Explore Hybrid Modeling for Moving Infrared Small Target Detection

Anonymous Authors

ABSTRACT

Moving infrared small target detection, crucial in contexts like traffic management and maritime rescue, encounters challenges from factors such as complex backgrounds, target occlusion, camera shake, and motion blur. Existing algorithms fall short in comprehensively addressing these issues by exploring hybrid modeling, impeding generalization in complex and dynamic motion scenes. In this paper, we propose a hybrid modeling method for moving infrared small target detection via smoothed-particle hydrodynamics (SPH) and Markov decision processes (MDP). SPH can simulate the motion trajectories of targets and background scenes, while MDP can optimize detection system strategies for optimal action selection based on contexts and target states. Specifically, we develop an SPH-inspired image-level enhancement algorithm which models the image sequence of infrared video as a 3D spatiotemporal graph in SPH. In addition, we design an MDP-guided temporal feature perception module. This module selects reference frames, aggregates features from both reference frames and the current frame. The previous and current frames are modeled as an MDP tailored for multi-frame infrared small target detection tasks, aiding in detecting the current frame. Conducted extensive experiments on two public dataset: DAUB and DATR, the proposed network surpasses the state-of-the-art methods in terms of objective metrics and visual quality.

CCS CONCEPTS

• **Computing methodologies** → **Object detection**; **Matching**.

KEYWORDS

Moving infrared small target detection, Deep Learning, Mathematical model, Smoothed-particle hydrodynamics, Markov decision processes

1 INTRODUCTION

Identifying moving targets in challenging weather conditions such as fog and heavy rain is often difficult with visible light videos. In contrast, infrared (IR) videos offer more reliable target detection, even in adverse weather, due to their unique imaging mechanism [26, 36, 39]. Accordingly, detecting small moving targets, derived from this unique modality, *i.e.*, IR videos, is a prominent subject in computer vision, widely applied in traffic management and maritime rescue [28, 37].

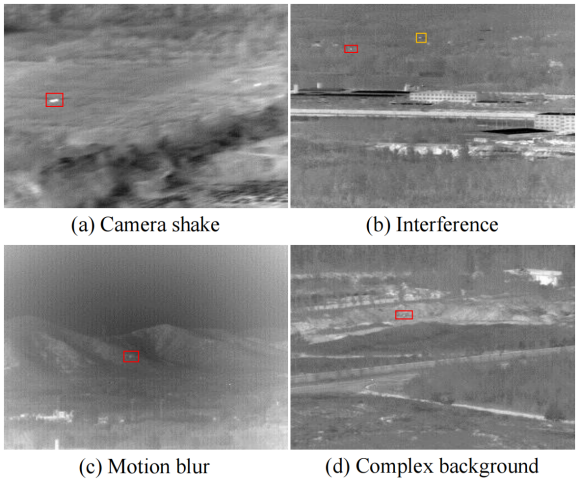

Figure 1: Typical issue with moving infrared small targets. Small targets marked by red bounding boxes, interferences by yellow.

As shown in Figure 1, the long-distance nature of infrared imaging results in insufficient information regarding target details such as size, shape, and texture. Challenges arise for moving infrared targets, including complex background, motion blur, interference and camera shake [9, 21]. Addressing these internal and external factors makes the task of detecting moving infrared small targets in video sequences exceptionally challenging.

In recent years, numerous algorithms have emerged for infrared small target detection [6, 12, 30, 38], categorized into single-frame and multi-frame methods. Single-frame methods focus on small target characteristics, utilizing complex nested network structures and attention modules to minimize information loss during pooling and downsampling processes. DNA-Net [20] employs densely nested interactive and spatial attention modules for feature fusion and enhancement, while UIU-Net [31] achieves multi-level learning by embedding a small U-Net into a larger one, yielding promising results in single-frame detection. However, limitations such as occlusion, motion blur, and camera shake hinder single-frame methods' efficacy in capturing moving infrared small targets. Human visual judgment can infer a blurry target's identity by leveraging information from adjacent frames in videos. Utilizing multiple frames provides rich temporal information compared to a single frame, enabling various multi-frame methods like *image-level target enhancement* and *temporal feature perception* [4, 25, 32, 45].

For image-level target enhancement, Du *et al.* [10] enhance small targets through inter-frame alignment, yet overlook spatial information's significance. Zhu *et al.* [45] leverage optical flow to enhance

Unpublished working draft. Not for distribution.

ACM MM, 2024, Melbourne, Australia

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

moving targets, but this method assumes constant target brightness, which may not hold for infrared small targets due to fluctuations caused by factors like temperature changes and occlusions. In addition, direct application of mainstream YOLO series enhancement algorithms[15, 18, 19] like mosaic enhancement to infrared small targets often leads to target loss. In summary, existing image-level enhancement methods have progressed but have limitations on specific modalities like IR videos, thus failing to achieve desired detection effects. *Can we view image-level target enhancement from a mathematical modeling perspective?* By scientifically modeling the motion state of the target, we can capture more temporal information. Smoothed Particle Hydrodynamics (SPH) simulates fluid behavior by dividing it into particles and simulating their interactions. SPH can model image-level enhancement, simulating information transmission and interaction within IR images, thus facilitating enhancement.

For temporal feature perception, existing methods often employ complex temporal feature aggregation networks. For instance, SSTNet [4] leverages LSTM's memory prediction and a cross-slice ConvLSTM structure to aggregate temporal information. Luo *et al.* [22] utilize dense nested structures and optical flow to design a multi-scale optical flow reconstruction network for capturing moving small targets. The above temporal feature aggregation networks utilize neural networks' nonlinear properties and parameter learning to handle complex temporal data, potentially boosting performance at the cost of increased training expenses. *Can we treat temporal feature perception as a prediction task and model state evolution in time-series IR data?* The Markov decision process (MDP) can offer a simplified approach to modeling time-series IR data by abstracting it into states and corresponding transition probabilities. This can allow us to move away from complex temporal aggregation networks and focus on prediction instead.

In this paper, we propose a method to explore hybrid models for detecting small moving infrared targets using SPH and MDP. Inspired by SPH, we represent motion as fluid dynamics at the image level, with the background and target modeled as stationary and moving particles, respectively. We develop an SPH-inspired image-level enhancement algorithm, using 3D spatiotemporal modeling and SPH Gaussian elliptical kernels for 3D sliding filtering. During sliding, it enhances local contrast and aggregates temporal information, improving efficiency and unifying spatiotemporal dimensions. In addition, we design an MDP-guided temporal feature perception module, comprising a lightweight feature aggregation network and a prediction propagation module. It enriches the temporal information of the target in the current frame by aggregating the reference frame, while reusing detection results from the previous frame and integrating current frame predictions to model the Markov decision. This assists in detecting the current frame across various modeling states and extracting temporal information from multiple frames. Experimental results on two public datasets DAUB[16] and DATR[13], incorporating multiple metrics, indicate that the proposed method outperforms the state of-the-art (SOTA) methods.

·We find hybird models for moving infrared dim-small target detection, with SPH and MDP playing a crucial role. These models describe target motion and background changes, optimize decision strategies, and boost detection system performance and efficiency. Experimental results on the DAUB and DATR datasets show that our method surpasses the SOTA methods.

· We pioneer a mathematical approach to image-level target enhancement and design an SPH-inspired image-level enhancement algorithm. Due to SPH's ability to establish strong spatiotemporal relationships, our enhancement algorithm effectively retains details and structure in IR videos, yielding enhanced images with greater accuracy and naturalness.

· We make the first attempt to treat temporal feature perception as a prediction task by designing an MDP-guided temporal feature perception module. This tightly connected and hierarchical module fully exploits temporal information and detection results from reference frames. Modeling the motion process as a Markov model follows explicit and interpretable design principles.

2 RELATED WORK

2.1 Infrared Small Target Detection in Image

Due to the characteristics of infrared imaging, traditional single-frame detection mainly focuses on modeling the relationship among the target, background, and noise. Traditional algorithms include filter-based methods such as maximum median/mean filters [9] and new top-hat filters [1], human visual system-based local contrast algorithms like LCM [3] and the improved algorithm MPCM [29] based on local contrast, as well as detection algorithms based on sparse representation, such as IPI [14] and improved RIPT [7], etc. However, these methods often lack balance between background suppression and target enhancement and rely heavily on manually extracted features, resulting in lower accuracy and higher false alarm rates.

Influenced by CNN, single-frame infrared small target detection based on deep learning has become mainstream. Dai *et al.* [8] utilize bottom-up attention modulation, integrating low-level features into deeper high-level features. Zhang *et al.* [40] design Taylor finite difference-inspired edge blocks and direction attention aggregation blocks, effectively addressing challenges in detecting the shape of infrared small targets. In addition, Zhang *et al.* [38] try to introduce pruning into small target detection, and used wavelet pruning rules and regularization methods to achieve infrared efficient pruning. Zhu *et al.* [44] design a group of cross stage partial networks and a spatial attention module with global average contrast to obtain local and global spatial semantics. Jia *et al.* [17] abandon the global transformer and the convolutional sliding window of CNN, regarde the local area of the image as a graph node, and apply the graph neural network to infrared small target detection. In order to improve the multi-scale perception ability of the network, Fang *et al.*[12] designed a scale-adaptive feature enhancement mechanism and an attention-guided cross-weighted feature aggregator. While these single-frame detections excel in feature extraction for stationary targets, applying them to moving small targets faces performance limitations due to unique challenges.

2.2 Infrared Small Target Detection in Video

Multi-frame detection, with its unique and rich temporal information, outperforms single-frame detection in both accuracy and speed. Traditional multi-frame detection algorithms typically employ methods like energy accumulation. For instance, Zhang *et al.*

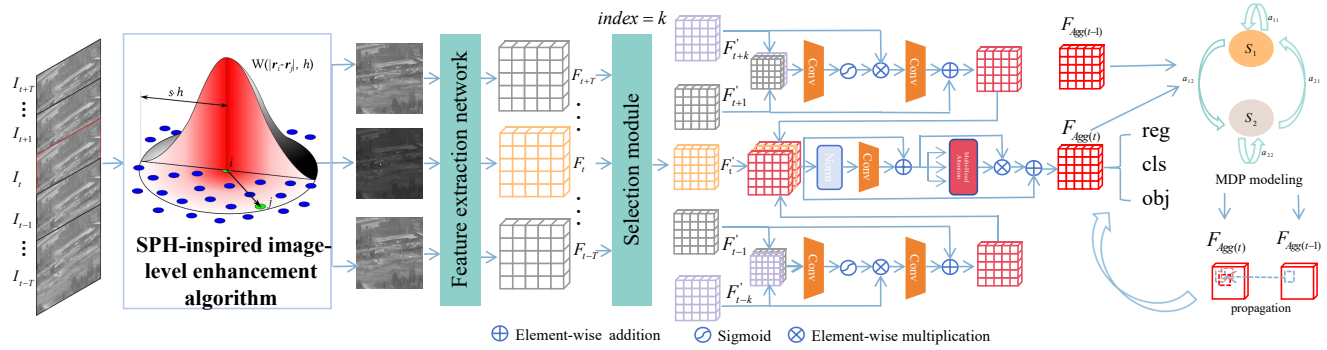

Figure 2: Overview of the proposed method. On the left is the designed SPH-inspired image-level enhancement algorithm (Section 3.2), which is used in the image preprocessing stage. On the right is the designed MDP-guided temporal feature perception module (Section 3.3), which consists of three parts, inculding the frame selection module, the temporal feature aggregation network and the frame propagation module based on Markov decision modeling.

[35] simplify the three-dimensional spatiotemporal information into a two-dimensional search, accumulating target energy based on motion direction. Background modeling methods include Zhou *et al.* [41] modeling the current frame and background in the Fourier domain, and Wang *et al.* [27] establishing a spatiotemporal tensor model, representing target extraction as a low-rank and sparse tensor decomposition problem. However, traditional algorithms still face challenges such as poor robustness and high complexity, despite performing well in specific scenarios.

With the development of deep learning, especially in video object detection, the focus of multi-frame infrared small target detection has shifted towards deep learning. The emergence of two-stage Faster-RCNN algorithms [23], and the widespread application of one-stage, YOLO series object detection algorithms have driven advancements in video object detection. FGFA [46] distorts and aggregates adjacent frames onto the current frame through optical flow networks, enhancing target information. MEGA [5] utilizes both local and global temporal information to enhance detection in the current frame. YOLOV [24], based on a one-stage detector, achieves significant success in inference speed by borrowing ideas from region proposals in two-stage detection. TransVOD++[43] proposes a temporal Transformer to aggregate spatial object queries and feature memories of frames. However, these general detection methods excel in learning capabilities for textured medium or large-sized targets but may not universally apply to multi-frame infrared small target detection due to infrared imaging characteristics.

Zhou *et al.* [42] propose an infrared image preprocessing and enhancement algorithm, using techniques like clahe, histogram stretching, and automatic gamma adjustment to enhance each channel separately and extract abundant feature information. But spatial domain enhancement alone is ineffective for moving target scenarios. Yan *et al.* [33] design a multiscale spatiotemporal difference attention network to aggregate more temporal information in feature extraction, achieving a good balance between target discovery and background suppression. Similarly, Bai *et al.* [2] introduce a cross-connected bidirectional pyramid structure and variable ROI pooling to enhance spatiotemporal information. Nevertheless, aggregating temporal information from complex network structures increases training and prediction costs. To mitigate this, Fan *et al.* [11] combine a lightweight target detection network with target tracking strategies to introduce motion target detection into tracking. Yuan *et al.* [34] design a dedicated module for infrared small target detection and use prior predictions during inference to guide the final output. However, the proposed method only improves upon CIOU but does not fully utilize temporal information during prediction. In summary, the above mentioned method lacks robust mathematical modeling due to the complexity of scenes and variations in this field, hindering comprehensive description with simple mathematical models. This limitation constrains the development of such detection methods, potentially resulting in subpar performance of existing algorithms in real-world applications.

3 METHODOLOGY

3.1 Overall Architecture

The proposed method, illustrated in Figure 2, is based on the YOLOX [15] framework. Given a set of input frames Q with a range of $2T+1$, where $Q = \{I_{t-T}, I_{t-(T-1)}, \cdots, I_t, I_{t+1}, \cdots, I_{t+T}\}$, before entering the network, it is first modeled as a 3D spatiotemporal graph using the proposed SPH-inspired image-level enhancement algorithm (Section 3.2) for enhancing targets and suppressing backgrounds. At this point, the current frame I_t in Q is significantly enhanced. Subsequently, the entire set of frames is transmitted to a feature extraction network for feature extraction. The feature extraction network adopts an FPN+PAN structure, with all input frames sharing convolutional weights. Later, the output feature set $\{F_i\}, i = t-T, t-T+1, \cdots t, t+1, \cdots t+T$, enters the proposed MDP-guided temporal feature perception module (Section 3.3). This module comprises three parts: selection, aggregation, and propagation. Firstly, a selection module picks out a reference frame feature map F_s ($s \in Q, s \neq t$) more effective for the current frame feature map F_t. Then, F_t and F_s are jointly input into an aggregation network for fusion, based on a multi-head attention mechanism and multi-scale fusion network. Finally, the propagation module transfers detection results from the previous frame F_{t-1} ($t \geq 1$) to the current frame F_t, aiding in current frame detection.

3.2 SPH-inspired image-level enhancement algorithm

The local contrast method is commonly used for enhancing infrared small targets, but it typically employs square-shaped kernels and operates on single-frame images. In this study, we introduce an SPH-inspired image-level enhancement algorithm that models sequences as 3D spatiotemporal grids. It replaces square-shaped kernels with Gaussian elliptical kernels from SPH to enhance targets in both temporal and spatial dimensions. Inspired by SPH density fields, this approach is combined with Gaussian difference, as illustrated in Figure 3.

Because of the fixed filter size and variable target sizes, using a square kernel may blend target and background information. Employing a Gaussian elliptical kernel in SPH for local contrast accommodates diverse target sizes. The Gaussian ellipse expression is as follows:

$$\Omega : \frac{(x\cos\theta - y\sin\theta)^2}{(2\sqrt{2}L_{max})^2} + \frac{(x\cos\theta + y\sin\theta)^2}{(\sqrt{2}L_{max})^2} = 1, \tag{1}$$

where L_{max} denotes the maximum value among all target sizes. θ represents the rotation angle of the ellipse, with this paper is $\pi/4$.

Inspired by SPH, we regard particles in the fluid as targets and background in the IR video, where the mass of particles corresponds to pixel values. The continuous density field computed by SPH is the ratio of the total mass of particles within a local sampling volume to the volume of the sampling volume. Similarly, we can approximate the pixel value of the central pixel in the elliptical kernel by the ratio of the sum of pixel values within the elliptical kernel to the area of the kernel. The calculation formula is as:

$$I_{avg} = \frac{1}{S_\Omega}\sum_{i=1}^{N}\omega_i I_i, \tag{2}$$

where N represents all the pixels within the elliptical kernel, I_i is their corresponding pixel values, S_Ω denotes the area of the elliptical kernel, and ω stands for the weighting coefficient, which depends on the distance between the pixel and the central pixel.

Subsequently, we divide the elliptical kernel into 9 sub-windows along its major and minor axes. During the sliding process in the spatial dimension, calculate the maximum pixel value I_{max} in the central sub-window. Compute the average grayscale value g_i and the maximum pixel value g_{max} for various sub-windows around the ellipse. The final enhancement for single-image is expressed as:

$$E_t = \min_i \frac{I_{avg}^2}{g_i} \times \varepsilon\left(I_{max} - G_{max}\right), \tag{3}$$

where ε represents the unit step function. As indicated by the formula, when the target is located in the central sub-window, the target is enhanced at that point.

Based on the rotational symmetry of the Gaussian ellipse, Gaussian ellipse filtering is performed in the temporal dimension. This approach efficiently achieves edge detection and key point detection, aligning features across frames, and ultimately enhancing the target. The final aggregation enhancement formula is as follows:

$$E_{final} = N\left(\sum_{i=t-T}^{t+T} sub(f_{warp}\left(G_t * E_t - G_i * E_i\right), E_t)\right), \tag{4}$$

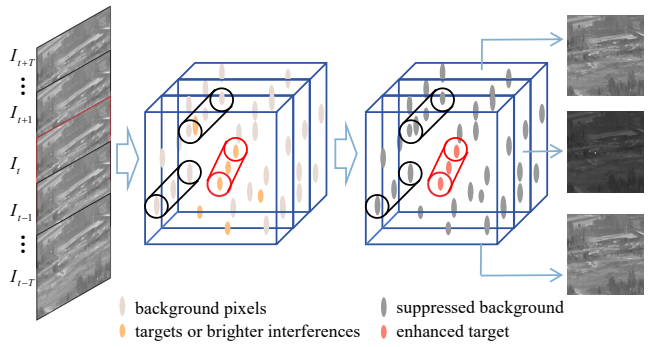

Figure 3: Overview of the SPH-inspired image-level enhancement algorithm. The black cylinder represents the motion trajectory of the background. It can be observed that, within a short-term frame, the background undergoes no significant changes, while the red cylinder exhibits a noticeable twist, allowing the capture of moving targets using temporal information.

where N denotes normalization, *sub* represents background subtraction, f_{warp} signifies feature alignment with the SIFT algorithm, i indicates the reference frame value, and G stands for the Gaussian difference function. The expression is as follows:

$$G\left(x, y\right) = \frac{1}{\sqrt{2\pi\sigma_1^2}} exp\left(-\frac{x^2+y^2}{2\sigma^2}\right). \tag{5}$$

3.3 MDP-guided temporal feature perception module

MDP display memorylessness, where the probability of future states depends only on the current state and is unaffected by past states. This property makes them well-suited for handling sequential data like image sequences, effectively capturing relationships between time intervals. Moreover, MDP effectively model the relationship between past data and future decisions, making them suitable for predicting target positions, states, and other information in time sequences. Accordingly, we can treat temporal feature perception as a prediction task by MPD. And we develop an MDP-guided temporal feature perception module to simplify modeling temporal data by abstracting it into states and corresponding transition probabilities. This reduction in complexity cuts computational costs and enhances interpretation of the model's behavior.

In fact, the MDP-guided temporal feature perception module is divided into three sub-modules: frame selection module, feature-level spatiotemporal information aggregation module, and frame-level prediction information propagation module.

Frame selection module. According to the backbone structure of YOLOX, the frame set $Q \in \mathbb{R}^{3\times T\times W\times H}$, after passing through the feature extraction network, produces outputs for three scales: $F_i^j \in \mathbb{R}^{C_j\times(2T+1)\times W_j\times H_j}$, $j = 1, 2, 3$, Here, C_j comprises three channels with values $[128, 256, 512]$, and $W_j = H_j = [64, 32, 16]$. For efficient reference frame selection, the second scale ($F_i^3 \in$

$\mathbb{R}^{256\times(2T+1)\times32\times32}$) is chosen to match the image scale required by the similarity calculation algorithm.

First, we compute the similarity between frames F_{t-1}, F_t, and F_{t+1} using the perceptual hash algorithm ($pHash$) as:

$$Sim_{(i)} = pHash\left(F_{t-1}, F_t, F_{t+1}\right), \tag{6}$$

where $pHash$ stands for the perceptual hash algorithm. This algorithm takes fixed-size 32×32 inputs and employs Discrete Cosine Transform (DCT) for pairwise image comparison, yielding the similarity value through Hamming distance.

After obtaining three sub-similarity values, they are normalized to the range (1,T), while acquiring the normalization weight ω_{nor}. The expression for ω_{nor} is:

$$\omega_{nor} = (T-1)\,/\left(Sim_{(imax)} - Sim_{(imin)}\right), \tag{7}$$

where ω_{nor} represents the normalized coefficient weight. The final index value is calculated as:

$$index = [1 + \omega_{nor} \times \frac{1}{N}\sum_{i} Sim_{(i)}]. \tag{8}$$

where [] represents rounding up, N is the total number of obtained similarity values. Frame selection effectively avoids redundancy in spatiotemporal information. High similarity between the current frame and adjacent frames may indicate occlusion, stationary targets, or slow motion. In such cases, the index obtains a larger value, directing attention to more distant and relevant frames. This aligns with human visual perception, extracting richer information from distant frames and eliminating redundant information from similar adjacent frames, facilitating target acquisition. These principles serve as the starting point for designing frame selection modules.

Feature aggregation module. After selecting reference frames, the chosen feature set $\left\{F_i^{'}\right\}$, where $i = t-k, t-1, t, t+1, t+k$ ($k = index$), is fed into a lightweight aggregation network for spatiotemporal fusion. This network aggregates features from adjacent frames $F_{t+1}^{'}$ and further reference frames $F_{t+k}^{'}$, followed by fine aggregation with the current frame $F_t^{'}$. The aggregation process resembles a transformer's encoder-decoder structure, incorporating residual connections and multi-head attention modules to enhance temporal features. Due to network symmetry, the explanation is based on one side's structure. Initially, the spatiotemporal information of the distant reference frame F_{t+k} is aggregated with F_{t+1}. The aggregation formula is as follows:

$$F_{Agg}^{1} = \sigma\left[f\left(F_{t+k}\right), f\left(F_{t+1}\right)\right] \otimes f\left(F_{t+k}\right) \oplus f\left(F_{t+1}\right), \tag{9}$$

where σ represents the sigmoid activation function, [] denotes concatenation, $\otimes$ is element-wise multiplication, $\oplus$ indicates element-wise addition, and f stands for the convolution operation. Similarly, F_{Agg}^{2} is derived from F_{t-k} and F_{t-1}. F_{Agg}^{1} and F_{Agg}^{2} now represent feature maps obtained by fusing temporal features from adjacent and distant reference frames. Subsequently, they are concatenated with the current frame in preparation for the final temporal feature aggregation. The formula is as follows:

$$F_{Agg}^{'} = \left[F_{Agg}^{1}, F_t^{'}, F_{Agg}^{2}\right] \oplus PE, \tag{10}$$

where PE is the added positional encoding, $F_t^{'}$ represents the current frame. The input $F_{Agg}^{'}$ undergoes initial feature extraction through normalization and convolutional layers. The formula is:

$$F_{Agg}^{''} = f\left(Norm\left(F_{Agg}^{'}\right)\right) \oplus F_{Agg}^{'}, \tag{11}$$

where $Norm$ denotes normalization, f represents the convolution operation. Following enhancement processing via the multi-head attention module, we obtain the final feature map F_{Agg}, defined as follows:

$$F_{Agg} = \Phi\left(F_{Agg}^{''}\right) \otimes F_{Agg}^{''} \oplus F_{Agg}^{'}. \tag{12}$$

where Φ represents the multi-head attention module.

Predictive propagation module. After feature aggregation, the aggregated frame F_{Agg} is fed into the decoupling head for information prediction. The predicted results, including target coordinates, classification information, and object presence, are then decoded. When $t \geq 1$, we model the motion of small IR targets using an MDP. We utilize the detection results from the previous frame to correct the detection results for the current frame. The MDP consists of quintuplicate variables: $MDP(S, A, P, R, \pi)$. Here, S represents target states, categorized as presence or absence. A stands for the actions performed between states, including detecting the IR target in both frames, not detecting the IR target in both frames, detecting the IR target in the previous frame but not in the subsequent frame, and not detecting the IR target in the previous frame but detecting it in the subsequent frame. P denotes the set of state transitions, while R signifies the reward function, indicating the rewards obtained when different actions are taken in a certain state. The policy π defines the actions A that the model may take under various states S, along with their corresponding probabilities. By learning each policy, we obtain locally maximal rewards and ultimately achieve the optimal result for the entire detection process. Additionally, we define a dynamic frame variable k, storing the most recent IR target detection outcome.

If the IR target is detected in the current frame, we match the policy with the result from the previous frame. The reward function at this point is defined as follows:

$$R_{t-1\to t}^{1} = \varepsilon\left(\min_{i,j}\left(\sqrt{\left(x_t^i - x_{t-1}^j\right)^2 + \left(y_t^i - y_{t-1}^j\right)^2}\right)\right), \tag{13}$$

where i and j are the top five highest-scoring results detected in the current frame and the previous frame, respectively. ε represents the IR target state of the previous frame, defines as follows:

$$\varepsilon = \begin{cases} 1, & if\ box\ exists; \\ 0, & if\ box\ is\ lost; \end{cases} \tag{14}$$

If the IR target remains undetected in the current frame, the reward function is as follows:

$$R_{t-1\to t}^{2} = \varepsilon \times \tau_{t-1} + (1-\varepsilon) \times \tau_k, \tag{15}$$

where τ_k represents the results detected in the most recent frame k. Actually, the overall reward function is defined as follows:

$$\begin{aligned} R_{t-1\to t} = p[q \times R_{t-1\to t}^{1} + (1-q) \times R_{t-1\to t}^{2}] + \\ (1-p)[q \times R_{t-1\to t}^{2} + (1-q) \times R_{t-1\to t}^{1}]. \end{aligned} \tag{16}$$

where p and q have the same expression as ε, representing the values of the previous frame and the current frame in two different states S.

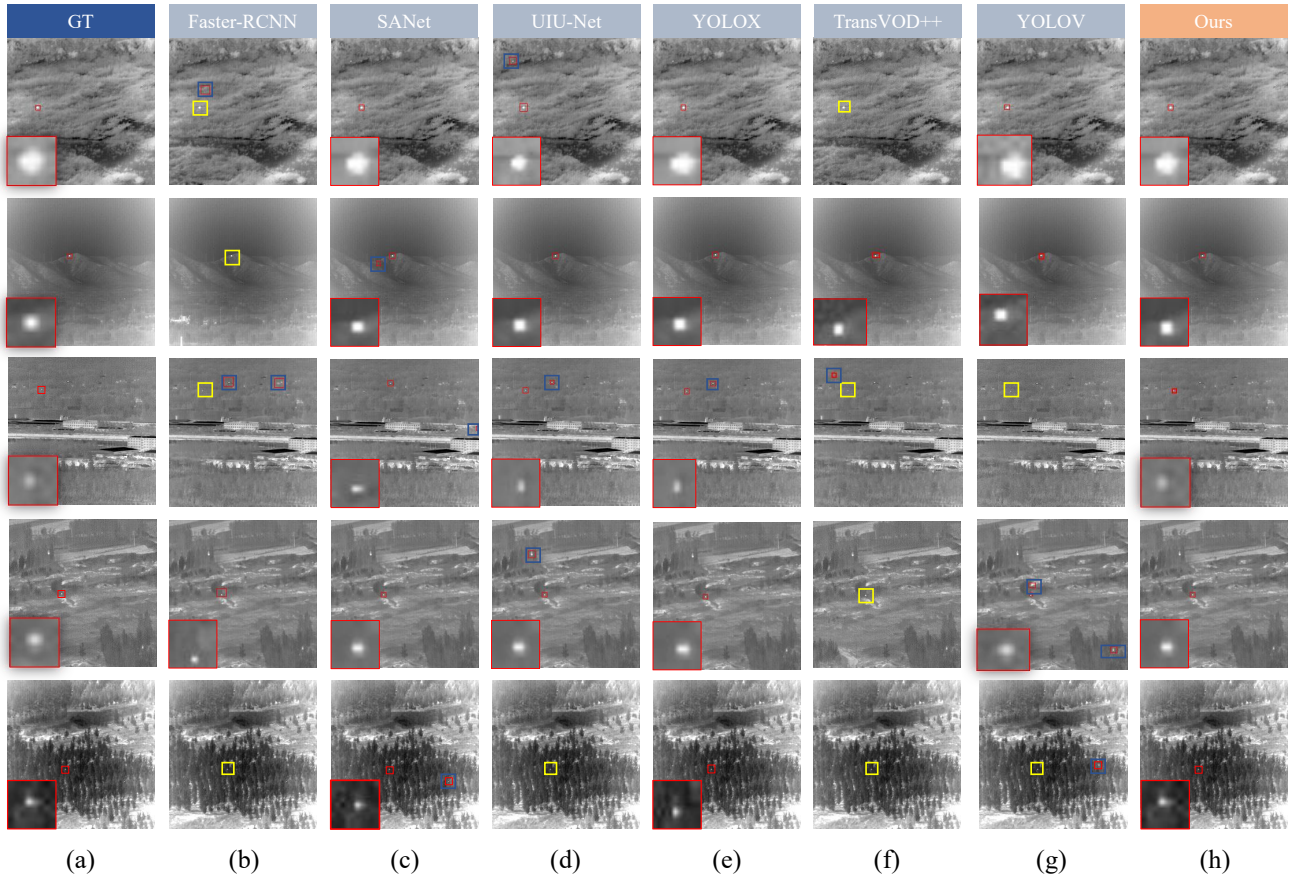

Figure 4: Visual results of different methods. The red boxes represent correctly detected targets, the blue boxes represent false alarms, and the yellow boxes represent missed detections. The first column represents the ground truth and the rest of the columns represent the visual result of Faster-RCNN, SANet, UIU-Net, YOLOX, TransVOD++, YOLOV, and Ours respectively.

Finally, the adjusted IR target box and score information are compared with the ground truth for loss calculation, and the ultimate loss expression is as follows:

$$L_{total} = \lambda L_{reg} + L_{cls} + L_{obj}. \tag{17}$$

where L_{reg} indicates the regression loss, L_{cls} is the classification loss, and L_{obj} represents the confidence loss.

4 EXPERIMENT

4.1 Dataset and Implementation Details

4.1.1 Dataset. We conduct extensive experiments on the proposed method using two publicly infrared small target datasets DAUB and DATR, along with comprehensive comparative and thorough ablation experiments. The DAUB dataset comprises various scenarios under sky and ground backgrounds, with a total of 22 video sequences. We select 18 sequences that meet the definition of small targets and divide the dataset into a 7:3 training-validation ratio. The training set includes 11 sequences with a total of 9734 frames, while the validation set comprises 6 sequences with 4044 images. The background of the DATR dataset is relatively simple, mainly for tracking and detecting vehicles, but it contains more targets per frame. The DATR dataset comprises 87 video sequences, with each sequence divided into 250 frames. Sequences 1-76 form the training set with 19000 images, and sequences 77-87 constitute the validation set with 2500 images.

4.1.2 Implementation Details. For all experiments, we standardize input images to 512×512 and apply the same data augmentation strategy. During training, we utilize the SGD optimizer with an initial learning rate of 0.01, momentum of 0.937, weight decay of 5×10^{-4}, and a learning rate reduction factor of 0.1. For DAUB dataset, the maximum training epochs are set to 100, with early termination if performance do not change over multiple epochs, while for DATR dataset, the maximum training epochs are set to 20. The batch size is set to 8, and during the training process, confidence threshold is set to 0.65, and non-maximum suppression is set to 0.3. All experiments are conducted on two NVIDIA RTX-3080 GPUs.

4.1.3 Evaluation Metrics. We use object-level evaluation metrics to assess our model's performance, including precision, recall, and F1 score. Precision represents the probability of correct predictions, recall represents the probability of accurate predictions, and the F1 score is the harmonic mean of precision and recall, reflecting the balance between the two. The definition of these metrics are as

Table 1: Comparison of different methods on the DAUB dataset.

Method	Pre(%)	Rec(%)	F1(%)	mAP50(%)
UIU-Net[31]	88.02	94.1	90.96	82.13
DNANet[20]	93.54	96.18	94.84	89.32
SANet[44]	92.99	96.11	94.52	83.3
Faster-RCNN[23]	45.28	57.16	50.57	40.9
YOLOv5[18]	91.45	95.82	93.58	88.83
YOLOX[15]	95.93	92.95	94.42	88.97
YOLOv8[19]	94.2	59.4	72.86	77.26
YOLOv[24]	91.58	80.85	85.88	72.62
TransVOD++[43]	83.78	65.34	73.42	54.48
Ours	**97.38**	**97.04**	**97.21**	**94.26**

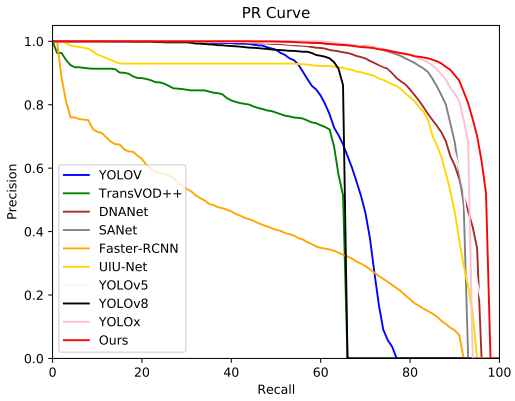

Figure 5: PR curves of different methods on DAUB dataset.

Table 2: Comparison of different methods on the DATR dataset.

Method	Pre(%)	Rec(%)	F1(%)	mAP50(%)
UIU-Net[31]	98.51	93.32	96.21	92.32
DNANet[20]	97.13	84.39	90.31	81.36
SANet[44]	98.39	92.55	95.38	91.63
Faster-RCNN[23]	75.82	88.92	81.85	66.13
YOLOv5[18]	89.44	95.74	92.48	84.27
YOLOX[15]	98.72	95.43	97.00	94.02
YOLOv8[19]	93.97	88.81	91.32	82.29
YOLOv[24]	98.56	96.12	97.63	93.60
TransVOD++[43]	70.65	62.91	66.56	43.50
Ours	**99.32**	**97.44**	**97.86**	**96.13**

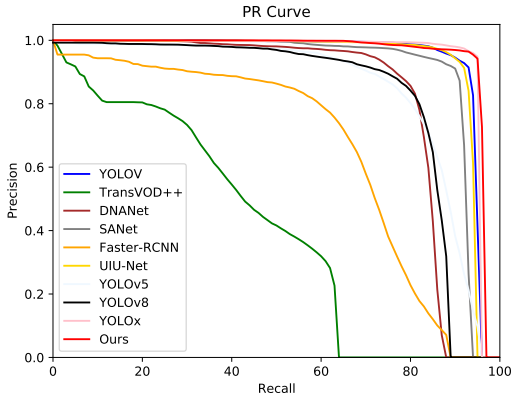

Figure 6: PR curves of different methods on DATR dataset.

follows:

$$Precision = \frac{TP}{TP + FP}, \tag{18}$$

$$Recall = \frac{TP}{TP + FN}, \tag{19}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{20}$$

where *TP*, *FP*, *FN* denote the true positive, false positive, false negative, respectively.

In addition, we compute the PR curve of the model. The PR curve reflects the relationship between precision and recall at different confidence levels. From this curve, we derive mAP using the following formula:

$$mAP = \frac{1}{n} \sum_{1}^{n} \int_{0}^{1} precision(recall) d(recall). \tag{21}$$

In this study, n denotes the number of target categories, which is 1. Additionally, we use an IOU threshold of 0.5 for computing mAP, referred to as mAP50.

4.2 Quantitative Results

We quantitatively compare the proposed method's performance based on mAP50, precision, recall rate, and F1 score, along with Precision-Recall (P-R) curves. Tables 1 and 2 present the metrics for the 9 object detection methods, including UIU-Net, DNA-Net, SANet, Faster-RCNN, YOLOv5, YOLOX, YOLOv8, TransVOD++ and YOLOV. Our method consistently outperforms others across all metrics, showing superior detection performance. For instance, on the DAUB dataset, our method improves precision, recall, F1, and mAP50 by 3.84%, 0.86%, 2.37%, and 4.94% over the second-ranked methods(DNA-Net), respectively. Notably, Faster-RCNN performs the poorest, potentially due to excessive candidate box generation. While YOLO series detectors show promising results, image enhancement algorithms in preprocessing may overshadow objects, affecting overall performance. Infrared small target detection algorithms achieve excellent results but are hampered by complex networks, leading to slow training and inference speeds. Video detection algorithms like YOLOV and TransVOD++ exhibit poor performance due to unsuitable temporal feature aggregation networks for infrared small targets. While on DATR dataset, our method likewise achieves the best results on four metrics. It improves precision, recall, F1, and mAP50 by 0.6%, 2.01%, 0.86%, and 2.11% over the second-ranked methods(YOLOX). Compared with the DAUB dataset, most of the methods have achieved better improvement on the DATR dataset, and we believe that the reason is that the small targets in the DATR dataset are relatively larger,

the background is relatively simple. It also has to do with how the dataset is divided.

The Precision-Recall (P-R) curve, illustrated in Figure 5 and 6, is a crucial comprehensive metric. It computes precision and recall at various thresholds, revealing the correlation between them and evaluating the relevance of detection results. A curve closer to the upper right corner signifies superior network performance. In Figure 5, our curve, highlighted in red, notably covers almost all the compared methods in the upper right corner, demonstrating our network's superior balance between accuracy and recall, resulting in the best overall performance. In addition, it can be seen that the curve of Faster-RCNN is the most flat, and the curve of YOLOV and YOLOv8 ends quickly with the increase of recall. The curves of the rest of the methods are not much different, but they are all below our curve as a whole. This is consistent with the above analysis and experimental results.

4.3 Visual Results

For a more intuitive comparison of contrastive effects, we select six existing methods for visual comparison with our network. And we choosed several typical scenarios in the DAUB dataset, including mountains, clouds, cities, and forests. As depicted in Figure 4, our method accurately locates small targets without producing missed detections or false positives at the same IOU threshold, and it can be clearly seen that the results of our method are highly consistent with the groundtruth. Faster-RCNN exhibits the highest number of missed detections and false positives. This is in line with the results of quantitative experiments. Among single-frame infrared small target detection networks, SANet achieves the highest accuracy but still encounters false positives and missed detections. In the case of the two video detectors, TransVOD++ demonstrates subpar visual results, possibly due to challenges in training as a transformer detector and limited applicability to multi-frame infrared small target detection as a general detection framework. YOLOV has a good detection effect in simple backgrounds, but missed detections and false detections occur in complex backgrounds and very small target situations. Because of the image preprocessing stage of YOLO causes confusion of complex backgrounds and targets, resulting in submerged targets. YOLOX showed good results, but their detection boxes did not have the best coincidence with the groundtruth.

4.4 Ablation Study

To validate the effectiveness of our proposed modules, we conduct ablation experiments, with results shown in Table 3. The base network solely utilizes YOLOX as the detector, without integrating the SPH-inspired image-level enhancement algorithm and MDP-guided temporal feature perception module. We then progressively add these components to YOLOX and observed their impact. Results indicate significant enhancements when incorporating both the MDP-guided temporal feature perception module and the SPH-inspired image-level enhancement algorithm into the base framework. Specifically, on both datasets, the baseline with the SPH-inspired image-level enhancement algorithm improves mAP50 by 1.38% and 1.52%, respectively. Similarly, the baseline with the MDP-guided temporal feature perception module improves mAP50 by 3.07% and 1.94%, respectively. Notably, combining both modules

Table 3: Ablation experiments for components of the proposed method on DAUB dataset.

Method	mAP50(%)	mAP0.50:0.95(%)
YOLOX	88.97	50.33
+MDP-guided	92.04	52.27
+SPH-inspired	90.35	51.85
+MDP-guided+SPH-inspired	94.26	54.32

results in synergistic effects, elevating mAP50 by 5.29% and 3.99%, respectively, demonstrating a greater performance boost than the sum of their individual contributions.

Through ablation experiments, we observe a notable trend: the MDP-guided temporal feature perception module outperforms the SPH-inspired image-level enhancement algorithm. Further analysis reveals that while the SPH-inspired algorithm enhances targets and suppresses background using spatiotemporal information, it faces challenges in detecting small, dark infrared targets against bright backgrounds and clutter. In contrast, the subsequent MDP-guided temporal feature perception module leverages Markov modeling to address moving small targets and eliminate static bright backgrounds, resulting in significantly enhanced detection performance. Essentially, the SPH-inspired algorithm provides coarse localization, reducing false negatives, while the MDP-guided module offers fine localization, reducing both false negatives and false positives. Consequently, comprehensive analysis supports the superiority of the MDP-guided temporal feature perception module in enhancing infrared small targets compared to the SPH-inspired image-level enhancement algorithm.

5 CONCLUSION

This paper presents a multi-frame infrared small target detection network by finding hybird models: SPH and MDP. SPH simulates fluid behavior by dividing it into particles and modeling their interactions. It can also simulate information propagation and interaction in images, enhancing them. MDP, known for their memorylessness, is effective for handling sequential data like image sequences, capturing relationships between time intervals. Accordingly, the MDP-guided temporal feature perception module effectively addresses challenges posed by complex backgrounds and occlusions, while the SPH-inspired image-level enhancement algorithm tackles issues arising from camera shake and motion blur. Results on the dataset demonstrate improved accuracy in target detection with reduced false positive and false negative rates. Ablation experiments highlight the contributions of the designed modules and algorithms to enhancing network performance. In the future, we can consider integrating data from different sensors (such as infrared and visible light) to enhance the performance and robustness of target detection. Further algorithm optimization is also essential to minimize power usage and enhance real-time capabilities, meeting the requirements of resource-constrained environments and real-time applications.

REFERENCES

[1] Xiangzhi Bai and Fugen Zhou. 2010. Analysis of new top-hat transformation and the application for infrared dim small target detection. *Pattern Recognition* 43, 6 (2010), 2145–2156.

[2] Yuanning Bai, Ruimin Li, Shuiping Gou, Chenchen Zhang, Yaohong Chen, and Zhihui Zheng. 2022. Cross-connected bidirectional pyramid network for infrared small-dim target detection. *IEEE Geoscience and Remote Sensing Letters* 19 (2022), 1–5.

[3] CL Philip Chen, Hong Li, Yantao Wei, Tian Xia, and Yuan Yan Tang. 2013. A local contrast method for small infrared target detection. *IEEE transactions on geoscience and remote sensing* 52, 1 (2013), 574–581.

[4] Shengjia Chen, Luping Ji, Jiewen Zhu, Mao Ye, and Xiaoyong Yao. 2024. SSTNet: Sliced spatio-temporal network with cross-slice ConvLSTM for moving infrared dim-small target detection. *IEEE Transactions on Geoscience and Remote Sensing* (2024).

[5] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. 2020. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10337–10346.

[6] De Cheng, Xiaojian Huang, Nannan Wang, Lingfeng He, Zhihui Li, and Xinbo Gao. 2023. Unsupervised visible-infrared person reid by collaborative learning with neighbor-guided label refinement. In *Proceedings of the 31st ACM International Conference on Multimedia*. 7085–7093.

[7] Yimian Dai and Yiquan Wu. 2017. Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection. *IEEE journal of selected topics in applied earth observations and remote sensing* 10, 8 (2017), 3752–3767.

[8] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. 2021. Asymmetric contextual modulation for infrared small target detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 950–959.

[9] Suyog D Deshpande, Meng Hwa Er, Ronda Venkateswarlu, and Philip Chan. 1999. Max-mean and max-median filters for detection of small targets. In *Signal and Data Processing of Small Targets 1999*, Vol. 3809. SPIE, 74–83.

[10] Jinming Du, Huanzhang Lu, Luping Zhang, Moufa Hu, Sheng Chen, Yingjie Deng, Xinglin Shen, and Yu Zhang. 2021. A spatial-temporal feature-based detection framework for infrared dim small target. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), 1–12.

[11] Jun Fan, Jingbiao Wei, Hai Huang, Dafeng Zhang, and Ce Chen. 2023. IRSDT: A Framework for Infrared Small Target Tracking with Enhanced Detection. *Sensors* 23, 9 (2023), 4240.

[12] Houzhang Fang, Zikai Liao, Lu Wang, Qingshan Li, Yi Chang, Luxin Yan, and Xuhua Wang. 2023. DANet: Multi-scale UAV Target Detection with Dynamic Feature Perception and Scale-aware Knowledge Distillation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 2121–2130.

[13] Ruigang Fu, Hongqi Fan, Yongfeng Zhu, Bingwei Hui, Zhilong Zhang, P Zhong, D Li, S Zhang, G Chen, and L Wang. 2022. A dataset for infrared time-sensitive target detection and tracking for air-ground application. *China Sci. Data* 7, 2 (2022), 206–221.

[14] Chenqiang Gao, Deyu Meng, Yi Yang, Yongtao Wang, Xiaofang Zhou, and Alexander G Hauptmann. 2013. Infrared patch-image model for small target detection in a single image. *IEEE transactions on image processing* 22, 12 (2013), 4996–5009.

[15] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430* (2021).

[16] Bingwei Hui, Zhiyong Song, Hongqi Fan, P Zhong, W Hu, X Zhang, J Lin, H Su, W Jin, Y Zhang, et al. 2019. A dataset for infrared image dim-small aircraft target detection and tracking under ground/air background. *Sci. Data Bank* 5 (2019), 12.

[17] Guimin Jia, Yu Cheng, and Tao Chen. 2024. IRGraphSeg: Infrared Small Target Detection Based on Hierarchical GNN. *IEEE Geoscience and Remote Sensing Letters* (2024).

[18] Glenn Jocher. 2020. *Ultralytics YOLOv5*. https://doi.org/10.5281/zenodo.3908559

[19] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. *Ultralytics YOLO*. https://github.com/ultralytics/ultralytics

[20] Boyang Li, Chao Xiao, Longguang Wang, Yingqian Wang, Zaiping Lin, Miao Li, Wei An, and Yulan Guo. 2022. Dense nested attention network for infrared small target detection. *IEEE Transactions on Image Processing* 32 (2022), 1745–1758.

[21] Lingyi Lu and Xin Xu. 2021. Visible-Infrared Cross-Modal Person Re-identification based on Positive Feedback. In *Proceedings of the 3rd ACM International Conference on Multimedia in Asia*. 1–6.

[22] Yihang Luo, Xinyi Ying, Ruojing Li, Yujun Wan, Bo Hu, and Qiang Ling. 2022. Multi-scale Optical Flow Estimation for Video Infrared Small Target Detection. In *2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*. IEEE, 129–132.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[24] Yuheng Shi, Naiyan Wang, and Xiaojie Guo. 2023. YOLOV: Making still image object detectors great at video object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 2254–2262.

[25] Xiaoliang Sun, Xiaolin Liu, Zhixuan Tang, Gucan Long, and Qifeng Yu. 2017. Real-time visual enhancement for infrared small dim targets in video. *Infrared physics & technology* 83 (2017), 217–226.

[26] Karasawa Takumi, Kohei Watanabe, Qishen Ha, Antonio Tejero-De-Pablos, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Multispectral object detection for autonomous vehicles. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*. 35–43.

[27] Guanghui Wang, Bingjie Tao, Xuan Kong, and Zhenming Peng. 2021. Infrared small target detection using nonoverlapping patch spatial–temporal tensor factorization with capped nuclear norm regularization. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), 1–17.

[28] Xing Wei, Diangang Li, Xiaopeng Hong, Wei Ke, and Yihong Gong. 2020. Co-attentive lifting for infrared-visible person re-identification. In *Proceedings of the 28th ACM international conference on multimedia*. 1028–1037.

[29] Yantao Wei, Xinge You, and Hong Li. 2016. Multiscale patch-based contrast measure for small infrared target detection. *Pattern Recognition* 58 (2016), 216–226.

[30] Tianhao Wu, Boyang Li, Yihang Luo, Yingqian Wang, Chao Xiao, Ting Liu, Jungang Yang, Wei An, and Yulan Guo. 2023. MTU-Net: Multilevel TransUNet for space-based infrared tiny ship detection. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–15.

[31] Xin Wu, Danfeng Hong, and Jocelyn Chanussot. 2022. UIU-Net: U-Net in U-Net for infrared small object detection. *IEEE Transactions on Image Processing* 32 (2022), 364–376.

[32] Yuyang Xi, Zhitao Zhou, Ying Jiang, Liuwei Zhang, Yunfei Li, Zhipeng Wang, Fanjiao Tan, and Qingyu Hou. 2023. Infrared moving small target detection based on spatial-temporal local contrast under slow-moving cloud background. *Infrared Physics & Technology* 134 (2023), 104877.

[33] Puti Yan, Runze Hou, Xuguang Duan, Chengfei Yue, Xin Wang, and Xibin Cao. 2023. STDMANet: Spatio-temporal differential multiscale attention network for small moving infrared target detection. *IEEE transactions on geoscience and remote sensing* 61 (2023), 1–16.

[34] Shudong Yuan, Bei Sun, Zhen Zuo, Honghe Huang, Peng Wu, Can Li, Zhaoyang Dang, and Zongqing Zhao. 2023. IRSDD-YOLOv5: Focusing on the Infrared Detection of Small Drones. *Drones* 7, 6 (2023), 393.

[35] Fei Zhang, Chengfang Li, and Lina Shi. 2005. Detecting and tracking dim moving point target in IR image sequence. *Infrared Physics & Technology* 46, 4 (2005), 323–328.

[36] Jing Zhang and Dacheng Tao. 2020. Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things. *IEEE Internet of Things Journal* 8, 10 (2020), 7789–7817.

[37] Mingjin Zhang, Haichen Bai, Jing Zhang, Rui Zhang, Chaoyue Wang, Jie Guo, and Xinbo Gao. 2022. Rkformer: Runge-kutta transformer with random-connection attention for infrared small target detection. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1730–1738.

[38] Mingjin Zhang, Handi Yang, Jie Guo, Yunsong Li, Xinbo Gao, and Jing Zhang. 2024. IRPruneDet: Efficient Infrared Small Target Detection viaWavelet Structure-Regularized Soft Channel Pruning. In *Proceedings of the 38th Annual AAAI Conference on Artificial Intelligence*. 1857–1865.

[39] Mingjin Zhang, Ke Yue, Jing Zhang, Yunsong Li, and Xinbo Gao. 2022. Exploring feature compensation and cross-level correlation for infrared small target detection. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1857–1865.

[40] Mingjin Zhang, Rui Zhang, Yuxiang Yang, Haichen Bai, Jing Zhang, and Jie Guo. 2022. ISNet: Shape matters for infrared small target detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 877–886.

[41] Anran Zhou, Weixin Xie, and Jihong Pei. 2021. Background modeling combined with multiple features in the Fourier domain for maritime infrared target detection. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), 1–15.

[42] Jinjie Zhou, Baohui Zhang, Xilin Yuan, Cheng Lian, Li Ji, Qian Zhang, and Jiang Yue. 2023. YOLO-CIR: The network based on YOLO and ConvNeXt for infrared object detection. *Infrared Physics & Technology* 131 (2023), 104703.

[43] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. 2022. TransVOD: end-to-end video object detection with spatial-temporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[44] Jiewen Zhu, Shengjia Chen, Lexiao Li, and Luping Ji. 2023. Sanet: Spatial attention network with global average contrast learning for infrared small target detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[45] Wenming Zhu and Yihua Tan. 2023. A moving infrared small target detection method based on optical flow-guided neural networks. In *2023 4th International conference on computer vision, image and deep learning (CVIDL)*. IEEE, 531–535.

[46] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. 2017. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE international conference on computer vision*. 408–417.