yongxian Liu, Jungang Yang, Boyang Li, Yingqian Wang, and Wei An. Graph laplacian regularization for fast infrared small target detection. *Pattern Recognition*, 158:111077, 2025. 2
- [24] Yajing Liu, Shijun Zhou, Xiyao Liu, Chunhui Hao, Baojie Fan, and Jiandong Tian. Unbiased faster r-cnn for single-source domain generalized object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28838–28847, 2024. 2

- [25] Yang Liu, Hongjin Wang, Zepu Wang, Xiaoguang Zhu, Jing Liu, Peng Sun, Rui Tang, Jianwei Du, Victor C. M. Leung, and Liang Song. Crcl: Causal representation consistency learning for anomaly detection in surveillance videos. *IEEE Transactions on Image Processing*, 34:2351–2366, 2025. 2
- [26] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20402–20411, 2023. 2
- [27] Yahao Lu, Yupei Lin, Han Wu, Xiaoyu Xian, Yukai Shi, and Liang Lin. SIRST-5K: Exploring Massive Negatives Synthesis with Self-supervised Learning for Robust Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 6
- [28] Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenying Yu. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9974–9983, 2025. 2, 4, 6
- [29] Shuang Peng, Luping Ji, Shengjia Chen, Weiwei Duan, and Sicheng Zhu. Moving infrared dim and small target detection by mixed spatio-temporal encoding. *Engineering Applications of Artificial Intelligence*, 144:110100, 2025. 3
- [30] Tushar Sangam, Ishan Rajendrakumar Dave, Waqas Sultani, and Mubarak Shah. Transvisdrone: Spatio-temporal transformer for vision-based drone-to-drone detection in aerial videos. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6006–6013. IEEE, 2023. 4
- [31] Heng Sun, Junxiang Bai, Fan Yang, and Xiangzhi Bai. 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, 2023. 5
- [32] Xiaozhong Tong, Zhen Zuo, Shaojing Su, Junyu Wei, Xiaoyong Sun, Peng Wu, and Zongqing Zhao. ST-Trans: Spatial-Temporal Transformer for Infrared Small Target Detection in Sequential Images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2, 3, 6
- [33] Guanghui Wang, Bingjie Tao, Xuan Kong, and Zhenming Peng. Infrared small target detection using nonoverlapping patch spatial-temporal tensor factorization with capped nuclear norm regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2021. 1, 2
- [34] Fengyi Wu, Anran Liu, Tianfang Zhang, Luping Zhang, Junhai Luo, and Zhenming Peng. Saliency at the helm: Steering infrared small target detection with learnable kernels. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025. 6
- [35] Xin Wu, Danfeng Hong, and Jocelyn Chanussot. UIU-Net: U-Net in U-Net for infrared small object detection. *IEEE Transactions on Image Processing*, 32:364–376, 2022. 6
- [36] Chao Xiao, Qian Yin, Xinyi Ying, Ruojing Li, Shuanglin Wu, Miao Li, Li Liu, Wei An, and Zhijie Chen. Dsfnet: Dynamic and static fusion network for moving object detection in satellite videos. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021. 8
- [37] Chang Xu, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. Detecting tiny objects in aerial images: A normalized wasserstein distance and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190: 79–93, 2022. 5
- [38] Bo Yang, Fengqian Li, Songliang Zhao, Wei Wang, Jun Luo, Huayan Pu, Mingliang Zhou, and Yangjun Pi. MTMLNet: Multi-task mutual learning network for infrared small target detection and segmentation. *IEEE Transactions on Image Processing*, 2025. 2, 6
- [39] Jiangnan Yang, Shuangli Liu, Jingjun Wu, Xinyu Su, Nan Hai, and Xueli Huang. Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9202–9210, 2025. 1, 2, 6
- [40] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8514–8523, 2021. 8
- [41] Hui Zhang, Zheng Wang, Dan Zeng, Zuxuan Wu, and Yungang Jiang. DiffusionAD: Norm-Guided One-Step Denoising Diffusion for Anomaly Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):7140–7152, 2025. 2, 4, 6
- [42] Liuwei Zhang, Yuyang Xi, Zhipeng Wang, Wang Zhang, Fanjiao Tan, and Qingyu Hou. ADSUNet: Accumulation-difference-based Siamese U-Net for inter-frame infrared dim and small target detection. *Pattern Recognition*, 169:111942, 2026. 2, 6
- [43] Mingjin Zhang, Rui Zhang, Yuxiang Yang, Haichen Bai, Jing Zhang, and Jie Guo. ISNet: Shape matters for infrared small target detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 877–886, 2022. 1, 6
- [44] Hang Zhou, Jiale Cai, Yuteng Ye, Yonghui Feng, Chenxing Gao, Junqing Yu, Zikai Song, and Wei Yang. Video anomaly detection with motion and appearance guided patch diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10761–10769, 2025. 2
- [45] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. Transvod: end-to-end video object detection with spatial-temporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7853–7869, 2022. 3
- [46] Sicheng Zhu, Luping Ji, Jiewen Zhu, Shengjia Chen, and Weiwei Duan. TMP: Temporal Motion Perception with spatial auxiliary enhancement for moving Infrared dim-small target detection. *Expert Systems with Applications*, page 124731, 2024. 2
- [47] Sicheng Zhu, Luping Ji, Jiewen Zhu, Shengjia Chen, and Haohao Ren. Spatio-temporal fusion with motion masks for the moving small target detection from remote-sensing videos. *Engineering Applications of Artificial Intelligence*, 138:109362, 2024. 8