SiamTDR: Time-Efficient RGBT Tracking via Disentangled Representations

Guorui Wang, Qian Jiang, Xin Jin, *Member, IEEE*, Yu Lin, Yuanyu Wang, and Wei Zhou, *Member, IEEE*

***Abstract*—The growing demand for vision tasks utilizing RGB-T (Red-Green-Blue-Thermal) imagery is attributed to the advantageous synergistic effect of combining RGB images with thermal (Tir) image information. Due to their exceptional real-time inference efficacy, siamese networks have garnered considerable attention in RGB-T object tracking as a leading solution. However, current RGB-T Siamese trackers still need to catch up with online training RGB-T trackers regarding accuracy and robustness due to ineffective utilization of valid information from both modes. To this end, this work proposes SiamTDR, a high-speed Siamese network-based RGB-T tracker with a disentangled representation and deconstructed features. Firstly, we introduce a single-modal feature extraction network into the Siamese network to capture cross-level information within unimodal features extracted from RGB or Tir images. Next, we employ a disentangled representation multi-modal feature fusion module (DP-MF) to extract cross-modal information between RGB and thermal features, thereby improving the information utilization of both modalities. Finally, a dual branch fusion module (DBF) significantly enhances the robustness of our tracker in the final bounding box selection stage. Besides, we also employ data augmentation techniques such as central random offset. Extensive experiments conducted on two RGB-T tracking benchmark datasets demonstrate the superior performance of our method, which achieves a tracking speed of over 127 frames per second (FPS) on the GTOT dataset.**

***Index Terms*—Disentangled representations, information fusion, RGB-T tracking, real-time tracking, Siamese network.**

Manuscript received 29 June 2023; accepted 17 August 2023. Date of publication 22 August 2023; date of current version 25 September 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 62101481, 62002313, 62261060, and 62162067, in part by the Basic Research Project of Yunnan Province under Grants 202301AW070007, 202201AU070033, 202201AT070112, 202301AU070210, 202001BB050076, and 202005AC160007, in part by the Major Scientific and Technological Project of Yunnan Province under Grant 202202AD080002, and in part by the Research and Application of Object Detection based on Artificial Intelligence. *(Corresponding authors: Xin Jin; Wei Zhou.)*

Guorui Wang, Qian Jiang, Xin Jin, and Wei Zhou are with the Engineering Research Center of Cyberspace, Yunnan University, Kunming 650000, China, and also with the School of Software, Yunnan University, Kunming 650000, China (e-mail: guoruiwang77@gmail.com; jiangqian@ynu.edu.cn; xinjin@ynu.edu.cn; zwei@ynu.edu.cn).

Yu Lin and Yuanyu Wang are with the Kunming Institute of Physics, Kunming 650223, China (e-mail: lwlinyu@163.com; wxyjin232425@163.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/TICPS.2023.3307340, provided by the authors.

Digital Object Identifier 10.1109/TICPS.2023.3307340

I. Introduction

With the deep integration and development of informatization and industrialization, cyber-physical systems integrating technologies such as computing, communication, and control have emerged [1], [2], [3]. The cyber-physical systems have been widely applied in critical infrastructure, such as the industrial internet, smart grids, and intelligent transportation systems [4]. Real-time monitoring and perceiving the external environment are fundamental for achieving system autonomy and collaboration in cyber-physical systems [5], [6]. In this context, object tracking is crucial to enabling real-time monitoring and perception in the system. Real-time information regarding position, velocity, and motion trajectories is obtained by using sensor data and other information to track the objects in the system, enhancing the system's perceptual capabilities and decision-making accuracy [7].

Currently, significant breakthroughs have been made in single-modal object tracking techniques [8], [9], [10], [11], [12], [13] based on RGB or TIR imaging. However, these algorithms still face challenges in complex scenarios or extreme conditions. RGB images provide rich color information and better differentiate and recognize different objects in austere environments. However, they struggle to provide sufficient effective features for tracking in complex environments such as low light, occlusion, and rainy conditions, resulting in unsatisfactory tracking precision. Consequently, RGB object tracking techniques' accuracy fails to meet the requirements in cyber-physical systems. On the other hand, TIR images are imaged by thermal radiation, which is insensitive to light and sensitive to temperature. Moreover, they exhibit excellent penetration capabilities in scenarios involving smoke obstruction. Nonetheless, TIR images are susceptible to thermal crosstalk interference and lack detailed texture and color information. Solely relying on TIR images for tracking in cyber-physical systems restricts the reliability of the algorithms. Considering the complementary nature of RGB and TIR image information, reasonably incorporating data from both modalities during the tracking process can achieve more robust tracking than using the single modality alone. RGBT object tracking algorithms carry significant research implications and practical value. Since the introduction of the Visible Light-Thermal Infrared multimodal dataset (OSU Color Thermal) by Davis et al. [14] in 2007, RGBT (RGB-Thermal) object tracking algorithms have emerged as a new research topic and have gained increasing attention from researchers.

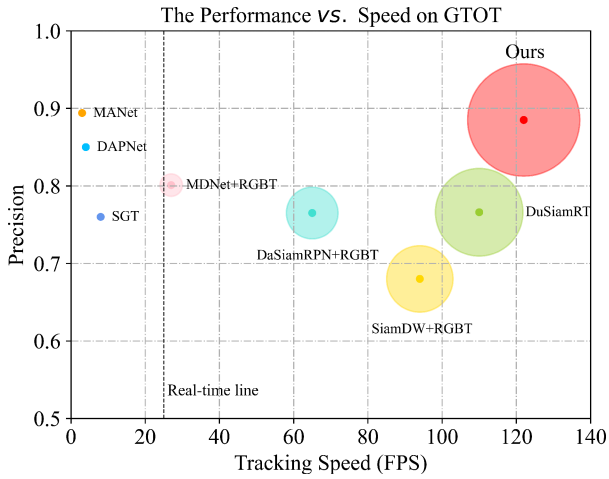

Fig. 1. Comparison of the proposed Siam with other state-of-the-art RGB-T trackers. the SiamTDR operates at twenty times the speed of DAPNet, with comparable performance.

Fig. 2. Illustrative of complementary information for RGB images and Tir images. (a) and (b) are two example frames of the RGB modality and the Tir modality, respectively. As shown in the area marked by the red box, there is remarkable image complementary information between the two modalities, which can improve the accuracy of object tracking.

In the field of RGBT object tracking, early research [15], [16], [17] relied on manual feature extraction. However, these methods fail to combine the two modalities effectively, resulting in inadequate tracking accuracy. Currently, deep learning methods have demonstrated good performance in RGBT objecting tracking, primarily based on two approaches: multi-domain networks and Siamese networks. MDNet [18] introduces the concept of multi-domain networks, which classifies regions randomly selected from each frame and uses the region with the highest confidence score as the object for tracking. However, these multi-domain network-based algorithms' significantly slower processing speed is much lower than in real-time. These trackers are ill-suited for performing real-time tasks such as cyber-physical systems and autonomous driving. The concept of Siamese networks is introduced to object tracking by SiamFC [19]. Siamese network algorithms continuously perform convolutional operations between the template frame and the search region frame to determine the object's location. Although Siamese networks [20], [21] demonstrate favorable real-time performance due to their fully convolutional structure, their accuracy falls short compared to RGBT trackers based on multi-domain networks [22], [23], as shown in Fig. 1.

To address the issue of insufficient accuracy in Siamese tracking, RGBT tracking algorithms extract features from modal images. Among these, simple multi-modal fusion algorithms tend to decompose the source images into different parts to extract features from each modality, often disregarding the varying contributions of each modality and region to the detection process, resulting in overfitting to noisy regions in the infrared image. In real-world scenarios, infrared imaging limitations, such as excessive noise, low image contrast, low signal-to-noise ratio, blurry edges, visual distortion, and limited grayscale range, result in different contributions of RGB and TIR images. Fig. 2 illustrates several images from a public RGB-T234 dataset [24], wherein certain target objects are lost or blurred due to imaging methods and noise interference, particularly near the background and object boundaries. To tackle these challenges, an attention-based multi-modal information fusion mechanism for RGB-T by using feature disentanglement is proposed in this article. This mechanism aims to mitigate the adverse effects of TIR image noise and maximize the complementarity of multi-modal features. Bengio et al. [25] proposed that disentangled representation learning is an important research direction in the next phase of deep learning, stating that the entanglement of multiple generating factors generates data. Disentangled representation learning aims to extract interpretable attributes from various data variations, generating meaningful representations that separate valuable interpretable attributes for downstream tasks from other interfering factors. This approach ultimately enhances the accuracy and robustness of the model.

To address the issues of algorithmic timeliness and accuracy, we have designed an RGB-T tracking framework based on the Siamese network, named SiamTDR. This framework achieves high performance while maintaining real-time operation speed. The RGB-T object tracking algorithm based on the Siamese network significantly enhances situational awareness, thereby optimizing cyber-physical systems' security. As shown in Fig. 1, our tracker achieves comparable performance to some state-of-the-art RGB-T trackers while being more efficient in tracking speed. We employ two identically structured feature extractors to extract features from different modalities of the template frame and the detection frame. The attention fusion module prevents introducing infrared noise during fusion and fully exploits multi-modal features. Subsequently, both single-modal and fused features are provided to the prediction module to

enhance the network's detection capability and noise resistance. The main contributions of this research can be summarized as follows:

1) A unified RGB-T tracking framework based on the Siamese network is proposed to balance the weights of different modal features and achieve high tracking performance.
2) A concise and small feature extraction method is designed to disentangle modal representations during extraction, combined with a dual neural network to construct the backbone network of the multi-modal object tracking algorithm.
3) A fusion module based on dynamic convolution is designed by adaptively focusing on the critical regions in both spatial and channel dimensions of different modalities, with efficient use of RGB and Tir image synergy features.
4) A fusion prediction module is designed and applied to track objects using RGB features and fused feature inspection. The tracking capability of this network is further enhanced.

II. Related Work

A. RGB Tracking Methods

In recent years, advancements in benchmark datasets and novel techniques have led to significant advancements in object tracking. Deep feature representation-based trackers [13], [26], [27], in particular, have achieved state-of-the-art performance in key tracking benchmarks, thanks to the success of Convolutional Neural Networks (CNNs) in various computer vision applications. These modern tracking algorithms can be broadly categorized into discriminative and generative trackers. Discriminative trackers, which require online model training, train a classifier to differentiate the target from the background. For instance, Object-aware Anchor-free Tracking (Ocean), a unique CNN architecture, was proposed in [28] to learn the tracked targets inside the predicted bounding box through the feature sampling locations module for tracking purposes.

Additionally, specialized trackers such as ATOM [29], and DiMP [30] have been developed to achieve even higher performance standards. Although these discriminative trackers have relatively lower speeds, they offer exceptional tracking performance. On the other hand, generative trackers [21], [21], [31] estimate the joint probability densities between features and search candidates to determine the best match for the target. Among generative trackers, Siamese network-based trackers [21], [31] have received significant attention, surpassing various benchmarks with their real-time performance.

B. Siamese Network Based RGB Trackers

The article "SiamFC" [19] pioneered using fully convolutional Siamese networks for object tracking. It framed object tracking as a similarity learning problem. Subsequently, "SiamRPN" [31] was developed by integrating region proposal networks [32] into the SiamFC framework, resulting in more accurate target bounding box predictions. Inspired by the success of SiamFC and SiamRPN, several subsequent studies [33], [34], [35] have further improved these models. For example, "Zhu et al." [33] introduced distractor-aware training into the SiamRPN framework. "C-RPN" [34] proposed a multi-stage tracking technique to improve localization accuracy. To improve the performance of Siamese network-based trackers, deeper networks such as ResNet [36], ResNeXt [37], and MobileNet [38] were incorporated as the backbone. The "SiamRPN++" [39] model used these modern deep neural networks and introduced a novel training strategy to overcome limitations by randomly changing the placement of training items within the search area. Another approach, "SiamDW" [40], used a residual network for visual tracking with customizable, receptive field size and network step, resulting in improved tracking accuracy. More recently, various Siamese trackers [21], [31] have adopted the regression of the distance between the projected target's center and the bounding box's borders, drawing inspiration from anchor-free detectors such as [41]. Additionally, some researchers [28] have utilized adversarial techniques from CNNs to increase the robustness of deep learning-based trackers.

C. RGB-T Tracking Methods

Various RGB-T tracking algorithms have recently been developed that leverage RGB and thermal information to improve tracking performance. Initially, these algorithms relied on hand-engineered features [15], [16], [17], [22], [23], [42]. With the advancement of deep learning, more RGB-T trackers have been introduced that are based on learned attributes [23], [24], [42]. The RGB trackers serve as the foundation for these RGB-T trackers. For instance, Li et al. [16] proposed a network that aggregates features from all layers and modalities before removing noise and redundant information. Li et al. [23] presented a multi-adapter architecture for learning target representations shared across modalities, specific to each modality, and instance-aware. Zhang et al. [43] compared different fusion techniques using DiMP [30] as the baseline tracker and demonstrated that their proposed fusion tracker outperforms the baseline and achieves state-of-the-art results in unimodal tracking. However, these discriminative RGB-T trackers have high computational complexity, as demonstrated by the low tracking speed of the MANET [23] tracker, which is only around two frames per second. To address this issue, some researchers have explored using Siamese networks in RGB-T tracking, as they have proven effective in RGB tracking. For example, SiamFT [35] used two Siamese networks to extract features from the RGB and thermal inputs and manually set the modality weights. DuSiamRT [20] improved upon this approach by using a joint modal channel attention module and improving the regional proposal subnetwork.

D. Disentangled Representation

The primary objective of disentangled representation learning is to effectively capture the fundamental factors responsible for variations in data by decomposing it into distinct and independent components [44]. These representations hold immense

value across diverse domains, such as machine learning, computer vision, and natural language processing, as they enhance comprehension, control, and data manipulation. Disentangled representation learning is currently being applied in numerous domains, including image fusion, image synthesis, and pose estimation. For instance, Xu et al. [45] proposed a visible and infrared image fusion network based on disentangled representation learning in the field of image fusion. This network effectively disentangles the sources of information from visible and infrared images, which can mitigate the issue of inappropriate extraction of specific information. In the context of image synthesis, Li et al. [46] introduced MixNMatch, it is a conditional generative model that is capable of encoding object pose shape and texture information from input images. These factors are then combined to compose the desired image. Furthermore, Xia et al. [47] presented a multi-domain adaptive learning model incorporating information-theoretic stimulus constraints in pose estimation. This algorithm improves the robustness of labeling variable samples by enabling the neural network to learn disentangled representations of multimodal sensor data.

E. Attention Mechanis

Several works [48], [49] developed attention mechanisms to evaluate the relative significance of various areas or modalities to circumvent these issues. The attention mechanism has been extensively employed in various applications [50], [51] to assist networks in extracting robust and distinguishable characteristics. Google DeepMind utilized the attention mechanism for the picture classification job and introduced a fresh recurrent neural network model in 2014. The model can extract information from an image or video by picking a series of areas or locations adaptively and analyzing just those regions at high resolution [50]. Hu et al. (2018). [52] developed the Squeeze-and-Excitation (SE) block to concentrate on the channel connection, which learns the reliance of each channel and adaptively recalibrates channel-wise feature responses to enhance the representation capability. The SE block solely analyzes the channel contribution of feature maps, ignoring the object's spatial position in pictures. The item's spatial placement plays a crucial role in object detection. Woo et al. (2018) [53] suggested the Convolutional Block Attention Module for this purpose (CBAM). CBAM successively infers attention maps in two dimensions (channel and spatial) and then performs adaptive feature refinement by multiplying the attention maps by the input feature maps. Chen et al. (2020b) [54] investigated attention processes for convolution kernels, in contrast to the studies of SE block and CBAM that address attention mechanisms for feature maps. They introduced a unique multi-dimensional attention mechanism with a concurrent technique for learning complementary attentions for convolutional kernels. The name of the block is Dynamic Convolution (DConv). Although these Siamese network-based RGB-T trackers can approach real-time performance, their accuracy is still lower than other state-of-the-art trackers, partly due to the lack of a module to enhance the utilization of information from both modalities. Thus, RGB-T tracking using Siamese networks is still room for improvement. In this article,

TABLE I
NUMBER OF SIAMTDR BACKBONE PARAMETERS COMPARED TO OTHER MAINSTREAM BACKBONE NETWORKS

Vgg11	ViT	Inception_v3	Resnet50	Googlenet	**Ours**
132.8M	86.5M	27.1M	25.5M	13M	**7.2M**

RGB-T multi-modal information is fused based on an attention mechanism to limit the detrimental impact of Tir image noise and maximize the complementary multi-modal characteristics.

III. METHOD

This work aims to improve object tracking accuracy by ensuring real-time and effective usage of multi-modal data (RGB and Tir). As a result, as illustrated in Fig. 3, a high-speed multi-stage noise-resistant feature fusion network is proposed. First, two feature extractors with the same structure (Extractor-RGB and Extractor-T) extract features from color and infrared images, respectively. The multi-modal features are then decoupled and separated, resulting in increased feature diversity and sequential transmission to the multi-modal feature fusion module for information fusion. This module allows the network to focus more on the significant regions and channels of the input features, allowing it to learn not only the standard and complementary features between the different modalities but also to eliminate the adverse effects of infrared image noise. The modal features are intercorrelated with the search region features and sent to the fusion prediction module to determine the target's location within the search region. Finally, the target's position is mapped to the original frame. Fig. 3 depicts the features extracted from the infrared and visible images, represented by the orange and blue lines. The orange line corresponds to the features extracted from the infrared image after undergoing processing by the backbone network, revealing the inherent characteristics of infrared radiation and transforming them into a more profound information representation. Conversely, the blue line represents the features extracted from the visible image after being processed by the backbone network, reflecting the scene's detailed texture and color information.

A. Siamese Single-Modal Feature Extraction Network

We developed a low-complexity two-stage feature extraction network in our tracking model proposal. The primary objective of this network is to extract discriminative features that hold high relevance for target tracking. Our algorithm achieves a favorable trade-off between accuracy and speed by concentrating on essential features and eliminating redundant details. Furthermore, our feature extraction network's parameter count is lower than most existing networks, resulting in faster processing times, as indicated in Table I. In the Siamese single-modal feature extraction network, the first stage employs a Siamese network to extract two distinct unimodal features. This network consists of two branches that share the same structure and parameters. One branch (the template branch) extracts features from the template image. The other branch (called the detection branch)

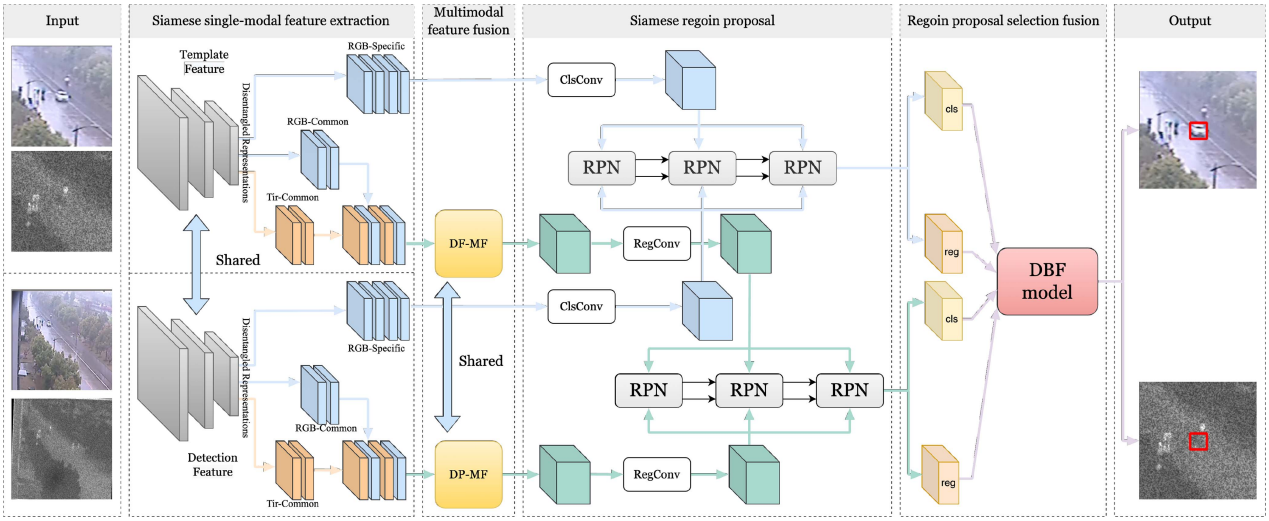

Fig. 3. Overview of the proposed SiamTDR. The overall network consists of five main parts: Siamese network for unimodal feature extraction, DP-MF module for multi-modal feature fusion, SiameseRPNs for region proposal generation, and DBF module for region proposal network.

extracts features from the search image. It ensures high speed, using fewer parameters to extract unimodal features from each input image. In the second stage, the synergetic and specific information of the RGT and Tir features are transformed into mode-common features and mode-specific features; because the feature space of the neural network has some information redundancy, and the coupling between redundant features, standard features, and unique features reduces the efficiency of using practical information and the sensitivity of the model to discriminative features of the target, making the model is more unstable in inter-modal information selection. Hence, this article introduces a network architecture designed to disentangle the aforementioned multimodal features into two distinct representations: mode-common and mode-specific. This approach effectively enhances the diversity and discriminative nature of the extracted features. In the following contents, we will discuss the construction of our template feature extraction module using the template branch in a Siamese network as an example, considering that both branches in a dual-stream Siamese network have the same structure. The RGB template feature extraction branch consists of a feature aggregation network in the first stage and a decoupling network in the second stage.

In the first stage, we improve on AlexNet as the backbone for feature extraction. We chose AlexNet as the baseline feature extraction network due to its high real-time performance, which is essential for target tracking. To make the network more lightweight, we modify it by utilizing only the first three layers of the AlexNet and removing the padding operation. The rationale behind removing all padding is that during feature extraction for tracking tasks. It is crucial to maintain the translational invariance of the target. The padding operation, however, constantly biases the model's attention towards the center of the image, which somewhat compromises the target's translational invariance.

In the second stage, considering that two modalities describing the same environment or target should exhibit common characteristics and can be described by the same or highly similar model to ensure consistent descriptions, we design an inter-modal common feature extraction branch. This approach gives the diversity of the modal feature space and the feature selection process. Consequently, we design RGB mode-specific, RGB mode-common, and Tir mode-common feature extraction branches to decouple the images from the two different modalities. As illustrated in Fig. 2, the Siamese single-modal feature extraction network accepts the visible RGB image ($\mathbf{image}_{RGB}$) and the infrared Tir image ($\mathbf{image}_{TIR}$) as inputs, generating three outputs. The improved backbone network is denoted as $\mathbf{F}_{PE}$, while the three feature extraction branches are labeled as $\mathbf{F}_{RS}$, $\mathbf{F}_{TS}$, and $\mathbf{F}_{MC}$. $\mathbf{F}_{PE}$ independently accepts visible and infrared images as inputs, $\mathbf{F}_{RS}$ and $\mathbf{F}_{MC}$ accept visible features, and $\mathbf{F}_{TS}$ accepts infrared features independently. The logical relationship between the input and output is represented in (1).

$$\begin{aligned} \mathcal{F}_R = F_{PE}\left(\mathbf{image}_{RGB}\right), \mathcal{F}_T = F_{PE}\left(\mathbf{image}_{TIR}\right) \\ \mathcal{F}_{RS} = F_{RS}\left(\mathcal{F}_R\right) \\ \mathcal{F}_{TS} = F_{TS}\left(\mathcal{F}_T\right) \\ \mathcal{F}_{MC} = F_{MC}\left(\mathcal{F}_R\right) \end{aligned} \tag{1}$$

where $\mathcal{F}_{RS}, \mathcal{F}_{TS}, \mathcal{F}_{MC}^{R}$ denote the visible-specific features, infrared-specific features, and common features extracted from visible images, respectively, of the ternary network output. $\mathcal{F}_{RS}$ and $\mathcal{F}_{TS}$ are completely independent to ensure that the modal feature extraction process is differentiated; $\mathcal{F}_{MC}$ accepts both visible and infrared images as input, meaning that the two modal images share the model structure and parameters, using the same model to produce consistent feature representations for both modal images.

B. DP-MF Module for Multi-Modal Feature Extraction and Fusion

The RGB and thermal (Tir) features for template and detection are acquired independently from the feature fusion branch of

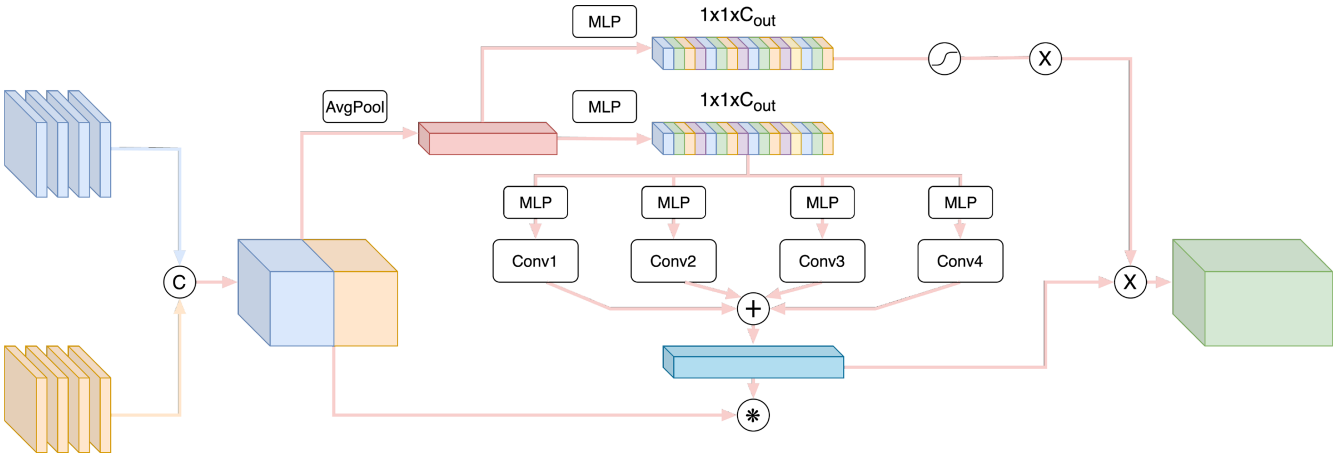

Fig. 4. Illustration of our proposed DP-MF module. First, the weight generation subnetwork takes the features maps from two-stream Siamese networks as input and produces weights that reflect how much additional information should be introduced from one modality data to another modality data. Then the enhanced multi-modal features are obtained by using dynamic convolution. Finally, the fused features are obtained by performing some attention mechanism operations on these enhanced multi-modal features.

the Siamese network. The following objective is to merge these features for target tracking. The fused template features for final tracking are obtained by combining the template branching features from the RGB Siamese network with the template features from the Siamese thermal network, consistent with the current approach of the RGB-T Siamese tracker. Similarly, the RGB detection features and their corresponding thermal detection features are combined to generate the tracking detection features. Successfully fusing these features to capture the complementary information between the RGB and thermal images is a crucial challenge the RGB-T tracking model faces.

RGB-T images possess complementary features that can enhance detection accuracy. However, the conventional methods of fusing multi-modal features, such as element summation and concatenation, do not adequately address the issue. The presence of noise and the varying contributions of different modal features to object detection make it unjust to fuse RGB and Tir images equally. Therefore, assigning appropriate weights when fusing multi-modal information is crucial, considering each modality's characteristics. Nevertheless, most existing fusion strategies fail to account for the feature disparities between the input multi-modal RGB and thermal images during fusion. In a prior study [22], a content-dependency weighting-based fusion strategy was proposed to fuse the multi-modal RGB and thermal features for tracking. This strategy has demonstrated superior performance compared to simple element-wise summation or concatenation-based methods, as it considers the feature reliability of each modality's data. This article introduces the dynamic perception of multi-modal features in the fusion module to address this limitation. The dynamic perception module adaptively selects relevant features from different modalities in both spatial and channel dimensions. By dynamically adjusting the convolution kernel and utilizing more appropriate convolution parameters for image features of different modalities, we mitigate the risk of over-fitting noisy areas while leveraging high-quality RGB images. Furthermore, in contrast to most other feature fusion modules that perform fusion at the feature map level, the DP-MF module focuses on optimizing memory utilization by increasing the feature extraction power of the convolution kernel. Moreover, the DP-MF module requires less computing power as the feature map size is much larger than the size of the convolutional kernel. It minimizes redundant computations and efficiently reuses intermediate results, thereby reducing memory footprint without compromising performance. It is particularly beneficial for resource-limited environments or large-scale models.

Moreover, leveraging the different dependencies on channels allows us to effectively utilize the complementary information from multi-modal sources, thereby aiding in the tracking of objects in scenarios involving overlap and occlusion. The module structure is depicted in Fig. 4. The proposed Dynamic Perception of Multi-Modal Features (DP-MF) model takes two inputs: the RGB feature map and the Tir feature map. It generates a more appropriate convolution kernel specific to each track by considering the characteristics of the respective feature maps, thereby improving feature extraction. To elaborate, we initially apply n filters with a kernel size of 3×3, dynamically adjusting the weights of the channel dimension for each filter and the weight assigned to each filter based on the unique input features. Subsequently, the adjusted filters are summed and subject to channel attention for the final refinement. Mathematically, the dynamic filter generation can be represented by:

$$\begin{aligned} \mathbf{X} &= \operatorname{cat}\left(\mathbf{F}_{rgb}, \mathbf{F}_{t}, dim\right) \\ \mathbf{W}_{c} &= \mathrm{F}_{sq}\left(\mathbf{X}\right) \\ &= \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X(i,j) \end{aligned} \tag{2}$$

Where $\operatorname{cat}(*, \operatorname{dim})$ denotes a feature map concatenation operation in the unique dimension, the weights $\mathbf{W}_c$ generated by using 1 are equivalent to indicating the distribution of values, or

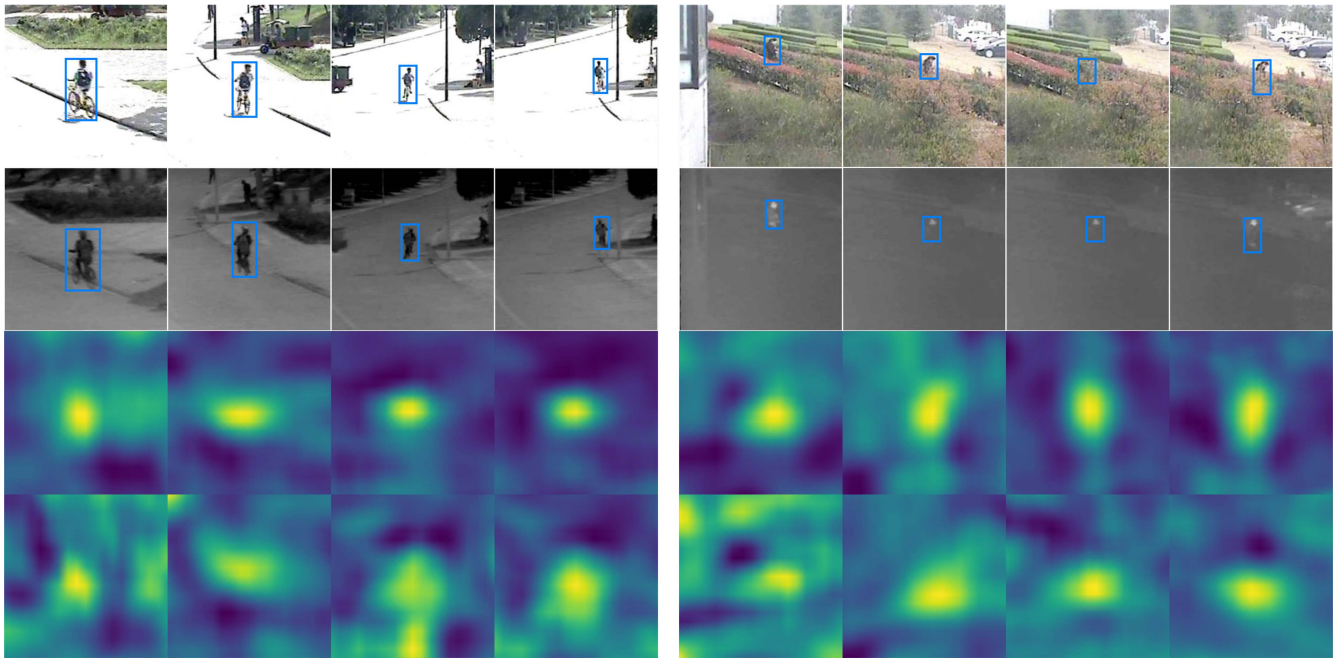

Fig. 5. Visualization of features maps. The first and second row is the RGB and thermal images respectively where the green bounding box indicates the ground truth. The third and fourth row is the feature maps of ours and baseline respectively

global information, for the $\mathbf{c}$ feature maps in that layer.

$$\begin{aligned}\mathbf{S}_c &= \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) \\ &= \sigma\left(\mathbf{W}_{c2}\delta\left(\mathbf{W}_{c1}\mathbf{W}_c\right)\right)\end{aligned} \tag{3}$$

The dimension of $\mathbf{W}_{*1}$ is $(\mathbf{C}/\mathbf{r}) \times \mathbf{C}$. This r is a scaling parameter intended to reduce the number of channels, hence the amount of computation. The dimension of $\mathbf{z}$ is $1 \times 1 \times \mathbf{C}$, so the result of $\mathbf{W}_{*1}\mathbf{W}_c$ is $1 \times 1 \times (\mathbf{C}/\mathbf{r})$; then it goes through a ReLU layer, and the output dimension remains the same; then it is multiplied by $\mathbf{W}_{c2}$, which is also fully connected layers. The dimension of $\mathbf{W}_{c1}$ is $\mathbf{K}_c \times \mathbf{K}_c$. Overall, this process establishes a channel attention mechanism for convolutional kernels, dynamically adjusting them in response to the yield of each feature map. Consequently, it enables the extraction of higher-quality feature information within the feature dimension subspace.

$$\begin{aligned}\mathbf{S}_d &= \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) \\ &= \sigma\left(\mathbf{W}_{d2}\delta\left(\mathbf{W}_{d1}\mathbf{W}_c\right)\right)\end{aligned} \tag{4}$$

The dimension of $\mathbf{W}_{p1}$ is $\mathbf{K}_p \times \mathbf{K}_c$. It produces a deep attention mechanism for convolutional kernels that dynamically adjusts the convolutional kernels in response to each feature map yield and extracts better quality feature information in the feature dimension subspace.

$$\begin{aligned}\mathbf{S}_s &= \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) \\ &= \sigma\left(\mathbf{W}_{s2}\delta\left(\mathbf{W}_{s1}\mathbf{W}_c\right)\right)\end{aligned} \tag{5}$$

The dimension of $\mathbf{W}_{s1}$ is $\mathbf{K}_s \times \mathbf{K}_c$. It produces a spatial attention mechanism with convolutional kernels, as not all regions in the perceptual field contribute equally to the task. Only task-relevant regions are the information to be attended to, so a spatial attention mechanism with convolutional kernels enhances the extraction of valid information.

$$\begin{aligned}\mathbf{S}_a &= \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) \\ &= \sigma\left(\mathbf{W}_{a2}\delta\left(\mathbf{W}_{a1}\mathbf{W}_c\right)\right)\end{aligned} \tag{6}$$

The dimension of $\mathbf{W}_{a1}$ is $\mathbf{K}_a \times \mathbf{K}_c$. It produces an attention mechanism for multiple convolutional kernels, as weighting is applied to the convolutional kernels. For different inputs, we use different convolutional kernels. Afterward, for these different convolutional kernels, attention is weighted.

$$\begin{aligned}\mathbf{S}_{se} &= \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) \\ &= \sigma\left(\mathbf{W}_{se2}\delta\left(\mathbf{W}_{se1}\mathbf{W}_c\right)\right)\end{aligned} \tag{7}$$

The dimension of $\mathbf{W}_{a1}$ is $\mathbf{K}_{se} \times \mathbf{K}_c$. It produces an attention mechanism for multiple convolutional kernels, as weighting is applied to the convolutional kernels. For different inputs, we use different convolutional kernels. Afterward, for these different convolutional kernels, attention is weighted. In order to demonstrate the efficacy of our proposed approach, we present a visual analysis of the modulated features across selected frames. As depicted in Fig. 6, our DF-MP model exhibits a discernible enhancement in feature representation for RGBT tracking tasks.

C. Dual Branch Fusion Module

The Region Proposal Network (RPN) plays a crucial role in classifying foreground and background elements and performing bounding box regression. In our SiamTDR framework, we introduce two classification branches and two regression

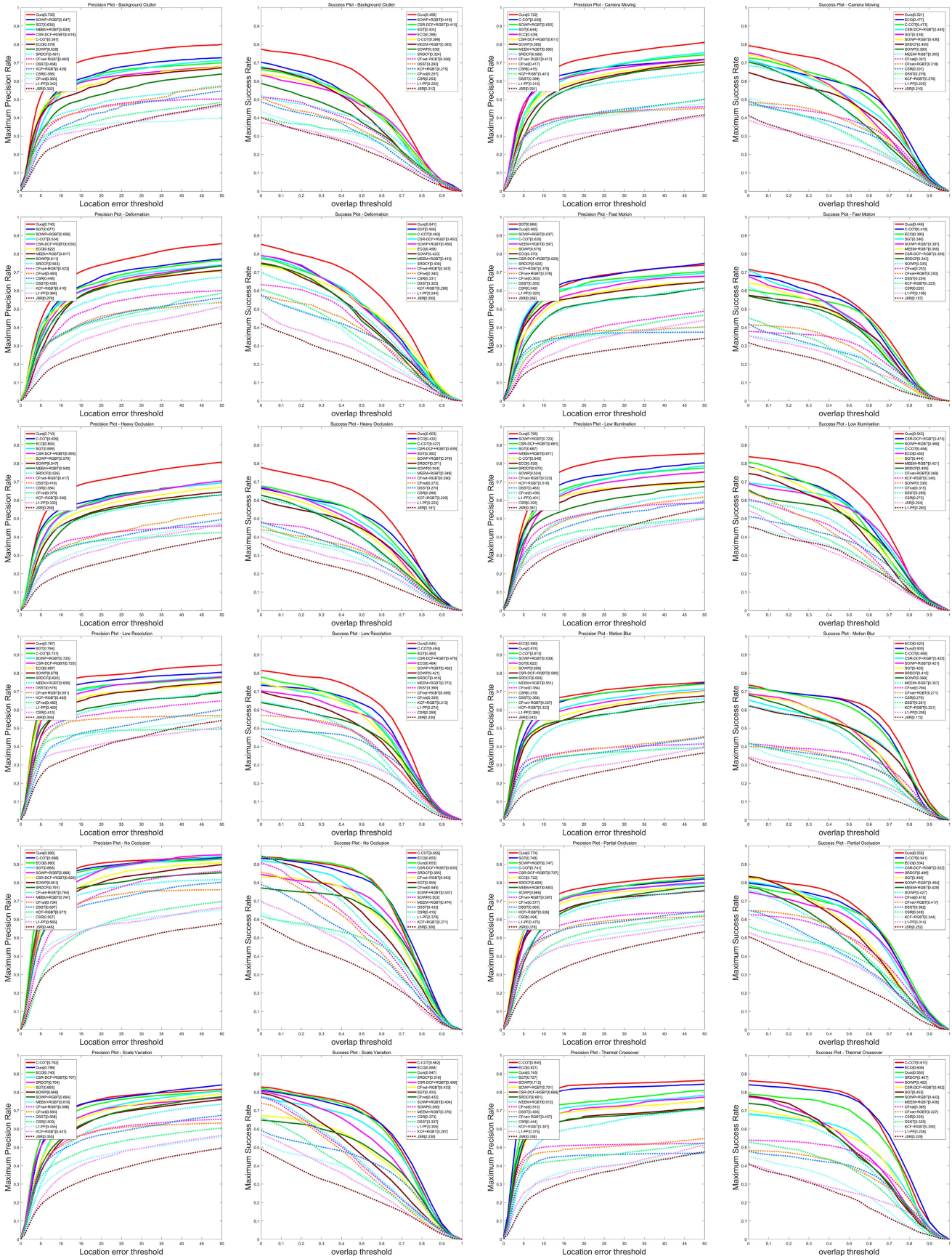

Fig. 6. Maximum success rate of tracking algorithm based on attributes on RGB-T234, our method achieves SOTA or lead results in most challenges.

branches to enhance the capabilities of the RPN. To further strengthen the robustness of the tracking inference phase, this study proposes a Dual Branch Fusion module (DBF) that combines the RGB and feature fusion branches to generate a final regression feature. Specifically, the two regression branches are integrated through feature summation, resulting in improved localization of bounding boxes. Similarly, the two classification branches are combined using feature summation to obtain a final classification feature, which is subsequently processed to produce the ultimate result. Additionally, the DBF module considers the different contributions of the RGB and TIR modalities in the target tracking algorithm and the need for speed. The introduction of RGB features helps to balance the amount of information input during the fusion process. Considering the information content ratio of the RGB image to the original TIR image is 3:1, incorporating RGB features with visual details based on the decomposition of the representation enables the fusion model to achieve a more balanced representation, mitigating the risk of over-reliance on the TIR modality. Furthermore, the DBF module does not require complex computations such as feature extraction or attention mechanisms, ensuring the real-time performance of the algorithm.

For the RGB and Tir modalities, two dedicated classification branches exist to classify the foreground and background of anchor boxes. Let us consider the scenario where 'n' anchor boxes with varying size ratios are present at each anchor point. Each classification branch has 2n output channels in such cases, representing n binary classifications for positive and negative samples. Taking the RGB modality as an example, when 'n' boxes of different scales are present at each anchor, the corresponding classification branch for this modality needs to expand the channel count to 2n through the convolutional layer. On the other hand, the input $\psi(x_r)$ requires size transformation via the convolutional layer but does not necessitate an increase in the number of channels. Denoting the operation through the convolutional layer as $(\cdot)cls$, we can express $\psi(z_r)$ and $\psi(x_r)$ after passing through the convolutional layer as $[\psi(z_r)]cls$ and $[\psi(x_r)]cls$, respectively. By utilizing $[\psi(z_r)]cls$ as the convolutional kernels and convolving it with $[\psi(x_r)]_{cls}$, we can calculate the correlation between the two features.

$$R_{cls} = [\psi(x_r)]_{cls} * [\psi(z_r)]_{cls} \tag{8}$$

$*$ stands for convolution operation. Similarly, the classification branch of the thermal infrared mode can be obtained:

$$M_{cls} = [\psi(x_m)]_{cls} * [\psi(z_m)]_{cls} \tag{9}$$

R_{cls} and M_{cls} denote the probability that each anchor frame at the corresponding location on the original map for the two modalities is predicted to be a background and a target, respectively. We then fuse the two classification results through a feature summarization operation so that the final classification result L_{cls} can be obtained:

$$L_{cls} = M_{cls} + R_{cls} \tag{10}$$

We mark all odd channels as L'_{cls} and record the predicted probabilities of all anchor boxes as set $P. = \{p'_i \mid i \in [0, k)\}$, where p_i represents the probability that a specific anchor box is predicted as a positive sample, k represents the total number of anchor boxes and make the bounding box whose confidence score is smaller than a threshold ve is first removed from the bounding box. The cross-entropy loss function used in Faster R-CNN [32] is used when training the two classification branches. Then, the loss classified for each anchor box is shown in Equation

$$\begin{aligned} \text{loss}_{cls}\left(y_i, p_i\right) &= -\log\left[y_i p_i + \left(1 - y_i\right)\left(1-\right)p_i\right] \\ y_i &= \begin{cases} \text{preserved} & \text{if } s_i^{cls} > v_e \\ \text{negative} & \text{if } s_i^{cls} < v_e \end{cases} \end{aligned} \tag{11}$$

Among them, $p_i = p_i^r + p_i^m$ combines the predictions of two modal classification branches. The total classification loss is:

$$L_{cls} = \sum_{i=1}^{k} \text{loss}_{cls}\left(y_i, p_i\right)/k \tag{12}$$

Where k represents the total number of anchor boxes.

The regression branch in this subnetwork regresses the anchor boxes to get a better bounding box. This algorithm employs information about the RGB and Tir modalities to regress the bounding boxes. Because dx, dy, dw, dh are required to calculate the distance between the anchor boxes and the ground truth, the number of channels of the regression branch is $4n$. The following formula can be obtained according to the classification branch::

$$\begin{aligned} R_{reg} &= [\psi(x_r)]_{reg} * [\psi(z_r)]_{reg} \\ M_{reg} &= [\psi(x_m)]_{reg} * [\psi(z_m)]_{reg} \\ L_{reg} &= M_{reg} + R_{reg} \end{aligned} \tag{13}$$

L_{reg} represents the predicted offset between each anchor box and the corresponding ground truth box. We transcribe it into the vector $c_i^*, i \in [0, k)$ according to the mathematical rules; meanwhile, the actual offset of each anchor box and the corresponding ground truth box is recorded as $c_i, i \in [0, k)$. We use the smooth L1 loss with normalized coordinates used in Faster r-cnn [32] to supervise the training of the regression branch:

$$\text{Lreg}\left(c_i^*, c_i\right) = \sum_{i=1}^{k} \text{smooth}_{L1}\left(c_i^* - c_i\right)/k \tag{14}$$

According to the content in (10) and (12), we get the total loss function used during training as:

$$L = L_{cls} + \gamma L_{reg} \tag{15}$$

where y is the hyper-parameter of the balanced two parts.

IV. Experiments

A. Datasets and Evaluation Metrics

The performance of SiamTDR is evaluated on two popular datasets: GTOT [57] and RGB-T234, which are captured by a visible-infrared camera. GTOT comprises 50 visible and thermal infrared video sequences, nine object classes, and 7.8 K aligned frame pairs. The GTOT dataset has seven challenge attributes, as shown in Table II. There are 234 visible and thermal infrared video sequences, 22 object classes, and 117 K aligned frame

TABLE II
LIST OF THE SEVEN CHALLENGE ATTRIBUTES ATTACHED TO GTOT

Attribute	Description
OCC	Occlusion - the target is partially or fully occluded.
LSV	Large Scale Variation - the ratio of the first bounding box and the current bounding box is out of the range [0.5, 1].
FM	Fast Motion - the motion of the ground truth is larger than 10 pixels.
LI	Low Illumination - the illumination in the target region.
TC	Thermal Crossover - the target has similar temperature with other objects or background.
SO	Small Object - the number of pixels in the ground truth bounding box is less than 400.
DEF	Deformation - non-rigid object deformation.

TABLE III
LIST OF THE ELEVEN CHALLENGE ATTRIBUTES ATTACHED TO RGBT234

Attribute	Description
NO	No Occlusion - the target is not occluded.
PO	Partial Occlusion - the target object is partially occluded.
HO	Heavy Occlusion - the target object is heavy occluded (over 80% percentage).
LI	Low Illumination - the illumination in the target region is low.
LR	Low Resolution - the resolution in the target region is low.
TC	Thermal Crossover - the target has similar temperature with other objects or background surroundings.
DEF	Deformation - non-rigid object deformation.
FM	Fast Motion - the motion of the ground truth between two adjacent frames is larger than 20 pixels.
SV	Scale Variation - the ratio of the first bounding box and the current bounding box is out of the range [0.5,1].
MB	Motion Blur - the target object motion results in the blur image information.
CM	Camera Moving - the target object is captured by moving camera.
BC	Background Clutter - the background information which includes the target object is messy.

pairs in RGB-T234. The RGB-T234 dataset has eleven challenge attributes, as shown in Table III. We use LasHeR, which consists of 1224 visible and thermal infrared video sequences and 730 K frame pairs for all, as a training dataset. We test it on GTOT and RGB-T234, respectively. Specifically, PR is the percentage of frames in which the Euclidean distance between the predicted position and the ground truth is less than the location error threshold. Take the PR with a location error threshold of 5 on GTOT as the PR score (because most of the targets in GTOT are small), and take the PR with a location error threshold of 20 on RGB-T234 as the PR score. SR is the percentage of frames whose overlap ratio between the predicted bounding box and the ground truth is greater than the overlap threshold, and the area under curves (AUC) is counted as the SR score.

B. Implementation Details

For each point of the final response map, our anchor boxes have five aspect ratios,i.e.,[0.33,0.5,1,2,3], and the anchor scale is set to 8. We determine the correspondences between the anchors and ground truth boxes in Siamese-RPN blocks based on IoU. Specifically, if the IoU between the anchor and ground-truth box is more significant than 0.6, the anchor is determined as positive. Meanwhile, if the IoU between the anchor and ground-truth box is less than 0.3, the anchor is determined to be negative. We collect at most 16 positive samples and 48 negative samples from one image pair.

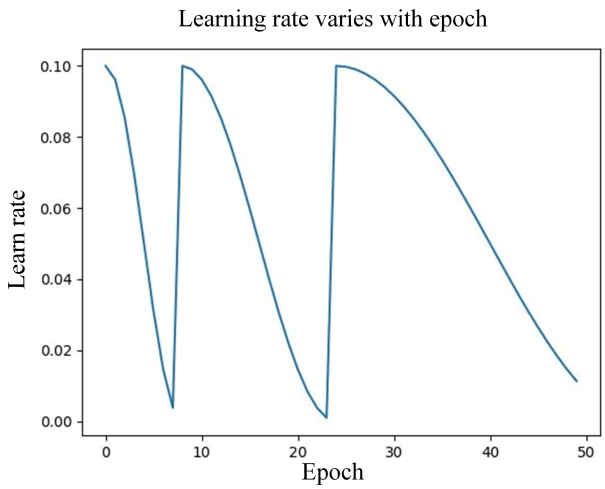

Fig. 7. To avoid the local optimum trap, the learning rate is continuously restarted as the epoch increases.

Our experiments use the SGD optimizer, which contains a certain degree of stochasticity that makes SGD not always oriented towards the overall optimum and not necessarily the global optimum in every iteration. Hence, the learning rate escapes from the local optimum using the CosineAnnealing strategy. The learning variation is shown in Fig. 7. We set batchsize as 28 and trained our model for 50 epochs. Moreover, the momentum is 0.9, and the weight decay is 5×10^{-4}.

C. Evaluation on GTOT Dataset

In this study, we sought to evaluate the efficacy of our proposed SiamTDR tracker through two comparison experiments conducted on the GTOT dataset. The experiments focused on overall performance comparison and challenge performance comparison and involved the inclusion of seven trackers, namely SiamBAN [40], SiamRPN++ [39], ATOM [29], DiMP [30], SiamFT [35], SGT [56], mfDiMP [43], and our proposed SiamTDR. Since there are limited existing RGBT trackers, we integrated the features of the RGB mode and the thermal infrared mode into a tensor to expand the single-mode tracker into an RGBT tracker for a fair comparison. Notably, the first six are derived from RGB trackers among the seven trackers.

As depicted in Table IV, the experimental results demonstrated that our proposed SiamTDR outperformed the other six trackers. Specifically, our proposed SiamTDR achieved a PR of 0.885 and an SR of 0.714, representing a 0.049 and 0.017 improvement, respectively, over the second-ranked tracker, mfDiMP (with a PR of 0.836 and an SR of 0.697). These outcomes provide evidence of the superior performance of our proposed algorithm. Furthermore, compared to the baseline tracker, SiamRPN, our proposed SiamTDR recorded a PR increase of 0.088 and an SR increase of 0.065, further affirming the robustness of our proposed SiamTDR tracker for tracking applications.

TABLE IV
PR/SR (%) RESULTS WITH DIFFERENT RGBT TRACKERS UNDER DIFFERENT CHALLENGES ON GTOT

Method	SiamBAN [40]	CMRT [54]	SiamRPN++ [39]	ATOM [29]	DiMP [30]	SiamFT [35]	SGT [55]	mfDiMP [42]	Ours
OCC	67.2/54.9	84.7/65.2	70.3/58.7	67.4/55.1	75.7/63.8	75.3/58.6	81.0/56.7	80.7/64.3	**85.2/67.7**
LSV	78.3/64.2	88.7/67.8	76.5/64.3	78.9/64.2	81,4/69.0	79,7/61.4	84.2/54.7	**90.5/73.9**	87.1/71.4
FM	74.3/62.0	**83.5**/65.0	75.9/65.9	74.8/63.0	78.9/68.0	72.1/60.1	79.9/55.9	81.3/**68.7**	82.8/68.5
LI	66.8/56.0	86.5/61.0	68.9/58.3	68.3/58.4	69.8/61.1	78.6/63.6	88.4/65.1	83.0/70.4	**88.7/76.4**
TC	76.3/61.0	85.3/66.7	76.6/64.0	79,0/63.3	84.2/68.7	76,0/59.3	84,8/61.5	80,4/65.2	**88.0/70.1**
DEF	66.1/55.5	71.1/62.2	71.0/59.3	69.1/58.8	69.9/59.9	72.5/61.9	**91.9/73.3**	80.7/67.1	87.9/72.7
SO	79.3/59.3	82.5/62.6	82.2/64.7	83.7/62.9	84.2/64.0	79.3/59.3	91.7/61.8	87.4/69.1	88.7/**70.0**
ALL	71.7/59.3	82.7/64.3	72.5/61.7	72.6/61.2	75.7/64.9	75.8/62.3	85.1/62.8	83.6/69.7	**88.5/71.4**
FPS	40	24	35	30	29	32	24	11	**127**

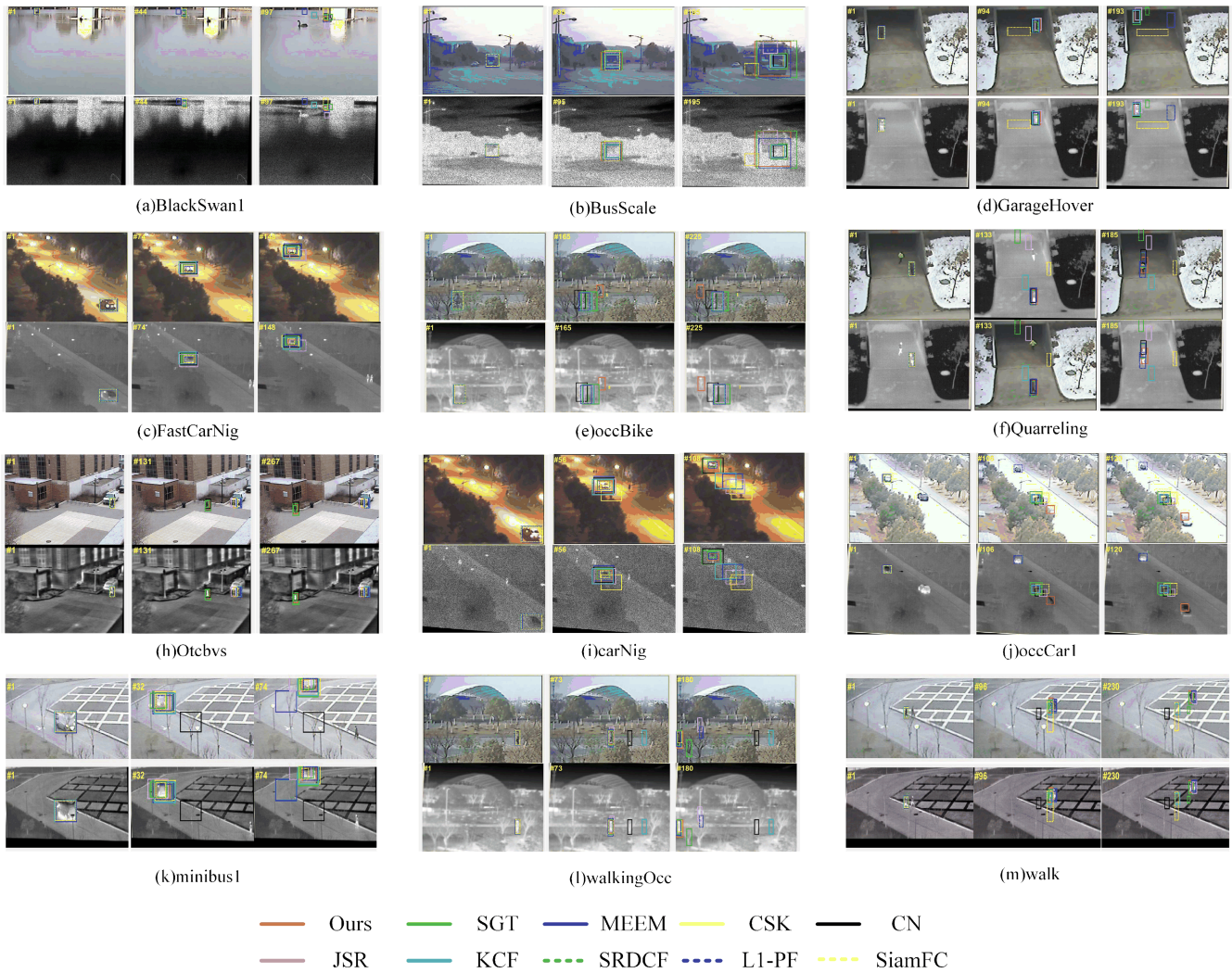

Fig. 8. Visual comparisons of our proposed tracker with another nine state-of-the-art trackers on six video sequences, i.e., lackSwan1, BusScale, FastCarNig, GarageHover, occBike, and Quarreling.

D. Evaluation on RGB-T234 Dataset

We evaluate the performance of SiamTDR, our proposed tracker, on the RGB-T234 dataset. To compare its effectiveness, we evaluate it against fifteen other trackers, which include CSR-DCF+RGBT [58], SOWP+RGBT [59], MEEM+RGBT [60], CFnet+RGBT [61], KCF+RGBT [10], C-COT [13], ECO [62], SGT [56], SOWP, DSST [11], SRDCF [12], CSR, CFnet, L1-PF [63], JSR [64]. The initial five trackers utilize RGBT data, while the remaining solely rely on RGB data. As illustrated in Fig. 5, our findings reveal that SiamTDR achieves superior performance on the RGB-T234 dataset compared to other

TABLE V
TRACKING OF THE RESULTS OBTAINED BY USING DIFFERENT MODULES

Fine-tune	Sign-model	MF-DP	DBF	Pr	Sr	ΔPr	ΔSr
✓				82.9	65.9	+4.5	+1.3
✓	✓			86.0	68.4	+7.6	+3.8
✓	✓	✓		86.9	69.9	+8.5	+5.3
✓	✓	✓	✓	88.5	71.3	+10.1	+6.7

algorithms. Precisely, our tracker attains a PR score of 0.772 and an SR score of 0.551, representing a 0.063 and 0.221 improvement over the second-ranked SGT and CFnet +RGBT, respectively, demonstrating the effectiveness of SiamTDR.

Additionally, our algorithm outperforms the standard Siamese framework that utilizes two modalities under the same training dataset. This further validates the effectiveness of our proposed approach, which efficiently integrates the information from both modalities to achieve robust object tracking.

E. Ablation Study

To validate the effectiveness of different components (or modules) in our proposed model, we construct simplified versions of our proposed tracker as the baseline.

1) In order to evaluate the effectiveness of the various components of our proposed model, a simplified version of our baseline tracker was first created. It involved using a feature extractor that lacked feature decoupling and a proposed network without feature fusion regions. Moreover, the DP-MF module was replaced with primary convolutional layers, and both the hot Siamese and RGB Siamese networks shared identical weights. After that, different modules or strategies were added to the baseline, and the performance of our tracker was evaluated using GTOT, with the results presented in Table V. As indicated in Table V, the baseline tracker achieved a Pr of 0.784 and Sr of 0.646. Subsequently, fine-tuning the Siamese Network feature extraction network improved Pr/Sr by 0.045/0.013, respectively. Further improvement was achieved by adding the feature decoupling mechanism module to the feature extractor, increasing Pr/Sr. Adding the DP-MF module to the baseline tracker, Pr/Sr improved by 0.085/0.053. Finally, adding the DBF module improved Pr/Sr by 0.103/0.08, indicating that our proposed module is highly effective and stable.

2) To establish the efficacy of our proposed DP-MF module, we have incorporated four distinct fusion strategies into our tracking system, allowing for multi-modal features to be fused. In this manner, we have developed four distinct versions of our approach, each with its unique fusion methodology, including Elementwise summation, Concatenation, Attention fusion strategy for channels, and The proposed DP-MF module. A comparative analysis of these four approaches was conducted, and the resulting experimental data, as presented in Table VI, clearly indicates that the proposed DP-MF module exhibits superior performance over the other fusion modules.

F. Efficiency Analysis

The proposed tracking algorithm is implemented on the PyTorch framework, and all experimental evaluations are conducted on the same server configuration consisting of an Intel(R) Xeon(R) Silver 4210 CPU @ 2.20 GHz and an NVIDIA Corporation GA102 GeForce RTX 3090 with 24 GB memory. Fig. 8 compares our algorithm's tracking performance with SiamFC, SGT, MEEM +RGBT, SRDCF+RGBT, CN, JSR, KCF, L1-PF, and CSK on twelve video sequences. The experimental results show that our algorithm outperforms the compared methods in handling challenging scenarios such as heavy occlusion, low visibility, thermal crossover, scale variations, and deformation. Before the experiments, data enhancement is carried out using the MSRCR method on the pictures in the dataset, enhancing image color, detail, and local contrast. The pictures are resized to the size of 640 x 640. During training, the SGD optimizer is used, the learning rate is warming up, the weight decay is set to 0.0005, and the number of epochs is 300.

Furthermore, our algorithm demonstrates high computational efficiency with an average frame per second (FPS) of 127, sufficient for real-time object tracking applications. Table VII presents the FPS comparison between our algorithm and four representative methods to provide a comprehensive comparison. These results highlight that our proposed algorithm achieves high performance and maintains high efficiency, making it a promising solution for practical tracking scenarios.

TABLE VI
TRACKING OF THE RESULTS OBTAINED BY USING DIFFERENT FUSED MODULES

Fosion model	Pr	Sr	ΔPr	ΔSr
Ours-ES	85.6	69.8	0	0
Ours-CO	85.2	69.1	-0.4	-0.7
Ours-SE	86.4	70.3	+1.8	+0.5
Ours	88.5	71.3	+2.9	+1.5

TABLE VII
COMPARISON OF THE AVERAGE FPS OF DUSIAMRT AND OTHER FOUR REPRESENTATIVE TRACKING ALGORITHMS WHEN RUNNING ON THE GTOT DATASET

SiamTDR	RT-MDNet	CFNet+RGBT	SiamDW+RGBT	SiamFC
127	16	33	95	38

V. THE VALUE OF SIAMTDR ON CYBER-PHYISCAL ASPECTS

Object tracking is a fundamental task in CPS, as it enables the system to monitor and predict the behavior of objects in real-time [4], [5]. Object tracking algorithms use computer vision and machine learning techniques to detect, locate, and track objects of interest. This information is crucial for tasks such as collision avoidance, surveillance, resource allocation, and decision-making in CPS applications. SiamTDR combines the advantages of RGB (color) and thermal (heat signature) imaging to provide a more comprehensive understanding of the environment. Combining the two modalities improves robustness and reliability, allowing more accurate and consistent tracking results in various environmental conditions. By fusing these modalities, it becomes possible to overcome limitations

like lighting conditions, camouflage, and occlusions. This enhanced perception is essential for accurate object tracking in CPS applications. SIamTDR has Real-Time Responsiveness. SiamTDR can provide timely and efficient updates on the location and trajectory of objects. This responsiveness is vital in CPS, where actions need to be taken based on the current state of the environment. By continuously tracking objects in real-time, the system can adapt and respond quickly to changes, ensuring the CPS's overall safety, efficiency, and reliability. Overall, SiamTDR offers improved perception, accuracy, and robustness in dynamic environments, making them highly valuable for cyber-physical systems. They enhance the capabilities of CPS in various domains, including security, safety, surveillance, and autonomous systems.Furthermore, our algorithm demonstrates high computational efficiency with an average frame per second (FPS) of 127, sufficient for real-time object tracking applications. Table VII presents the FPS comparison between our algorithm and four representative methods to provide a comprehensive comparison. These results highlight that our proposed algorithm achieves high performance and maintains high efficiency, making it a promising solution for practical tracking scenarios.

VI. Conclusion

In this article, we present a newly designed RGB-T Siamese tracker that exhibits state-of-the-art performance and operates in real-time, thanks to innovative modules' integration. Our proposed DP-MF enables the tracker to take full advantage of the complementary benefits of multi-modal features, resulting in satisfactory results in challenging scenarios such as heavy occlusion and illumination variations. Furthermore, the proposed DBF model enhances the tracker's resilience against distractors, such as semantic backgrounds, significantly improving overall tracking performance. We extensively evaluated our proposed tracker on two benchmark datasets, where it considerably outperformed existing RGB-T Siamese trackers. Our proposed tracker performs competitively compared to other state-of-the-art methods, with slightly better tracking accuracy and superior tracking speed.

References

[1] S. Yin, X. Li, H. Gao, and O. Kaynak, "Data-based techniques focused on modern industry: An overview," *IEEE Trans. Ind. Electron.*, vol. 62, no. 1, pp. 657–667, Jan. 2015.

[2] A. W. Colombo, S. Karnouskos, Y. Shi, S. Yin, and O. Kaynak, "Industrial cyber–physical systems [scanning the issue]," *Proc. IEEE*, vol. 104, no. 5, pp. 899–903, May 2016.

[3] A. Fisher, C. A. Jacobson, E. A. Lee, R. M. Murray, A. Sangiovanni-Vincentelli, and E. Scholte, "Industrial cyber-physical systems–iCyPhy," in *Proc. Complex Syst. Des. Manage.: Proc. 4th Int. Conf. Complex Syst. Des. Manage.*, pp. 21–37.

[4] A. W. Colombo, S. Karnouskos, O. Kaynak, Y. Shi, and S. Yin, "Industrial cyberphysical systems: A backbone of the fourth industrial revolution," *IEEE Ind. Electron. Mag.*, vol. 11, no. 1, pp. 6–16, Mar. 2017.

[5] Y. Jiang, S. Yin, and O. Kaynak, "Data-driven monitoring and safety control of industrial cyber-physical systems: Basics and beyond," *IEEE Access*, vol. 6, pp. 47374–47384, 2018.

[6] Y. Jiang et al., "Secure data transmission and trustworthiness judgement approaches against cyber-physical attacks in an integrated data-driven framework," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 52, no. 12, pp. 7799–7809, Dec. 2022.

[7] N. Nikolakis, V. Maratos, and S. Makris, "A cyber physical system (CPS) approach for safe human-robot collaboration in a shared workplace," *Robot. Comput.- Integr. Manuf.*, vol. 56, pp. 233–243, 2019.

[8] X. Wang et al., "Greedy batch-based minimum-cost flows for tracking multiple objects," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4765–4776, Oct. 2017.

[9] W. Luo, P. Sun, F. Zhong, W. Liu, T. Zhang, and Y. Wang, "End-to-end active object tracking and its real-world deployment via reinforcement learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 6, pp. 1317–1332, Jun. 2020.

[10] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[11] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–5.

[12] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4310–4318.

[13] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 472–488.

[14] A. Leykin, Y. Ran, and R. Hammoud, "Thermal-visible video fusion for moving target tracking and pedestrian classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[15] C. Ó. Conaire, N. E. O'Connor, and A. Smeaton, "Thermo-visual feature fusion for object tracking using multiple spatiogram trackers," *Mach. Vis. Appl.*, vol. 19, no. 5-6, pp. 483–494, 2008.

[16] C. Li, Z. Xiang, J. Tang, B. Luo, and F. Wang, "RGBT tracking via noise-robust cross-modal ranking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 5019–5031, Sep. 2022.

[17] C. Li, X. Wu, N. Zhao, X. Cao, and J. Tang, "Fusing two-stream convolutional neural networks for RGB-T object tracking," *Neurocomputing*, vol. 281, pp. 78–85, 2018.

[18] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4293–4302.

[19] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. Comput. Vis. Workshops*, 2016, pp. 850–865.

[20] C. Guo, D. Yang, C. Li, and P. Song, "Dual siamese network for RGBT tracking via fusing predicted position maps," *Vis. Comput.*, vol. 38, no. 7, pp. 2555–2567, 2022.

[21] P. Voigtlaender, J. Luiten, P. H. Torr, and B. Leibe, "Siam R-CNN: Visual tracking by re-detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6578–6588.

[22] Y. Zhu, C. Li, B. Luo, J. Tang, and X. Wang, "Dense feature aggregation and pruning for RGBT tracking," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 465–472.

[23] C. L. Li, A. Lu, A. H. Zheng, Z. Tu, and J. Tang, "Multi-adapter RGBT tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 2262–2270.

[24] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: Benchmark and baseline," *Pattern Recognit.*, vol. 96, 2019, Art. no. 106977.

[25] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[26] L. A. Varga, B. Kiefer, M. Messmer, and A. Zell, "Seadronessee: A maritime benchmark for detecting humans in open water," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 2260–2270.

[27] D. Gordon, A. Farhadi, and D. Fox, "Re^3: Re al-time recurrent regression networks for visual tracking of generic objects," *IEEE Robot. Automat. Lett.*, vol. 3, no. 2, pp. 788–795, Apr. 2018.

[28] Z. Zhang, H. Peng, J. Fu, B. Li, and W. Hu, "Ocean: Object-aware anchor-free tracking," in *Proc. 16th Eur. Conf. Comput. Vis.*2020: Glasgow, U.K., 2020, pp. 771–787.

[29] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "Atom: Accurate tracking by overlap maximization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4660–4669.

[30] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6182–6191.

[31] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with Siamese region proposal network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8971–8980.

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, vol. 28, pp. 1–9.
[33] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware Siamese networks for visual object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 101–117.
[34] H. Fan and H. Ling, "Siamese cascaded region proposal networks for real-time visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7952–7961.
[35] X. Zhang, P. Ye, S. Peng, J. Liu, K. Gong, and G. Xiao, "SiamFT: An RGB-infrared fusion tracking method via fully convolutional Siamese networks," *IEEE Access*, vol. 7, pp. 122122–122133, 2019.
[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[37] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1492–1500.
[38] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, pp. 1–9, 2017.
[39] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN: Evolution of siamese visual tracking with very deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4282–4291.
[40] Z. Zhang and H. Peng, "Deeper and wider siamese networks for real-time visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4591–4600.
[41] Y. Liang, J. Feng, X. Zhang, J. Zhang, and L. Jiao, "MidNet: An anchor-and-angle-free detector for oriented ship detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–13, 2023, doi: 10.1109/TGRS.2023.3280973.
[42] C. Li, C. Zhu, S. Zheng, B. Luo, and J. Tang, "Two-stage modality-graphs regularized manifold ranking for RGB-T tracking," *Signal Process.: Image Commun.*, vol. 68, pp. 207–217, 2018.
[43] L. Zhang, M. Danelljan, A. Gonzalez-Garcia, J. V. D. Weijer, and F. S. Khan, "Multi-modal fusion for end-to-end RGB-T tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 1–10.
[44] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 35–51.
[45] H. Xu, X. Wang, and J. Ma, "DRF: Disentangled representation for visible and infrared image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 5006713.
[46] Y. Li, K. K. Singh, U. Ojha, and Y. J. Lee, "Mixnmatch: Multifactor disentanglement and encoding for conditional image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8039–8048.
[47] S. Xia, L. Chu, L. Pei, Z. Zhang, W. Yu, and R. C. Qiu, "Learning disentangled representation for mixed-reality human activity recognition with a single IMU sensor," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 2514314.
[48] G. Li, Y. Gan, H. Wu, N. Xiao, and L. Lin, "Cross-modal attentional context learning for RGB-D object detection," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1591–1601, Apr. 2019.
[49] C. Li et al., "ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection," *IEEE Trans. Cybern.*, vol. 51, no. 1, pp. 88–100, Jan. 2021.
[50] V. Mnih et al., "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, vol. 27.
[51] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.
[52] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[53] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[54] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11030–11039.
[55] C. Li, C. Zhu, Y. Huang, J. Tang, and L. Wang, "Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 808–823.
[56] C. Li, N. Zhao, Y. Lu, C. Zhu, and J. Tang, "Weighted sparse representation regularized graph learning for RGB-T object tracking," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 1856–1864.
[57] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal tracking," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5743–5756, Dec. 2016.
[58] A. Lukezic, T. Vojir, L. Č. Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6309–6318.
[59] H.-U. Kim, D.-Y. Lee, J.-Y. Sim, and C.-S. Kim, "SOWP: Spatially ordered and weighted patch descriptor for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3011–3019.
[60] J. Zhang, S. Ma, and S. Sclaroff, "MEEM: Robust tracking via multiple experts using entropy minimization," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 188–203.
[61] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2805–2813.
[62] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6638–6646.
[63] Y. Wu, E. Blasch, G. Chen, L. Bai, and H. Ling, "Multiple source data fusion via sparse representation for robust visual tracking," in *Proc. IEEE 14th Int. Conf. Inf. Fusion*, 2011, pp. 1–8.
[64] H. Fan and H. Ling, "Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5486–5494.

Guorui Wang received the B.S. degree in software engineering from the Shandong University of Science and Technology, Qingdao, China, in 2022. He is working toward the M.S. degree in software engineering from the School of Software, Yunnan University, Kunming, China. His research interests include image processing, object tracking, and information fusion.

Qian Jiang received the B.S. degree in thermal energy and power engineering and the M.S. degree in power engineering and engineering thermo-physics from the Central South University, Changsha, China, in 2012 and 2015, respectively. She is currently an Associate Professor with the School of Software, Yunnan University, Kunming, China. She was a Postdoctoral Fellow with the School of Software, Yunnan University, from 2019 to 2021. Her research interests include machine learning, bio-informatics, and image processing.

Xin Jin (Member, IEEE) received the B.S. degree in electronics and information engineering from the Henan Normal University, Xinxiang, China, in 2013, and the Ph.D. degree in communication and information systems from Yunnan University, Kunming, China, in 2018. He is currently an Associate Professor with the School of Software, Yunnan University. He was a Postdoctoral Fellow with the School of Software, Yunnan University, from 2018 to 2020. His current research interests include pulse coupled neural networks theory and its applications, image processing, information fusion, optimization algorithm, and bio-informatics.

Yu Lin received the B.S. degree from Yunnan University, Kunming, China, and the Ph.D. degree in electron physics from the China Academy of Electronic Sciences, Beijing, China. He is currently a Professor with the Kunming Institute of Physics, Kunming, China. His research interests include information fusion, infrared image processing, and object detection.

Yuanyu Wang received the B.S. degree in material science and engineering and the Ph.D. degree in material science and engineering from Tsinghua University, Beijing, China, in 2012 and 2017, respectively. He is currently a Senior Engineer with the Kunming Institute of Physics, Kunming, China. His current research interests include neural networks theory and its application, infrared image processing, infrared target tracking, and detection.

Wei Zhou (Member, IEEE) received the Ph.D. degree in computer theory and software from the University of Chinese Academy of Science, Beijing, China, in 2008. He is currently a Full Professor with the School of Software, Yunnan University, Kunming, China. He is a Fellow of the China Institute of Communications. His research interests include distributed data intensive computing, image processing, bioinformatics, neural networks, and information security.