Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Modality-Guided Dynamic Graph Fusion and Temporal Diffusion for Self-Supervised RGB-T Tracking

Shenglan Li[1,2], **Rui Yao**[1,2,*], **Yong Zhou**[1,2], **Hancheng Zhu**[1,2], **Kunyang Sun**[1,2], **Bing Liu**[1,2], **Zhiwen Shao**[1,2], **Jiaqi Zhao**[1,2]

[1]School of Computer Sciences and Technology, China University of Mining and Technology
[2]Mine Digitization Engineering Research Center of the Ministry of Education, China
{shenglanli, ruiyao, yzhou, zhuhancheng, kunyang_sun, liubing, zhiwen_shao}@cumt.edu.cn

Abstract

To reduce the reliance on large-scale annotations, self-supervised RGB-T tracking approaches have garnered significant attention. However, the omission of the object region by erroneous pseudo-label or the introduction of background noise affects the efficiency of modality fusion, while pseudo-label noise triggered by similar object noise can further affect the tracking performance. In this paper, we propose GDSTrack, a novel approach that introduces dynamic graph fusion and temporal diffusion to address the above challenges in self-supervised RGB-T tracking. GDSTrack dynamically fuses the modalities of neighboring frames, treats them as distractor noise, and leverages the denoising capability of a generative model. Specifically, by constructing an adjacency matrix via an Adjacency Matrix Generator (AMG), the proposed Modality-guided Dynamic Graph Fusion (MDGF) module uses a dynamic adjacency matrix to guide graph attention, focusing on and fusing the object's coherent regions. Temporal Graph-Informed Diffusion (TGID) models MDGF features from neighboring frames as interference, and thus improving robustness against similar-object noise. Extensive experiments conducted on four public RGB-T tracking datasets demonstrate that GDSTrack outperforms the existing state-of-the-art methods. The source code is available at https://github.com/LiShenglana/GDSTrack.

1 Introduction

The RGB-T tracking task, which leverages the complementary strengths of RGB images (rich texture information) and thermal infrared images (enhanced nighttime perception), has gained increasing attention in recent years [Lai *et al.*, 2024; Hu *et al.*, 2024]. However, the reliance on manual annotation for both modalities is labor-intensive and resource-demanding [Zhang *et al.*, 2024], and the inherent inconsistency in cross-modal annotation further complicates the creation of high-quality RGB-T datasets. Consequently, the development of robust self-supervised methods for RGB-T tracking [Li *et al.*, 2024] has become crucial and offers significant practical advantages.

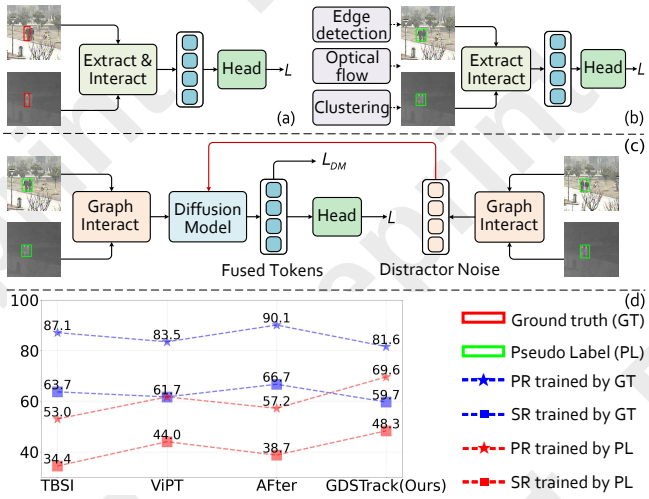

Figure 1: (a) Existing RGB-T tracking methods rely heavily on a large number of ground truth annotations, making it challenging to handle larger-scale RGB-T datasets under supervised settings. (b) The discrepancy between the pseudo label and ground truth makes accurate object tracking difficult, causing performance degradation when used for the fully supervised training of sota trackers. (c) We leverage modality-guided dynamic graph fusion and temporal diffusion to address the issues of object irrelevant fusion and distractor noise caused during the training process supervised by pseudo-label. (d) Compared with the three trackers, our proposed method achieved the best performance under pseudo-label supervision.

To fully leverage information in images [Xi *et al.*, 2023], a class of mainstream methods relies on pseudo-labels [Zheng *et al.*, 2021; Shen *et al.*, 2022; Zhang *et al.*, 2025] for supervision. USOT [Zheng *et al.*, 2021] relies on unsupervised optical flow to detect moving objects and applies a dynamic programming algorithm to establish interframe associations, thereby generating pseudo-labels to provide supervision. However, a discrepancy between the pseudo-labels and ground truth is inevitable. For example, comparing the accurate red bounding box in Fig.1(a) with the pseudo green bounding box in Fig.1(b), it is evident that the pseudo-label incorrectly includes both the object and a similar moving ob-

*Corresponding author.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

ject, thereby further exacerbating performance degradation. As illustrated in Fig. 1(d), when training with pseudo-labels, a significant performance degradation is observed in TBSI [Hui *et al.*, 2023], ViPT [Zhu *et al.*, 2023], and AFter [Lu *et al.*, 2024]. Obviously, this performance degradation stems from pseudo-label noise, which arises from two primary factors: (1) inaccurate pseudo-labels may fail to capture whole object regions and always introduce excessive background noise, and (2) inaccurate annotation caused by interference from similar objects will significantly degrade the model performance. To address these two problems, we designed a self-supervised RGB-T tracking method as shown in Fig. 1(c).

In fact, label noise also exists in fully supervised tasks, where misalignment between modality labels can introduce label noise. GMMT [Tang *et al.*, 2024] models noise as Gaussian noise, and uses a generative model [Song *et al.*, 2020; Cao *et al.*, 2025] to enhance its ability to perceive noise. However, when dealing with pseudo-labels, simple Gaussian noise is insufficient for modeling the two types of labeling errors mentioned above. To address the pseudo-label noise issue, we propose our method. Specifically, to tackle the first issue, we introduce the Modality-Guided Dynamic Graph Fusion (MDGF) module, which dynamically adjusts the adjacency matrix based on the similarity between the input RGB and thermal infrared modalities. By leveraging a dynamic adjacency matrix to guide graph attention in mining and focusing on object regions, a fusion that supports tracking is achieved, even in the presence of inaccurate annotations. Moreover, to study the second issue, we draw inspiration from GMMT [Tang *et al.*, 2024] and propose a Temporal Graph-Informed Diffusion (TGID) module, which treats fused information from neighboring frames as noise within a diffusion model. This approach helps the model to identify interference from similar objects and train it to be more robust to such noise. Furthermore, the MDGF module complements the TGID module to prevent information loss and enhance overall model performance. In summary, our main contributions are as follows:

- We propose a novel self-supervised RGB-T tracker GDSTrack, which leverages multi-modal optical flow to extract pseudo-labels for coarse localization, significantly reducing the cost of manual annotation.
- We design the MDGF module to simulate interference from similar objects in pseudo-labels and utilize the generative capability of the TGID module for denoising and fine localization, achieving a balance between performance and cost.
- Ablation studies and comparative experiments on four benchmarks demonstrate the state-of-the-art performance of our method and its ability to enhance self-supervised learning.

2 Related Work

Self-supervised Tracking. RGB-T tracking has advanced significantly owing to the complementary nature of RGB and thermal infrared modalities. [Lai *et al.*, 2024] leverage spatio-temporal contextual modeling through long-range cross-frame integration and short-term historical trajectory prompts. [Hu *et al.*, 2024] introduce a temporal state generator that produces temporal information tokens to guide the localization of the object in the next time state. Self-supervised RGB-T tracking [Li *et al.*, 2024] addresses the challenge of requiring large-scale manual annotations for tracking tasks. However, this topic has not received sufficient research attention. Most self-supervised tracking methods rely primarily on the RGB modality alone. The UDT [Wang *et al.*, 2019] exploits the discriminative power of correlation filtering for tracking and employs cycle consistency as a self-supervised signal. Existing deep trackers can be trained using synthesized data in routine ways, without requiring human annotation. [Sio *et al.*, 2020] exploit the fact that an image and any cropped region of it naturally form a pair for self-training. To replace naive cropping methods such as center cropping, [Zheng *et al.*, 2021] propose a more accurate pseudo-label generation method based on optical flow and dynamic programming techniques. Building on this, [Zhang *et al.*, 2025] learn a prompt representation of an object using a pre-trained diffusion model. However, these methods do not account for interference caused by similar objects in the absence of ground-truth labels. Our method tackles the root cause of the noise by simulating distractor noise using the modal fusion results of similar objects, thereby enhancing the robustness of the model to such interference.

RGB-T Fusion aims to fully utilize the complementary characteristics of RGB and thermal infrared images to enhance feature extraction and improve performance in downstream tasks. In terms of the fusion level, [Zhang *et al.*, 2019] consider several fusion mechanisms at the pixel, feature, and response levels. In terms of fusion methods, [Xu *et al.*, 2021] use a channel attention mechanism to implement the adaptive calibration of feature channels before realizing hierarchical feature fusion. [Tang *et al.*, 2023] propose a novel model for infrared and visible image fusion via a dual attention Transformer to represent long-range context information. [Tang *et al.*, 2024] seek to uncover the potential of generative techniques to address the critical challenge in multi-modal tracking. [Chen *et al.*, 2024] predict bounding boxes through a sequence-to-sequence framework. [Zhu *et al.*, 2023] introduce prompt learning for multi-modal fusion tracking method. In terms of frequency domain fusion, [Zhao *et al.*, 2023b] decompose modality-specific and modality-shared features to better reflect the semantic information contained in the high-frequency and low-frequency features. However, these methods do not account for the fusion process and focus on irrelevant areas because of the lack of ground truth labels. Our method can dynamically generate adjacency matrices, guiding the graph attention model to focus more on object regions.

3 Methodology

3.1 Framework Overview

Our model performs self-supervised tracking by fully leveraging the advantages of both RGB and infrared features through the Modality-Guided Dynamic Graph Fusion and Temporal Graph-Informed Diffusion modules, as shown in

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

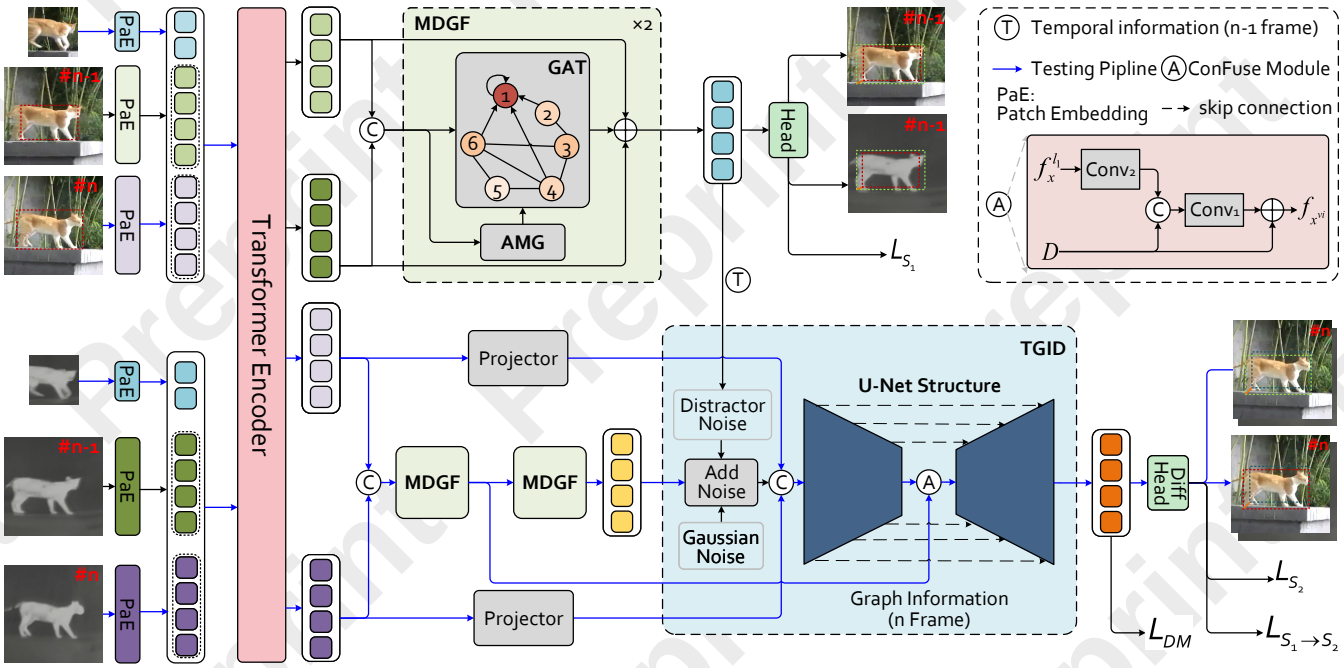

Figure 2: The pipeline of GDSTrack model. First, the encoder obtains the search frame features that interact with the template frame features. Subsequently, the MDGF module fuses the features of the two modalities using graph attention guided by dynamic adjacency matrix. Then the features obtained from the MDGF module are used as distractor noise and input for the TGID module to perform denoising and obtain the final tracking results. We use TGID to enhance the model's robustness to noise.

Fig. 2. GDSTrack obtains the interaction features through the encoder. The RGB and infrared images share encoder parameters. We first use the AMG module to dynamically adjust the adjacency matrix based on the features of the two modalities, guiding graph attention in MDGF to fuse the information from both modalities. The TGID module then denoises similar object noise simulated by the MDGF module. Finally, the output is fed into the tracking head for tracking. GDSTrack framework takes two RGB images and their corresponding infrared images as input: the visible template frame $\boldsymbol{z}^v \in \mathbb{R}^{3\times W_{z_0}\times H_{z_0}}$ cropped from the first visible frame $\boldsymbol{x}_1^v \in \mathbb{R}^{3\times W_{x_0}\times H_{x_0}}$, and the infrared template frame $\boldsymbol{z}^i \in \mathbb{R}^{1\times W_{z_0}\times H_{z_0}}$ cropped from the first infrared frame $\boldsymbol{x}_1^i \in \mathbb{R}^{1\times W_{x_0}\times H_{x_0}}$. The second visible search frame $\boldsymbol{x}_2^v \in \mathbb{R}^{3\times W_{x_0}\times H_{x_0}}$, the second infrared search frame $\boldsymbol{x}_2^i \in \mathbb{R}^{1\times W_{x_0}\times H_{x_0}}$, where W_{z_0} and H_{z_0} are half of W_{x_0} and H_{x_0}, respectively. We obtain the RGB and infrared search frame features, denoted as $\boldsymbol{f}_{x_1}^v \in \mathbb{R}^{W_x H_x\times d_{model}}$ and $\boldsymbol{f}_{x_1}^i \in \mathbb{R}^{W_x H_x\times d_{model}}$, respectively, after the interaction between the template and the search frame through the encoder. The MDGF module performs an initial fusion of the inputs $\boldsymbol{f}_{x_1}^v \in \mathbb{R}^{W_x H_x\times d_{model}}$ and $\boldsymbol{f}_{x_1}^i \in \mathbb{R}^{W_x H_x\times d_{model}}$, producing two layers of fused results $\boldsymbol{f}_{x_1}^{l_1}$ and $\boldsymbol{f}_{x_1}^{S_1}$.

$$\boldsymbol{f}_{x_1}^{l_1}, \boldsymbol{f}_{x_1}^{S_1} = MDGF(\boldsymbol{f}_{x_1}^v, \boldsymbol{f}_{x_1}^i), \tag{1}$$

where $MDGF(\cdot)$ indicates the Modality-guided Adaptive Graph Fusion Module (see Sect. 3.2). $\boldsymbol{f}_{x_1}^{S_1}$ is processed through the tracking head to obtain the tracking score $Score^{S_1}$ and the predicted bounding box B^{S_1}:

$$Score^{S_1}, B^{S_1} = Head(\boldsymbol{f}_{x_1}^{S_1}), \tag{2}$$

where we use a series of Conv-BN-ReLU layers to independently estimate the top-left and bottom-right corners following [Cui *et al.*, 2022]. We employ the Generalized Intersection over Union (GIoU) loss and L_1 loss to guide tracker training. The loss function in stage one is defined as follows:

$$L_{S_1} = \lambda_1 L_{GIoU}(B^{S_1}, \hat{B}) + \lambda_2 L_1(B^{S_1}, \hat{B}). \tag{3}$$

Where $\hat{B}$ is our pseudo label, and we set the loss weight λ_1 to 2 and set λ_2 to 5 following [Zhao *et al.*, 2023a]. The third search frame is then introduced to obtain the corresponding visible and thermal infrared features, named as $\boldsymbol{f}_{x_2}^v \in \mathbb{R}^{W_x H_x\times d_{model}}$ and $\boldsymbol{f}_{x_2}^i \in \mathbb{R}^{W_x H_x\times d_{model}}$.

We first use the MDGF module to obtain the first-layer and second-layer fusion results, denoted as $\boldsymbol{f}_{x_2}^{l_1}$ and $\boldsymbol{f}_{x_2}^{S_1}$:

$$\boldsymbol{f}_{x_2}^{l_1}, \boldsymbol{f}_{x_2}^{S_1} = MDGF(\boldsymbol{f}_{x_2}^v, \boldsymbol{f}_{x_2}^i). \tag{4}$$

We then utilize the TGID module to further denoise $\boldsymbol{f}_{x_2}^{S_1}$ from the MDGF to obtain more refined fused features $\boldsymbol{f}_{x_2}^{S_2}$:

$$\boldsymbol{f}_{x_2}^{S_2} = TGID(\boldsymbol{f}_{x_1}^{S_1}, \boldsymbol{f}_{x_2}^{l_1}, \boldsymbol{f}_{x_2}^{S_1}, \boldsymbol{f}_{x_2}^v, \boldsymbol{f}_{x_2}^i), \tag{5}$$

where $TGID(\cdot)$ indicates the Temporal Graph-Informed Diffusion Module (see Sect. 3.3). We use $\boldsymbol{f}_{x_2}^v$ and $\boldsymbol{f}_{x_2}^i$ as the initial conditions for the diffusion model, with $\boldsymbol{f}_{x_2}^{l_1}$ from the first layer of MDGF serving as the conditions for the intermediate layers of the denoising process. $\boldsymbol{f}_{x_1}^{S_1}$ is set as distractor noise,

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

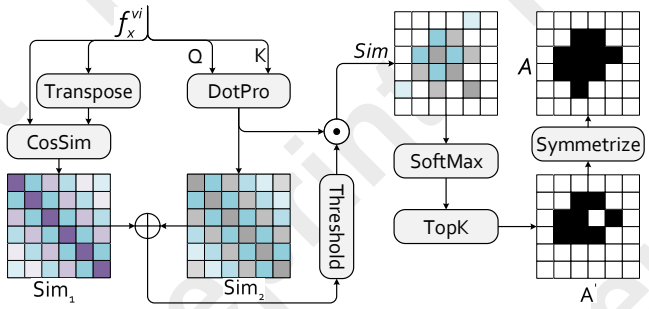

Figure 3: The pipeline of AMG. The AMG module concatenates visible and infrared features along the sequence dimension, then computes cosine similarity and dot-product attention with themselves. After summation and threshold filtering, they guide the dot-product attention to generate similarity mask. Finally, SoftMax, TopK, and Symmetrize are applied to obtain the adjacency matrix.

contributing to the noise during the added noise process. $\boldsymbol{f}_{x_2}^{S_2}$ is used to obtain the tracking results through DiffHead:

$$Score^{S_2}, B^{S_2} = DiffHead(\boldsymbol{f}_{x_2}^{S_2}), \tag{6}$$

where DiffHead shares the same structure as Head in MDGF. Then, we employ pseudo-labels $\hat{B}$ to supervise the training:

$$L_{S_2} = \lambda_1 L_{GIoU}(B^{S_2}, \hat{B}) + \lambda_2 L_1(B^{S_2}, \hat{B}). \tag{7}$$

We use the predicted bounding boxes $\hat{B}^{S_1}$ from the MDGF to supervise the training process of the diffusion model:

$$L_{S_1 \to S_2} = \lambda_1 L_{GIoU}(B^{S_2}, \hat{B}^{S_1}) + \lambda_2 L_1(B^{S_2}, \hat{B}^{S_1}). \tag{8}$$

The final loss function in stage two is defined as follows:

$$L = L_{S_2} + L_{S_1 \to S_2} + \lambda L_{DM}, \tag{9}$$

where L_{DM} is the generative loss.

3.2 Modality-guided Dynamic Graph Fusion

The MDGF consists of an Adjacency Matrix Generator and a Graph Attention Network. AMG generates a dynamic adjacency matrix based on the similarity between RGB and infrared modalities to guide the graph attention network. This similarity-based fusion method enables GDSTrack to focus more on the coherent regions of the object, even in the absence of ground truth labels.

Adjacency Matrix Generator. The AMG module employs Cosine Similarity to create a mask that guides the cross-attention mechanism in generating the corresponding adjacency matrix according to feature variations (see Fig. 3).

We concatenate $\boldsymbol{f}_x^v$ and $\boldsymbol{f}_x^i$ along the sequence dimension:

$$\boldsymbol{f}_x^{vi} = Concat(\boldsymbol{f}_x^v, \boldsymbol{f}_x^i), \tag{10}$$

where we get $\boldsymbol{f}_x^{vi} \in \mathbb{R}^{2W_xH_x \times d_{model}}$. Then, we compute the similarity Sim_1 between $\boldsymbol{f}_x^v$ and $\boldsymbol{f}_x^i$ using the scaled dot product attention [Vaswani *et al.*, 2017]:

$$Q = \boldsymbol{f}_x^{vi} W^Q, K = \boldsymbol{f}_x^{vi} W^K, S_1 = \frac{QK^{\mathrm{T}}}{\sqrt{d_k}}, \tag{11}$$

where the projections are the parameter matrices $W^Q \in \mathbb{R}^{d_{model} \times d_k}$ and $W^K \in \mathbb{R}^{d_{model} \times d_k}$. $\frac{1}{\sqrt{d_k}}$is a scaling factor and d_k is 2048. We then calculate the Cosine Similarity Sim_2 between $\boldsymbol{f}_x^v$ and $\boldsymbol{f}_x^i$:

$$S_2 = CosineSim(\boldsymbol{f}_{x^{vi}}, \boldsymbol{f}_{x^{vi}}^{\mathrm{T}}). \tag{12}$$

We sum the two similarities and then shift their values to the range $[0, 1]$:

$$M = (S_1 + S_2 + 1)/2, \tag{13}$$

where $M \in \mathbb{R}^{2W_xH_x \times 2W_xH_x}$.

We use M as a mask, setting the similarity values to less than the threshold θ to negative infinity, and retaining the similarity values from Sim_1 otherwise:

$$\begin{cases} S_{i,j} = S_{1_{i,j}}, if M > \theta, \\ S_{i,j} = -\infty, if M \le \theta. \end{cases} \tag{14}$$

Then we get the final similarity:

$$S' = softmax(S). \tag{15}$$

We then set the Top-K positions with the largest similarity values in the final similarity matrix to 1, while setting all other positions to zero:

$$\begin{cases} A'_{i,j} = 1, if\ S'_{i,j} \ge \text{the } k\text{-th largest values of } S', \\ A'_{i,j} = 0, otherwise. \end{cases} \tag{16}$$

To ensure the symmetry of adjacency A':

$$A = (A' + A'^{\mathrm{T}})/2, \tag{17}$$

where $A \in \mathbb{R}^{2W_xH_x \times 2W_xH_x}$.

Graph Attention Networks. We utilize a two-layer Graph Attention Network (GAT) [Velickovic *et al.*, 2017] to perform graph-based attention fusion of visible and thermal infrared features under the guidance of the adjacency matrix generated by the AMG module:

$$\boldsymbol{f}_x^{l_1}, \boldsymbol{f}_x^{S_1} = GAT(\boldsymbol{f}_x^{vi}, A). \tag{18}$$

GAT takes each pixel in the concatenated modalities $\boldsymbol{f}_x^{vi}$ as a graph node, and the adjacency matrix A preserves both intra-modal and inter-modal relationships. The input features are processed using weighted learning, where the attention mechanism adaptively assigns weights based on the correlations between features.

In the first layer, a learnable matrix $W_1 \in \mathbb{R}^{d_{model} \times nhid}$ is used to perform a linear transformation of features $\boldsymbol{f}_x^{vi}$. Then, a shared attentional mechanism and softmax function is performed following the GAT model. Similarly, a learnable vector $W_2 \in \mathbb{R}^{nhid \times out_{model}}$ is used to perform a linear transformation of the output of the first layer, $\boldsymbol{f}_x^{layer1}$.

To reduce information loss, the output of the second layer $\boldsymbol{f}_x^{l2}$ is added to the visible and infrared features after the dimension reduction using a 1×1 convolution as $Proj(\cdot)$:

$$\boldsymbol{f}_x^{S_1} = Add(split(\boldsymbol{f}_x^{l_2})) + Proj(\boldsymbol{f}_x^v + \boldsymbol{f}_x^i), \tag{19}$$

where $split$ is the inverse of Concatenation.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

3.3 Temporal Graph-Informed Diffusion

To address the noise from similar objects during pseudo-label training, we designed the TGID module. It treats the MDGF results of nearby frames as noise using the first-layer MDGF features to improve robustness to noise and prevent information loss. Specifically, after obtaining the fusion results $\boldsymbol{f}_{x_2}^{S_1}$ from MDGF, we introduce a diffusion model DDIM [Song *et al.*, 2020] to denoise $\boldsymbol{f}_{x_2}^{S_1}$. By leveraging the generative capabilities of the diffusion model, we achieved fusion results that are more conducive to tracking. We use $\boldsymbol{f}_{x_2}^{S_1}$ as the input and train the model with the features of the two modalities $\boldsymbol{f}_{x_2}^{v}$ and $\boldsymbol{f}_{x_2}^{i}$ as conditions following [Tang *et al.*, 2024], where we use a 1×1 convolution as a projector for feature dimensionality reduction. We also use the first-layer output of MDGF $\boldsymbol{f}_{x_2}^{l_2}$ as another condition. Additionally, distractor object noise $\boldsymbol{f}_{x_1}^{S_1}$ is introduced during the diffusion process to enhance the robustness of the model against similar distractor.

Diffusion process. Due to the motion in video frames, the fusion result of the neighboring frame $\boldsymbol{f}_{x_1}^{S_1}$ is offset at the object location compared to the fusion result of the current frame $\boldsymbol{f}_{x_2}^{S_1}$. This natural offset can serve as a similarity distractor for the current frame, enhancing the robustness of the model. Based on this observation, in addition to the original Gaussian noise, we incorporate the first-stage fusion result $\boldsymbol{f}_{x_1}^{S_1}$ as part of the noise in the diffusion process.

In the forward diffusion process, we use the MDGF fusion output $\boldsymbol{f}_{x_2}^{S_1}$ as x_0. Then x_0 undergoes diffusion through the random Gaussian noise $\bar{z}_t$ and the distractor noise from the first-stage fusion result of the neighboring frame $\boldsymbol{f}_{x}^{S_1}$ as d_t:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + (1-\beta)\sqrt{1-\bar{\alpha}_t}\bar{z}_t + \beta d_t. \tag{20}$$

Denoising process. Some recent works [Cao *et al.*, 2025] utilize additional information as conditions for generative models to guide the denoising process of the model. We employ a UNet [Ronneberger *et al.*, 2015] network to perform the denoising process in the model following [Tang *et al.*, 2024]. We design a simple module named the Condition Fuse (ConFuse) model to serve as supplementary information for the intermediate layers of the UNet network. We use the first-layer graph convolution results $\boldsymbol{f}_{x}^{l_1}$ generated in MDGF as middle conditions in the denoising process:

$$\boldsymbol{f}_{x^{vi}} = ConFuse(\boldsymbol{f}_{x}^{l_1}, D) \tag{21}$$

$$= Conv_1(concat(Conv_2(\boldsymbol{f}_{x}^{l_1}), D)) + D. \tag{22}$$

The features are downsampled using $Conv_2(\cdot)$ to match the size of the intermediate layers of the UNet network named D, then concatenated with D along the channel dimension. Subsequently, $Conv_1(\cdot)$ is used to reduce the dimensionality of the features, and finally, they are added to D to prevent information loss.

We calculate the L_2 loss between noise and the output of the denoising process:

$$L_{DM} = L_2(noise, output). \tag{23}$$

4 Experiments

Datasets. GTOT dataset [Li *et al.*, 2016] comprises 50 pairs of visible and thermal infrared video sequences and corresponding ground truth annotations. These sequences are designed to examine diversity and biases across seven specific challenges. Attribute-based annotations allow for a more detailed evaluation of tracker performance. **RGBT234** [Li *et al.*, 2019] encompassing 234 sequences, with the longest sequence containing up to 8k frames. This dataset is annotated with 12 attributes, making it a robust benchmark for assessing the effectiveness of different trackers. **LasHeR** dataset [Li *et al.*, 2021] featuring 1,224 pairs of visible and thermal infrared videos, with a total of 730k frame pairs. It introduces 19 real-world challenge attributes, providing a more extensive evaluation framework than GTOT and RGBT234. **VTUAV** [Zhang *et al.*, 2022] is a large-scale benchmark specifically designed for visible-thermal UAV tracking, including 500 sequences and 1.7 million high-resolution frame pairs. Its diversity encompasses various categories and scenes, supporting exhaustive evaluations that cover short-term tracking, long-term tracking, and segmentation mask prediction.

Tracker	RGBT234		LasHeR			VTUAV		GTOT	
	PR↑	SR↑	PR↑	NPR↑	SR↑	PR↑	SR↑	PR↑	SR↑
KCF†	46.3	30.5	-	-	-	-	-	-	-
MEEM†	63.6	40.5	-	-	-	-	-	64.8	52.3
UDT†	56.8	42.0	-	-	-	-	-	73.7	61.4
USOT†	50.8	31.8	24.8	20.6	16.0	36.3	26.5	70.6	58.3
UDT-FF	56.1	41.4	28.1	22.8	21.5	51.0	40.1	79.3	65.1
TBSI*	53.0	34.4	29.8	24.5	25.9	24.4	23.4	57.2	49.1
ViPT*	61.7	44.0	38.2	34.1	32.4	47.8	42.4	61.1	54.0
AFter*	57.2	38.7	30.7	25.9	26.6	25.1	22.1	30.7	25.7
S2OTFormer	68.4	47.7	39.8	35.4	29.5	56.7	44.5	**83.1**	**70.2**
GDSTrack	**70.9**	**48.5**	**45.9**	**39.3**	**35.4**	**72.3**	**59.8**	73.9	59.8

Table 1: PR, NPR, and SR evaluation results compared to state-of-the-art self-supervised methods on four benchmarks. Among them, KCF and MEEM are correlation filtering-based methods. Trackers marked with † denote the results obtained by directly combining RGB and infrared features using existing self-supervised RGB tracking methods. For a fair comparison, trackers with * refer to the results trained with the same pseudo-labels as GDSTrack (Ours). Performance of which achieved best are marked in ***bold***.

Metric. Following [Wang *et al.*, 2023; Lu *et al.*, 2024], we evaluate our tracker using three metrics: **Precision rate** (PR), **Normalized precision rate** (NPR), and **Success rate** (SR). PR measures the proportion of frames where the Euclidean distance between the predicted object's center and the ground truth center is below a predefined threshold. Specifically, we use a threshold of 5 pixels for GTOT and 20 pixels for RGBT234 and LasHeR. PR is further normalized as NPR. SR assesses the intersection ratio between the predicted bounding box and the ground truth bounding box, also referred as the overlap ratio. It quantifies the percentage of frames where the overlap exceeds a certain threshold. For all four datasets, we calculate the area under the success rate curve (AUC) to derive the success score.

Implementation Details. Our model is implemented using the PyTorch platform. We use LasHeR [Li *et al.*, 2021] as our training dataset. Following [Zheng *et al.*, 2021], we generate pseudo labels and confidence scores for both the visible and infrared datasets. The label with the higher confidence score is selected as the final pseudo label. We use ViT-B/16 [Dosovitskiy *et al.*, 2020] as the feature encoder for interaction between the template frame and the infrared frame. The en-

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

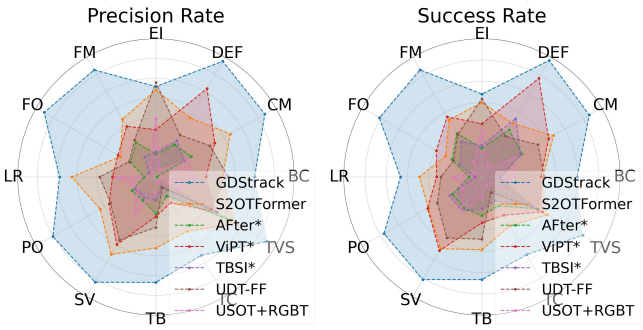

Figure 4: Attribute-based evaluation on VTUAV dataset compared against five self-supervised RGBT trackers. The radial axis for PR ranges from 0.2 to 0.8, while SR is scaled from 0.1 to 0.7.

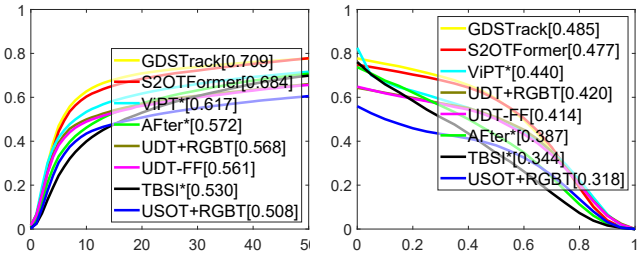

Figure 5: Precision Rate and Success Rate on RGBT234 dataset compared against other self-supervised RRateGBT trackers.

coder is initialized with the pre-trained parameters proposed in [Zhao *et al.*, 2023a]. The template size and search region size are set to 128×128 and 256×256. Training is carried out using the AdamW optimizer [Loshchilov, 2017]. The batch size is 16, and the learning rate is set to 0.00003. We first train the MDGF module and its corresponding tracking head. Starting from the 10th epoch, the parameters of the ViT encoder are unfrozen. From the 22nd epoch, the encoder, MDGF, and tracking head parameters are frozen, and only the parameters related to the diffusion model and the tracking head (diffuse-head) are trained. The training is guided by the MDGF tracking results and pseudo-labels. The training concludes at the 50th epoch. The weight decay coefficient is set to 0.0001, and the momentum value is set to 0.9.

4.1 Main Results

We present the main experimental results here, and additional results can be found in the appendix.

Comparisons with the State-of-the-Art. As shown in Tab. 1, we compared our method with other state-of-the-art self-supervised RGB-T tracking methods, including KCF [Henriques *et al.*, 2014]+ RGBT, MEEM [Zhang *et al.*, 2014] + RGBT, TBSI* [Hui *et al.*, 2023], AFter* [Lu *et al.*, 2024], ViPT* [Zhu *et al.*, 2023], UDT [Wang *et al.*, 2019] + RGBT, UDT-FF [Li *et al.*, 2023], USOT [Zheng *et al.*, 2021] + RGBT and S2OTFormer [Li *et al.*, 2024]. We re-trained AFter [Lu *et al.*, 2024], ViPT [Zhu *et al.*, 2023] and TBSI [Hui *et al.*, 2023] on the LasHeR dataset using the same RGB-T pseudo-labels as our method, named as AFter*, ViPT* and TBSI* for comparison. It can be observed that our model achieves state-of-the-art performance on the RGBT234, LasHeR, and VTUAV datasets.

KCF†	MEEM†	UDT†	USOT†	UDT-FF	TBSI*	ViPT*	AFter*	S2OTFormer	GDSTrack
124.1	4.9	74.3	58.4	71.9	50.4	92.3	27.3	38.2	37.6

Table 2: Speed comparison (fps) between GDSTrack and other self-supervised RGB-T tracking methods.

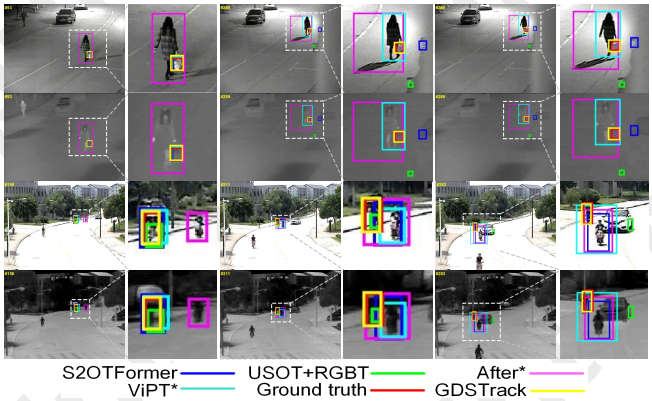

Figure 6: Visualization results of GDSTrack alongside other self-supervised state-of-the-art methods on the RGBT234 dataset. The first two rows are the RGB and infrared frames of sequence *baginhand*, and the last two rows are the RGB and infrared frames of sequence *cycle2*. For complete sequence visualization details, please refer to the appendix.

It can be seen that our method exceeds the self-supervised method S2OTFormer (68.4% / 47.7%) by 2.5% on PR and 0.8% on SR on the RGBT234 dataset. It also surpasses the S2OTFormer (39.8% / 35.4% / 29.5%) by 6.1% on PR, 3.9% on NPR, and 5.9% on SR on the LasHeR dataset. It also surpasses the S2OTFormer (56.7% / 44.5%) by 15.6% on the PR score and 15.3% on the SR score on the VTUAV dataset. We hypothesize that the suboptimal performance on the GTOT dataset is due to the smaller number and shorter sequence lengths in this dataset. We compared our model with other state-of-the-art self-supervised RGB-T trackers on the RGBT234 dataset. As shown in Fig. 5, our model achieved the best performance on both the PR plots and the SR plots.

Attribute Analysis. The performance of our method GDSTrack, along with other sota self-supervised methods on different attribute challenges of VTUAV dataset is shown in Fig. 4. There is no OV attribute in the Short-term Evaluation of the VTUAV dataset, so we only draw twelve attributes. As observed, our method achieves the best performance on most attributes of the VTUAV dataset with only 2.1% of PR lower on extreme illumination (EI) attribute than UDT-FF.

Speed Analysis. As shown in Tab. 2, GDSTrack achieves better performance than S2OTFormer while maintaining comparable speed.

Visualization Analysis. We conducted visualization on the RGBT234 dataset. As shown in Fig. 6, for better visualization, we magnify the objects within the dashed boxes. Through comparison, we can observe that, thanks to our

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

MDGF	TGID	RGBT234	LasHeR	VTUAV	GTOT
		69.1/47.1	43.8/37.6/33.7	69.0/56.2	68.6/56.0
✓		70.9/48.3	45.9/39.1/35.2	70.3/58.0	73.5/59.5
✓	✓	**70.9/48.5**	**45.9/39.3/35.4**	**72.3/59.8**	**73.9/59.8**

Table 3: PR, NPR, and SR of the model with or without MDGF or TGID, evaluated on four datasets.

Adj. Matrix	RGBT234	LasHeR	VTUAV	GTOT
Identity	69.8/**48.7**	42.3/37.5/33.1	68.7/57.3	70.8/58.3
QKV	67.6/46.2	42.5/36.5/32.7	70.0/57.6	71.9/57.5
Cosine	69.2/46.4	42.8/36.4/33.0	68.7/56.5	70.6/58.6
AMG (Ours)	**70.9**/48.3	**45.9/39.1/35.2**	**70.3/58.0**	**73.5/59.5**

Table 4: PR, NPR, and SR of the model with different adjacency matrix generation methods, evaluated on four datasets.

MDGF module, our method can locate the object area more accurately. For example, in sequence *baginhand*, as the object is tracked, the ViPT* method gradually focuses on the pedestrian, while the After* method simultaneously tracks the shadow region. In contrast, our method precisely tracks the pedestrian's bag in hand. Furthermore, benefiting from our TGID module, our method demonstrates stronger resistance to similar object noise. For instance, in sequence *cycle2*, when tracking the pedestrian, other methods gradually get distracted by similar objects and begin tracking a different pedestrian with similar features, whereas our method consistently tracks the same pedestrian.

4.2 Ablation Study

Component Analysis. We conducted ablation studies on four datasets to evaluate the effectiveness of the MDGF and TGID modules, with the results presented in Tab. 3. We constructed our baseline model by directly incorporating RGB and infrared features into the RGB tracker, using a model trained with pseudo-labels. The baseline results on the four datasets are presented in the first row of Tab. 3.

The ablation of MDGF. The tracker with MDGF demonstrates a notable improvement over the baseline model, with PR scores of 1.8%, 1.2% on the RGBT234 dataset, 2.1%, 1.5%, and 1.5% on the LasHeR dataset, 1.3% and 1.8% on the VTUAV dataset, and 4.9% and 3.5% on the GTOT dataset. As shown in the first and second rows of this table, MDGF can dynamically adjust the adjacency matrix, thereby better integrating the advantages of modalities for subsequent tracking.

The ablation of TGID. The tracker with both MDGF and TGID achieves further improvement over the MDGF-only model, with SR of 0.2% on the RGBT234 dataset, NPR and SR of 0.2%, 0.2% on the LasHeR dataset, PR and SR of 2% and 0.8% on the VTUAV dataset, and 0.4% and 0.3% on the GTOT dataset. From the second and third rows of this table, we can see that the TGID module further enhances the model's performance by improving robustness to interfering noise, building on the MDGF module.

The ablation of AMG. To evaluate the effectiveness of our proposed adjacency matrix generation method, we compared it with other generation methods, as shown in Tab. 4. We employed the identity matrix, self-attention, and cosine similarity methods to generate adjacency matrices, which serve as comparative methods for our proposed AMG module. We can observe that training with the adjacency matrix generated by the AMG method yields the best performance on the LasHeR, VTUAV, and GTOT datasets. On the RGBT234 dataset, the AMG method showed a little decrease of 0.4% in SR, compared to the identity matrix method. The excellent performance of the AMG module proves that, compared to other adjacency matrix generation methods, the AMG module can dynamically capture the inter-modal similarity, direct the model to focus on the object-related regions through graph attention and obtain fusion results that are favorable for tracking.

k	5	10	15	20	128	256
GTOT	69.9/57.0	70.6/56.6	69.5/54.2	69.5/54.7	71.9/58.1	**73.5/59.5**
RGBT234	68.9/47.7	67.8/46.2	69.0/46.3	66.8/45.2	69.6/47.0	**70.9/48.3**

Table 5: PR and SR of the model with different k value evaluated on GTOT and RGBT234 datasets.

Naive DM	condition	distractor	RGBT234	LasHeR	VTUAV	GTOT
✓			68.5/47.5	44.0/37.9/34.0	70.8/58.7	69.0/56.5
✓	✓		68.9/47.6	43.8/37.4/33.7	71.3/58.9	69.8/56.6
✓	✓	✓	**69.6/ 48.3**	**44.7/38.3/34.5**	**72.3/59.8**	**70.0/57.5**

Table 6: PR, NPR, and SR of the model with or without middle conditions or distractor noise, evaluated on four datasets. To save computational time, we conducted this ablation experiment under the condition that K is set to 5 in Tab. 5.

The ablation of TopK. Tab. 5 shows the performance of different values of K in the AMG module on two datasets. We observe that the model's performance is positively correlated with the value of TopK. When K is set to 256, the model performs the best on GTOT and RGBT234 datasets.

The ablation of Graph-informed condition. As shown in Tab. 6, compared to [Song *et al.*, 2020], using the first layer of graph convolution from the MDGF module as a condition for the intermediate layers of the UNet in the diffusion model resulted in a clear improvement on three datasets.

The ablation of distractor noise. As shown in Tab. 6, adding similar distractor noise resulted in increases of 0.7% in PR and SR on the RGBT234 dataset, 1% and 0.9% on the VTUAV dataset, and 0.2% and 0.9% on the GTOT dataset.

5 Conclusion

In conclusion, we have designed a Dynamic Graph Diffusion framework to address the issues in self-supervised RGB-T tracking tasks. Specifically, to tackle the problem of incorrect fusion of non-object regions, the MDGF module has been proposed to dynamically adjust the adjacency matrix based on the similarity between modalities, guiding graph attention to focus on the fusion of coherent object regions. To address the noise interference caused by similar objects, the TGID module has incorporated the MDGF fusion results from neighboring frames as noise, training the model to improve its robustness against interference from similar objects. Experiments on four datasets have demonstrated that GDSTrack achieves state-of-the-art results in the field of self-supervised RGB-T tracking, validating the effectiveness of our model.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62172417, 62272461, 62276266 and 62472424.

References

[Cao *et al.*, 2025] Congqi Cao, Hanwen Zhang, Yue Lu, Peng Wang, and Yanning Zhang. Scene-dependent prediction in latent space for video anomaly detection and anticipation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):224–239, 2025.

[Chen *et al.*, 2024] Xin Chen, Ben Kang, Jiawen Zhu, Dong Wang, Houwen Peng, and Huchuan Lu. Unified sequence-to-sequence learning for single- and multi-modal visual object tracking, 2024.

[Cui *et al.*, 2022] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13608–13618, 2022.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Henriques *et al.*, 2014] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583–596, 2014.

[Hu *et al.*, 2024] Xiantao Hu, Ying Tai, Xu Zhao, Chen Zhao, Zhenyu Zhang, Jun Li, Bineng Zhong, and Jian Yang. Exploiting multimodal spatial-temporal patterns for video object tracking. *arXiv preprint arXiv:2412.15691*, 2024.

[Hui *et al.*, 2023] Tianrui Hui, Zizheng Xun, Fengguang Peng, Junshi Huang, Xiaoming Wei, Xiaolin Wei, Jiao Dai, Jizhong Han, and Si Liu. Bridging search region interaction with template for rgb-t tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13630–13639, 2023.

[Lai *et al.*, 2024] Simiao Lai, Chang Liu, Jiawen Zhu, Ben Kang, Yang Liu, Dong Wang, and Huchuan Lu. Mambavt: Spatio-temporal contextual modeling for robust rgb-t tracking. *arXiv preprint arXiv:2408.07889*, 2024.

[Li *et al.*, 2016] Chenglong Li, Hui Cheng, Shiyi Hu, Xiaobai Liu, Jin Tang, and Liang Lin. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing*, 25(12):5743–5756, 2016.

[Li *et al.*, 2019] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: Benchmark and baseline. *Pattern Recognition*, 96:106977, 2019.

[Li *et al.*, 2021] Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, Jin Tang, and Dengdi Sun. Lasher: A large-scale high-diversity benchmark for rgbt tracking. *IEEE Transactions on Image Processing*, 31:392–404, 2021.

[Li *et al.*, 2023] Shenglan Li, Rui Yao, Yong Zhou, Hancheng Zhu, Bing Liu, Jiaqi Zhao, and Zhiwen Shao. Unsupervised rgb-t object tracking with attentional multi-modal feature fusion. *Multimedia Tools and Applications*, pages 1–19, 2023.

[Li *et al.*, 2024] Shenglan Li, Rui Yao, Yong Zhou, Hancheng Zhu, Jiaqi Zhao, Zhiwen Shao, and Abdulmotaleb El Saddik. Motion-aware self-supervised rgbt tracking with multi-modality hierarchical transformers. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

[Loshchilov, 2017] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Lu *et al.*, 2024] Andong Lu, Wanyu Wang, Chenglong Li, Jin Tang, and Bin Luo. After: Attention-based fusion router for rgbt tracking. *arXiv preprint arXiv:2405.02717*, 2024.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015.*, pages 234–241. Springer, 2015.

[Shen *et al.*, 2022] Qiuhong Shen, Lei Qiao, Jinyang Guo, Peixia Li, Xin Li, Bo Li, Weitao Feng, Weihao Gan, Wei Wu, and Wanli Ouyang. Unsupervised learning of accurate siamese tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8101–8110, 2022.

[Sio *et al.*, 2020] Chon Hou Sio, Yu-Jen Ma, Hong-Han Shuai, Jun-Cheng Chen, and Wen-Huang Cheng. S2siamfc: Self-supervised fully convolutional siamese network for visual tracking. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1948–1957, 2020.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[Tang *et al.*, 2023] Wei Tang, Fazhi He, Yu Liu, Yansong Duan, and Tongzhen Si. Datfuse: Infrared and visible image fusion via dual attention transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7):3159–3172, 2023.

[Tang *et al.*, 2024] Zhangyong Tang, Tianyang Xu, Xiaojun Wu, Xue-Feng Zhu, and Josef Kittler. Generative-based fusion mechanism for multi-modal tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5189–5197, 2024.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

need. *Advances in neural information processing systems*, 30, 2017.

[Velickovic *et al.*, 2017] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

[Wang *et al.*, 2019] Ning Wang, Yibing Song, Chao Ma, Wengang Zhou, Wei Liu, and Houqiang Li. Unsupervised deep tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1308–1317, 2019.

[Wang *et al.*, 2023] Xiao Wang, Xiujun Shu, Shiliang Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Mfgnet: Dynamic modality-aware filter generation for rgb-t tracking. *IEEE Transactions on Multimedia*, 25:4335–4348, 2023.

[Xi *et al.*, 2023] Yuling Xi, Hao Chen, Ning Wang, Peng Wang, Yanning Zhang, Chunhua Shen, and Yifan Liu. A dynamic feature interaction framework for multi-task visual perception. *Int. J. Comput. Vision*, 131(11):2977–2993, July 2023.

[Xu *et al.*, 2021] Qin Xu, Yiming Mei, Jinpei Liu, and Chenglong Li. Multimodal cross-layer bilinear pooling for rgbt tracking. *IEEE Transactions on Multimedia*, 24:567–580, 2021.

[Zhang *et al.*, 2014] Jianming Zhang, Shugao Ma, and Stan Sclaroff. Meem: robust tracking via multiple experts using entropy minimization. In *European Conference on Computer Vision*, pages 188–203. Springer, 2014.

[Zhang *et al.*, 2019] Lichao Zhang, Martin Danelljan, Abel Gonzalez-Garcia, Joost Van De Weijer, and Fahad Shahbaz Khan. Multi-modal fusion for end-to-end rgb-t tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[Zhang *et al.*, 2022] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8886–8895, 2022.

[Zhang *et al.*, 2024] Lei Zhang, Jiangtao Nie, Wei Wei, and Yanning Zhang. Unsupervised test-time adaptation learning for effective hyperspectral image super-resolution with unknown degeneration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):5008–5025, 2024.

[Zhang *et al.*, 2025] Zhengbo Zhang, Li Xu, Duo Peng, Hossein Rahmani, and Jun Liu. Diff-tracker: text-to-image diffusion models are unsupervised trackers. In *European Conference on Computer Vision*, pages 319–337. Springer, 2025.

[Zhao *et al.*, 2023a] Haojie Zhao, Dong Wang, and Huchuan Lu. Representation learning for visual object tracking by masked appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18696–18705, June 2023.

[Zhao *et al.*, 2023b] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5906–5916, 2023.

[Zheng *et al.*, 2021] Jilai Zheng, Chao Ma, Houwen Peng, and Xiaokang Yang. Learning to track objects from unlabeled videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13546–13555, 2021.

[Zhu *et al.*, 2023] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9516–9526, 2023.