Bridging Search Region Interaction with Template for RGB-T Tracking

Tianrui Hui[1,2,4] Zizheng Xun[3,5] Fengguang Peng[3,5] Junshi Huang[4] Xiaoming Wei[4]
Xiaolin Wei[4] Jiao Dai[1,2*] Jizhong Han[1,2] Si Liu[3,5]
[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] School of Cyber Security, University of Chinese Academy of Sciences
[3] Institute of Artificial Intelligence, Beihang University
[4] Meituan [5] Hangzhou Innovation Institute, Beihang University

Abstract

RGB-T tracking aims to leverage the mutual enhancement and complement ability of RGB and TIR modalities for improving the tracking process in various scenarios, where cross-modal interaction is the key component. Some previous methods concatenate the RGB and TIR search region features directly to perform a coarse interaction process with redundant background noises introduced. Many other methods sample candidate boxes from search frames and conduct various fusion approaches on isolated pairs of RGB and TIR boxes, which limits the cross-modal interaction within local regions and brings about inadequate context modeling. To alleviate these limitations, we propose a novel Template-Bridged Search region Interaction (TBSI) module which exploits templates as the medium to bridge the cross-modal interaction between RGB and TIR search regions by gathering and distributing target-relevant object and environment contexts. Original templates are also updated with enriched multimodal contexts from the template medium. Our TBSI module is inserted into a ViT backbone for joint feature extraction, search-template matching, and cross-modal interaction. Extensive experiments on three popular RGB-T tracking benchmarks demonstrate our method achieves new state-of-the-art performances. Code is available at https://github.com/RyanHTR/TBSI.

1. Introduction

Given the initial state of a single target object in the first frame, the goal of single object tracking (SOT) is to localize the target object in successive frames. As a fundamental task in the computer vision community, SOT has drawn the great attention of researchers. However, current SOT methods built on only visible light (RGB) data become vulnerable under extreme imaging conditions (*e.g.*, low illumination and adverse weather, *etc*), which motivates the incorporation of thermal infrared (TIR or T) data for mutual enhancement and complement. Benefiting from the strong nocturnal photosensitivity and penetration ability of thermal infrared data, RGB-T tracking enjoys wide potential applications such as video surveillance processing [1], intelligent robotics [5], and autonomous driving [8].

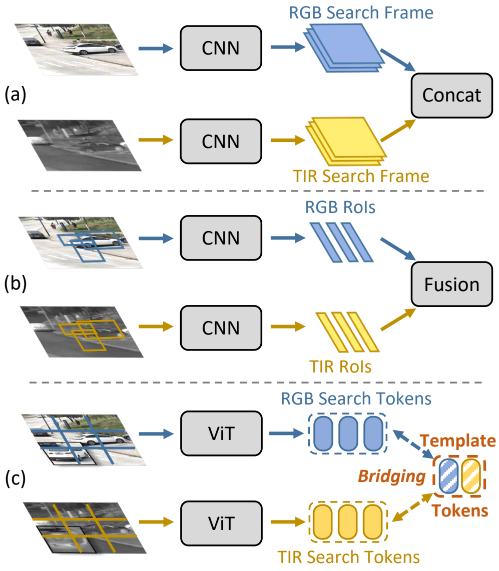

Figure 1. Comparison between our cross-modal interaction approach and previous ones. (a) Features of RGB and TIR search frames are directly concatenated. (b) Candidate boxes (RoIs) are sampled from RGB and TIR search frames and fused in pairs with gating or attention mechanisms. (c) Our approach exploits template tokens as the medium to bridge the cross-modal interaction between RGB and TIR search region tokens.

As a multimodal vision task, the key to RGB-T tracking is how to perform effective cross-modal interaction. Since the tracking process occurs in successive frames guided by the annotated initial frame, cross-modal interaction be-

[*] Corresponding author

tween search frames of RGB and TIR modalities becomes the main focus. As illustrated in Figure 1 (a), some previous methods [16, 44] directly concatenate features of the whole RGB and TIR search frames from the encoders of strong base trackers [4, 40]. This simple manner tends to introduce redundant background noise information, making cross-modal interaction too coarse and hence harming the model's discriminative ability. In addition, there are many other methods [14,27,28,37,39,49] which sample candidate boxes (RoIs) from the Gaussian distribution in the search frames and conduct various fusion operators based on attention, gating mechanism, or dataset attributes, *etc*, to fuse each pair of RoI features of RGB and TIR modalities as shown in Figure 1 (b). Then, fused RoI features are separately fed into a binary classifier to distinguish the target object. However, each pair of RoIs merely crops a small portion of local features from the search frames, containing limited foreground and background information. Thus, cross-modal interaction between each isolated pair of RoIs may bring about inadequate modeling of the global environment context in the search frame and restrict the mutual enhancement and complement effect of the two modalities.

Given the above discussion, we argue that direct cross-modal interaction between RGB and TIR search frames or candidate RoIs still has limitations in comprehensively leveraging complementary multimodal clues to facilitate the tracking process. Therefore, we propose *a novel scheme which exploits the target templates as the medium to bridge the cross-modal interaction between RGB and TIR search regions*, as illustrated in Figure 1 (c). The major superiority motivating our advocate of this scheme is that the templates contain original multimodal information of the target object, which can serve as strong guidance to extract *target-relevant* object and environment contexts from search regions for adaptive and precise information enhancement and complement. The background noises of other distractors in search regions can also be reduced by template bridging during the cross-modal interaction process.

In order to implement the above scheme, we design a Template-Bridged Search region Interaction (TBSI) module. Concretely, our TBSI module first fuses features of RGB and TIR templates to obtain the multimodal context medium. Since the cross-attention mechanism [36] is an effective and widely-adopted practice for context aggregation, our TBSI also utilizes it with the fused template as query and TIR search region feature as key and value to gather target-relevant TIR context information into the template medium. Then, the RGB search region feature serves as query and the fused template serves as key and value to distribute target-relevant TIR context from the medium to the RGB search region. Similarly, target-relevant RGB context is also gathered and distributed to the TIR search region through the template medium in a reverse direction. Finally, comprehensive multimodal information aggregated in the fused template is transferred back to the original RGB and TIR templates to update them with the enriched multimodal contexts gathered from search regions.

In addition, most existing RGB-T tracking methods [14, 27, 28, 37, 39, 49] employ MDNet [32] with VGG-M [34] as the base tracker, whose number of classification branches equals the number of training sequences, which largely limits their capacity and scalability. Inspired by the powerful ability of Vision Transformer (ViT) [12] to capture long-range dependencies and its recent success on SOT [7, 24, 42], we also extend ViT to RGB-T tracking for joint feature extraction, search-template matching, and cross-modal interaction. Our TBSI module is inserted into the ViT base tracker to bridge the intra-modal information flow within the Transformer layers for effective RGB-T tracking.

Our contributions are summarized as follows: (1) We propose a novel Template-Bridged Search region Interaction (TBSI) module which exploits the fused target template as the medium to bridge the cross-modal interaction between RGB and TIR search regions and update original templates as well, forming adaptive and precise information enhancement. (2) We extend the ViT architecture with the proposed TBSI module to RGB-T tracking for joint feature extraction, search-template matching, and cross-modal interaction, which has not been explored by previous methods to our best knowledge. (3) Extensive experiments demonstrate that our method achieves new state-of-the-art performances on three popular RGB-T tracking benchmarks.

2. Related Work

2.1. Single Object Tracking

As one of the fundamental vision tasks, notable progress has been achieved on SOT for accurate and stable target object tracking in various scenarios. Siamese-based methods [3, 17, 18, 40, 52] utilize correlation operator to compute matching responses between template and search region with the Siamese network. Some online updating methods [4, 9, 10, 29, 32] learn a target-dependent discriminative classifier to distinguish the target object from the background in search frames. Recently, some Transformer-based methods [6, 38, 41] leverage self-attention and cross-attention to integrate search region and template information for matching relationship modeling, and explore to jointly extract their features via ViT backbones [7, 24, 42]. In this paper, we extend ViT to a multimodal base tracker equipped with our TBSI module for joint feature extraction, search-template matching, and cross-modal interaction.

2.2. RGB-T Tracking

General SOT methods are trained only on visible light data so that they are inclined to encounter failures under

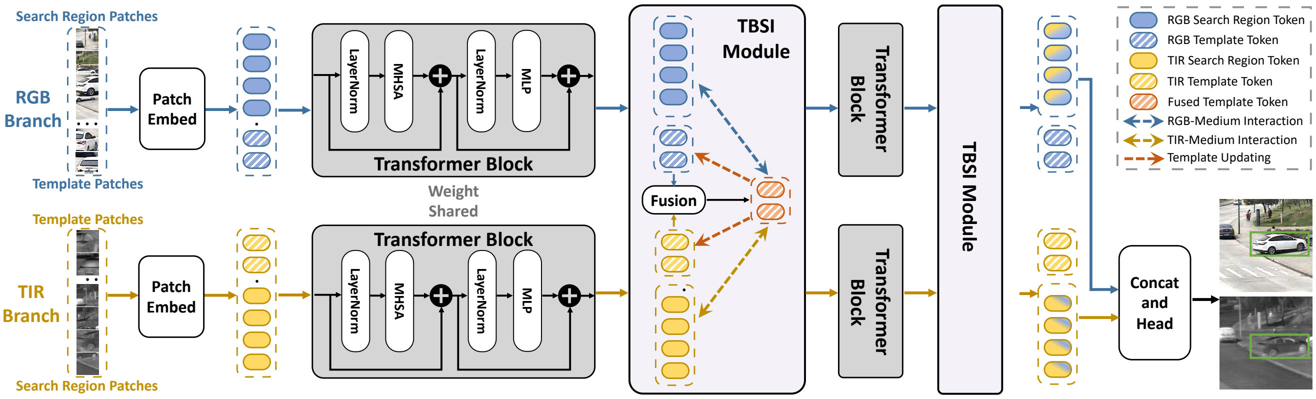

Figure 2. The overall framework of our method. RGB and TIR image patches are embedded as tokens and fed into Transformer blocks for joint feature extraction and intra-modal search-template matching. In our proposed TBSI module, bidirectional RGB and TIR search region interaction are bridged by the fused template, which serves as a medium to gather and distribute target-relevant contexts to enhance RGB and TIR search region features. The two original templates are also updated with the enriched contexts of the template medium. Finally, RGB and TIR search region features are concatenated and fed into the tracking head to predict the target's current state.

extreme imaging conditions. Therefore, thermal infrared data has become a widely-adopted information source [14, 16, 20, 27, 28, 37, 39, 44, 46, 47, 49] for mutual complement with visible light data to enhance the robustness of trackers. To deploy the complementarity of features in all layers, Zhu *et al.* [49] propose a recursive strategy to densely aggregate these features that yield robust representations of target objects in each modality. mfDiMP [44] embeds the multimodal feature concatenation process into the framework of a strong tracker DiMP [4] for RGB-T tracking. Zhang *et al.* [46] propose a late fusion method to obtain both global and local weights for multimodal fusion, taking both appearance and motion information into account and dynamically switching between appearance and motion cues. SiamCDA [47] presents a complementarity-aware multimodal feature fusion module to enhance the discriminability of the fused features by first reducing the modality differences between unimodal features and then fusing them. To make full use of training data and cope with different challenges (*e.g.*, illumination variation, occlusion, thermal crossover, fast motion, *etc*), CAT [20] mines modal-shared information and modal-specific information with different challenges, and all challenge-aware branches are embedded into the backbone to form more discriminative target representations. APFNet [39] designs an attribute-based aggregation fusion model to adaptively aggregate all attribute-specific fused features and proposes an attribute-based progressive fusion network to disentangle the fusion process via the challenge attributes and increase the fusion capacity. However, previous RGB-T tracking methods conduct fusion between RGB and TIR search frames or candidate RoIs, which inevitably introduces background noises and restricts the multimodal complementary effect, yielding coarse and insufficient cross-modal interaction. To alleviate these limitations, we propose a TBSI module that exploits the target templates as the medium to bridge the cross-modal interaction between RGB and TIR search regions, achieving adaptive and precise information enhancement and complement using the target-relevant object and environment contexts.

3. Method

The overall framework of our method is shown in Figure 2. The input RGB and TIR search region and template images are first split and flattened as sequences of patches (tokens), then fed into a series of shared Transformer blocks for joint feature extraction and search-template matching within each modality. Our proposed TBSI module is inserted between Transformer blocks to bridge the cross-modal search region interaction with the fused template tokens as the medium for target-relevant context gathering and distribution. Finally, the tracking head takes the concatenated RGB and TIR search region features from the backbone as input to predict the target's current state.

3.1. Multimodal ViT for RGB-T Tracking

Considering the powerful ability of ViT to capture long-range dependencies, we follow recent SOT methods [7, 24, 42] to extend ViT as a multimodal backbone of our base tracker for jointly extracting features and performing search-template matching within the Transformer blocks. Let $\boldsymbol{I}_r^x, \boldsymbol{I}_t^x \in \mathbb{R}^{H_x \times W_x \times 3}$ denote the RGB and TIR search region images and $\boldsymbol{I}_r^z, \boldsymbol{I}_t^z \in \mathbb{R}^{H_z \times W_z \times 3}$ denote the RGB and TIR template images respectively, where different modalities have the same image resolutions. We first spatially partition these images into patches with the size of $P \times P$ and flatten them as four sequences of patches

$\boldsymbol{P}^x_r, \boldsymbol{P}^x_t \in \mathbb{R}^{N_x\times(3P^2)}$ and $\boldsymbol{P}^z_r, \boldsymbol{P}^z_t \in \mathbb{R}^{N_z\times(3P^2)}$, where $N_x = H_xW_x/P^2$, $N_z = H_zW_z/P^2$ denote the number of search region patches and template patches. Then, the patch embedding layers with linear projections are applied to these sequences to obtain the initial features of RGB and TIR search regions and templates as follows:

$$\begin{aligned}\boldsymbol{X}^0_r = \boldsymbol{P}^x_r\boldsymbol{W}^0_r, \boldsymbol{Z}^0_r = \boldsymbol{P}^z_r\boldsymbol{W}^0_r,\\ \boldsymbol{X}^0_t = \boldsymbol{P}^x_t\boldsymbol{W}^0_t, \boldsymbol{Z}^0_t = \boldsymbol{P}^z_t\boldsymbol{W}^0_t,\end{aligned} \tag{1}$$

where $\boldsymbol{W}^0_r, \boldsymbol{W}^0_t \in \mathbb{R}^{(3P^2)\times C}$ are learnable parameters of linear projections, $\boldsymbol{X}^0_r, \boldsymbol{X}^0_t \in \mathbb{R}^{N_x\times C}$ denote the embedded features of RGB and TIR search region patches (referred as *tokens* in the following text), $\boldsymbol{Z}^0_r, \boldsymbol{Z}^0_t \in \mathbb{R}^{N_z\times C}$ denote the embedded features of RGB and TIR template tokens, C is the number of feature channels. Following [12], we also add the learnable positional encoding matrices $\boldsymbol{E}_x \in \mathbb{R}^{N_x\times C}$ and $\boldsymbol{E}_z \in \mathbb{R}^{N_z\times C}$ with the token features $\boldsymbol{X}^0_r, \boldsymbol{X}^0_t$ and $\boldsymbol{Z}^0_r, \boldsymbol{Z}^0_t$ to provide positional prior information. Note that we share the same positional encoding matrices between RGB and TIR modalities since the raw frames are carefully aligned by dataset constructors.

Afterward, the RGB and TIR tokens are concatenated as $\boldsymbol{H}^0_r = [\boldsymbol{X}^0_r; \boldsymbol{Z}^0_r] \in \mathbb{R}^{(N_x+N_z)\times C}$ and $\boldsymbol{H}^0_t = [\boldsymbol{X}^0_t; \boldsymbol{Z}^0_t] \in \mathbb{R}^{(N_x+N_z)\times C}$ to separately fed into a series of Transformer [36] blocks for multimodal joint feature extraction and search-template matching. Since the operations on RGB and TIR modalities are similar, here we take the RGB tokens as an example to elaborate on how the ViT backbone works for the tracking process. The subscript r and superscript 0 are omitted for simplicity. In each Transformer block, three projections are first performed on $\boldsymbol{H}$ to obtain the query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$. Then, matrix multiplications are conducted to aggregate features, in which the attention weights are generated as follows:

$$\begin{aligned}\boldsymbol{A} &= \mathrm{Softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{C}}) = \mathrm{Softmax}(\frac{[\boldsymbol{X}_q; \boldsymbol{Z}_q][\boldsymbol{X}_k; \boldsymbol{Z}_k]^{\mathrm{T}}}{\sqrt{C}})\\ &= \mathrm{Softmax}(\frac{[\boldsymbol{X}_q\boldsymbol{X}_k^{\mathrm{T}}, \boldsymbol{X}_q\boldsymbol{Z}_k^{\mathrm{T}}; \boldsymbol{Z}_q\boldsymbol{X}_k^{\mathrm{T}}, \boldsymbol{Z}_q\boldsymbol{Z}_k^{\mathrm{T}}]}{\sqrt{C}}).\end{aligned} \tag{2}$$

From the above formulation, we can observe that search region tokens and template tokens simultaneously refine their own features and aggregate features from each other based on the joint attention weights. Through successive Transformer blocks, features of search region and template tokens are gradually extracted and the matching relationships between them are captured as well to locate the target object in each modality respectively. The parameters of the Transformer blocks are shared between RGB and TIR tokens to avoid redundancy.

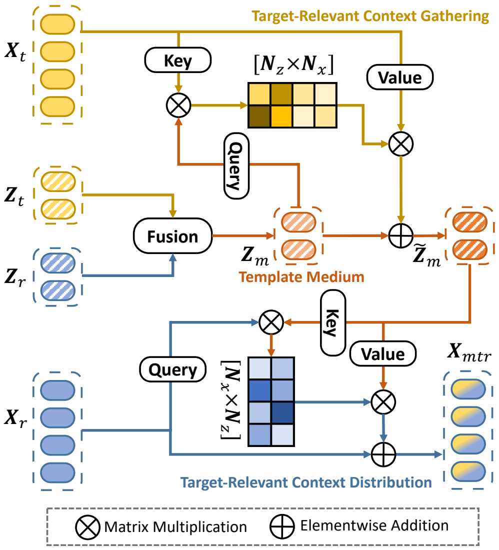

Figure 3. Conceptual illustration of the *TIR→Medium→RGB* search region interaction process in our TBSI module. Interaction in the reverse direction is conducted similarly. We omit template updating and operations like LN and MLP for clear presentation.

3.2. Template-Bridged Search Region Interaction

Our proposed TBSI module aims to bridge the cross-modal interaction between RGB and TIR search regions with the templates as the medium, where target-relevant object and environment contexts are mutually complemented to each modality. We insert TBSI module between the Transformer blocks of ViT backbone multiple times for joint feature extraction, search-template matching, and cross-modal interaction. We take the i-th Transformer block to elaborate the bridging process of our TBSI module. Let $\boldsymbol{X}^i_r$, $\boldsymbol{Z}^i_r$, $\boldsymbol{X}^i_t$, $\boldsymbol{Z}^i_t$ denote the search region and template token features of RGB and TIR modalities respectively, and we omit the superscript i for simplicity. Figure 3 conceptually illustrates the *TIR→Medium→RGB* search region interaction with detailed operations omitted.

Template Fusion. We first fuse the features of two templates to obtain a sequence of multimodal template tokens as the bridging medium $\boldsymbol{Z}_m \in \mathbb{R}^{N_z\times C}$:

$$\boldsymbol{Z}_m = [\boldsymbol{Z}_r; \boldsymbol{Z}_t]\boldsymbol{W}_m, \tag{3}$$

where $\boldsymbol{W}_m \in \mathbb{R}^{2C\times C}$ is the parameter of a linear layer. $\boldsymbol{Z}_m$ contains the target object clues of both RGB and TIR modalities, thus serving as an appropriate medium to excavate target-relevant contexts in a bidirectional manner.

Bidirectional Template-Bridged Interaction. Since cross-attention [36] is a common and widely-adopted practice for information aggregation, we also apply it as cross-modal attention between TIR search region tokens $\boldsymbol{X}_t$ and template medium tokens $\boldsymbol{Z}_m$ to first gather target-relevant

TIR context as follows:

$$\boldsymbol{D}_t = \mathrm{Softmax}(\frac{(\boldsymbol{Z}_m\boldsymbol{W}_q^1)(\boldsymbol{X}_t\boldsymbol{W}_k^1)^{\mathrm{T}}}{\sqrt{C}})(\boldsymbol{X}_t\boldsymbol{W}_v^1), \quad (4)$$

where $\boldsymbol{W}_q^1$, $\boldsymbol{W}_k^1$, $\boldsymbol{W}_v^1$ denote parameters of the query, key, value projection layers. Then, the target-relevant TIR context $\boldsymbol{D}_t$ is refined and integrated with $\boldsymbol{Z}_m$ to enrich the medium with the required TIR information:

$$\begin{aligned} \boldsymbol{Z}'_m &= \mathrm{LN}(\boldsymbol{Z}_m + \boldsymbol{D}_t), \\ \tilde{\boldsymbol{Z}}_m &= \mathrm{LN}(\boldsymbol{Z}'_m + \mathrm{MLP}(\boldsymbol{Z}'_m)), \end{aligned} \quad (5)$$

where LN and MLP represent LayerNorm [2] and Multilayer Perceptron. Afterward, the gathered target-relevant context from TIR search region tokens, along with the multimodal target prior information contained in the bridging template medium, are further distributed to the RGB search region tokens adaptively for enhancing RGB target features. Concretely, RGB search region tokens serve as query and the template medium serves as key and value to distribute information via similar cross-modal attention:

$$\boldsymbol{D}_{mt} = \mathrm{Softmax}(\frac{(\boldsymbol{X}_r\boldsymbol{W}_q^2)(\tilde{\boldsymbol{Z}}_m\boldsymbol{W}_k^2)^{\mathrm{T}}}{\sqrt{C}})(\tilde{\boldsymbol{Z}}_m\boldsymbol{W}_v^2). \quad (6)$$

Then, $\boldsymbol{D}_{mt}$ is further refined and integrated with $\boldsymbol{X}_r$ to enhance the corresponding target-relevant RGB search region tokens as follows:

$$\begin{aligned} \boldsymbol{X}'_r &= \mathrm{LN}(\boldsymbol{X}_r + \boldsymbol{D}_{mt}), \\ \boldsymbol{X}_{mtr} &= \mathrm{LN}(\boldsymbol{X}'_r + \mathrm{MLP}(\boldsymbol{X}'_r)). \end{aligned} \quad (7)$$

In the reverse direction, target-relevant RGB context is similarly gathered from the RGB search region tokens $\boldsymbol{X}_r$ to the enriched template medium $\tilde{\boldsymbol{Z}}_m$, then distributed to the TIR search region tokens $\boldsymbol{X}_r$ along with the target prior information to obtain the enhanced TIR search features $\boldsymbol{X}_{mrt}$.

Template Updating. Instead of only enhancing the search region tokens with template-bridged target-relevant contexts, we also transfer features of the template medium back to the original RGB and TIR templates to update them with enriched multimodal target information. For the architecture consistency in our TBSI module, we also adopt the cross-modal attention mechanism to implement this information transfer process. Let $\hat{\boldsymbol{Z}}_m$ denote the multimodal template medium after the bidirectional search region interaction, we use $\hat{\boldsymbol{Z}}_m$ as key and value, and original templates $\boldsymbol{Z}_r$ and $\boldsymbol{Z}_t$ as queries to perform the information transfer. The outputs $\boldsymbol{Z}'_r$ and $\boldsymbol{Z}'_t$ are then refined and integrated with $\boldsymbol{Z}_r$ and $\boldsymbol{Z}_t$ using similar LN and MLP layers. The updated template tokens $\boldsymbol{Z}_{mr}$ and $\boldsymbol{Z}_{mt}$ along with interacted search region tokens $\boldsymbol{X}_{mtr}$ and $\boldsymbol{X}_{mrt}$ serve as the input data of the next $(i+1)$-th Transformer block in our ViT backbone.

4. Experiments

4.1. Datasets and Evaluation Metrics

We conduct experiments on three RGB-T tracking benchmarks including LasHeR [22], RGBT234 [19], and RGBT210 [23]. Following prior works [22, 39, 44], we utilize three widely-adopted metrics to evaluate our method. Precision rate measures the percentage of frames whose distance between the predicted position and the ground-truth is less than a certain threshold. Considering the sensitivity to target size, normalized precision rate is calculated by normalizing the precision rate on the size of the ground truth bounding box. Success rate is the ratio of successfully tracked frames whose IoU overlaps are larger than thresholds. The area under the curve plotted by different thresholds measures the representative success score.

4.2. Implementation Details

Our model is implemented using PyTorch [33] and experiments are conducted on four NVIDIA A100 GPUs. The total training batch size is 128 image pairs. We train our model for 15 epochs on LasHeR dataset with 60k image pairs per epoch and directly evaluate our model on RGBT234 and RGBT210 datasets without further finetuning. The learning rate is set as 4e-5 for the backbone and 4e-4 for other parameters, which is decayed by 10× after 10 epochs. We adopt AdamW [26] as the optimizer with 1e-4 weight decay. The search regions are resized to 256 × 256 and templates are resized to 128 × 128. Our TBSI module is inserted in the 4-th, 7-th, and 10-th layers of the ViT backbone. As a common practice, the threshold is set to 20 pixels to compute the representative precision score.

4.3. Comparison with State-of-the-art Methods

We compare our method with previous state-of-the-art RGB-T tracking methods on three benchmarks including LasHeR [22], RGBT234 [19], and RGBT210 [23]. As shown in Table 1, our methods with different ViT backbones consistently outperform previous RGB-T trackers on all metrics. Since mfDiMP [44] exploits pretraining on the joint splits of COCO [25], LaSOT [13], GOT-10k [15], and TrackingNet [31] as SOT methods do (referred as SOT pretraining), we also adopt this setting for a fair comparison. Our ViT-Base model with SOT pretraining achieves significant performance improvements over previous methods and our ViT-Tiny model with ImageNet [11] pretraining still outperforms mfDiMP, demonstrating the effectiveness of our template-bridged search region interaction scheme. However, previous MDNet-based [32] methods with VGG-M [34] backbones could not benefit from the powerful pretrained SOT models since the number of binary classification branches in MDNet equals the number of training sequences to conduct multi-domain learning. Therefore,

	Method	Backbone	Pretraining	Precision	NormPrec	Success	FPS
Online	DAPNet [49]	VGG-M	ImageNet	43.1	38.3	31.4	-
	FANet [50]	VGG-M	ImageNet	44.1	38.4	30.9	-
	DAFNet [14]	VGG-M	ImageNet	44.8	39.0	31.1	20.5
	CAT [20]	VGG-M	ImageNet	45.0	39.5	31.4	-
	MANet [21]	VGG-M	ImageNet	45.5	-	32.6	2.1
	MANet++ [27]	VGG-M	ImageNet	46.7	40.4	31.4	-
	MaCNet [43]	VGG-M	ImageNet	48.2	42.0	35.0	1.6
	DMCNet [28]	VGG-M	ImageNet	49.0	43.1	35.5	-
	APFNet [39]	VGG-M	ImageNet	50.0	43.9	36.2	1.9
	mfDiMP [44]	ResNet-50	SOT	59.9	-	46.7	34.6
Offline	TBSI	ViT-Tiny	ImageNet	61.7	57.8	48.9	**40.3**
	TBSI	ViT-Small	ImageNet	62.4	58.6	49.4	39.1
	TBSI	ViT-Base	ImageNet	63.8	60.2	50.6	36.2
	TBSI	ViT-Base	SOT	**69.2**	**65.7**	**55.6**	36.2

Table 1. Comparison with state-of-the-art methods on LasHeR testing set. "SOT" denotes pretraining on the joint splits of COCO, LaSOT, GOT-10k, and TrackingNet, which is a common practice for training SOT methods. We also adopt this setting for a fair comparison. We only report the inference speeds of previous methods whose codes are available.

	Method	Precision	Success
Online	MDNet+RGBT [32]	72.2	49.5
	MaCNet [43]	76.4	53.2
	DAPNet [49]	76.6	53.7
	MANet [21]	77.7	53.9
	HDINet [30]	78.3	55.9
	FANet [50]	78.7	55.3
	JMMAC [46]	79.0	57.3
	M5L [35]	79.5	54.2
	MANet++ [27]	79.5	55.9
	DAFNet [14]	79.6	54.4
	CAT [20]	80.4	56.1
	ADRNet [45]	80.7	57.0
	CMPP [37]	82.3	57.5
	APFNet [39]	82.7	57.9
	DMCNet [28]	83.9	59.3
	mfDiMP [44]	84.2	59.1
Offline	SiamCDA [47]	76.0	56.9
	SiamIVFN [16]	81.1	63.2
	TBSI	**87.1**	**63.7**

Table 2. Comparison with state-of-the-art methods on RGBT234 dataset. Our method outperforms both online and offline ones.

	Method	Precision	Success
Online	TFNet [51]	77.7	52.9
	CAT [20]	79.2	53.3
	DMCNet [28]	79.7	55.5
	mfDiMP* [44]	84.9	59.3
Offline	DSiamMFT [48]	64.2	43.2
	TBSI	**85.3**	**62.5**

Table 3. Comparison with state-of-the-art methods on RGBT210 dataset. * means results are reproduced by us.

the scalability of MDNet-based methods on modern large-scale SOT datasets is severely limited, thus causing their performance to lag behind. In terms of efficiency, we report the FPS values of our method and previous ones on the same machine with an NVIDIA RTX 3080Ti GPU. Previous MDNet-based methods tend to have slower inference speeds since they all rely on heavy online updating operations to finetune the testing sequences, while our offline-learned models possess high efficiency with real-time inference speeds. In Table 2 and 3, our performance superiority on RGBT234 and RGBT210 datasets also demonstrates the generalization ability of our method on small datasets.

4.4. Ablation Studies

Component Analysis. In Table 4, we conduct ablation studies on the LasHeR dataset to evaluate different designs of our TBSI module with ImageNet pretraining.

RGB Baseline denotes feeding only RGB image pairs as input to the ViT-Base backbone and tracking head to perform single-modal tracking.

RGB-T Baseline denotes both RGB and TIR image pairs are embedded as patches and fed into the shared ViT-Base backbone for joint feature extraction and search-template matching without cross-modal interaction. RGB and TIR search region features outputted from the backbone are concatenated to serve as the input of the tracking head. We can observe the simple RGB-T baseline without complicated cross-modal interaction yields better performance than the RGB-only baseline, showing that introducing the TIR modality is beneficial to the tracking process.

w/o Template Bridging denotes directly conducting bidirectional cross-modal search region interaction between

Method	Precision	NormPrec	Success
RGB Baseline	50.1	45.4	40.1
RGB-T Baseline	53.5	49.1	42.5
w/o Template Bridging	59.6	55.9	47.4
w/o RGB→TM→TIR	58.7	55.1	46.6
w/o Template Updating	62.7	58.9	49.7
Full Model (TBSI)	**63.8**	**60.2**	**50.6**

Table 4. Ablation studies of our proposed TBSI module. "TM" denotes the template medium for bridging interaction.

RGB and TIR modality by the cross-attention mechanism without using the fused template as a medium. Compared with the RGB-T baseline, large performance gains are witnessed to show the importance of cross-modal interaction. Compared with our full model with TBSI module, we can also observe that template bridging is able to further notably improve the performance of cross-modal interaction by enhancing search region features with target-relevant contexts gathered by the template medium. The background noises of other distractors in the search region are also reduced through template bridging to highlight the target area. These results well demonstrate the effectiveness of our proposed template-bridged search region interaction scheme.

w/o RGB→TM→TIR denotes the uni-directional version of our TBSI module where only the target-relevant TIR contexts are gathered by the template medium to enhance the RGB search region features. This experiment shows that uni-directional interaction can outperform the baseline but bi-directional interaction is able to further boost the performance by RGB-TIR mutual enhancement.

w/o Template Updating denotes removing the template updating step after bidirectional template-bridged search region interaction. The performance drops compared with our full model (TBSI), which indicates that updating original templates with enriched multimodal contexts from the template medium also benefits the tracking process.

Layers			Precision	NormPrec	Success
4	7	10			
			53.5	49.1	42.5
✓			60.5	56.9	47.8
✓	✓		62.7	59.2	49.8
✓	✓	✓	**63.8**	**60.2**	**50.6**

Table 5. Inserting layers of the proposed TBSI module.

Inserting Layers of TBSI module. We evaluate different inserting layers of our proposed TBSI module and summarize the experimental results in Table 5. It can be observed that inserting the TBSI module in the 4-th layer of ViT backbone yields a large performance boost against the RGB-T baseline model, which shows the importance of proper cross-modal interaction between search regions. When inserting the TBSI module into the 7-th and 10-th layers of ViT backbone, tracking performance is further elevated by interacting with deep semantic features. Marginal improvements are found by inserting more layers so we adopt the setting of three TBSI modules as our final model.

	APFNet† [39]	CMPP [37]	mfDiMP* [44]	TBSI
NO	93.4/66.4	95.6/67.8	**96.2**/69.4	96.1/**72.8**
PO	85.0/58.7	85.5/60.1	86.6/60.9	**88.7/64.7**
HO	72.9/49.0	73.2/50.3	76.1/53.2	**81.5/58.6**
LI	82.3/54.4	86.2/58.4	84.2/58.0	**89.2/63.6**
LR	82.9/54.8	**86.5**/57.1	82.1/53.0	85.1/**60.0**
TC	82.1/57.3	83.5/58.3	84.8/58.9	**85.8/63.2**
DEF	77.1/54.6	75.0/54.1	81.5/60.2	**84.1/63.7**
FM	78.2/49.2	78.6/50.8	77.3/54.8	**81.4/58.7**
SV	82.1/56.5	81.5/57.2	87.1/63.7	**89.9/66.8**
MB	72.8/53.0	75.4/54.1	80.1/58.0	**88.1/64.9**
CM	76.3/54.5	75.6/54.1	84.0/60.3	**88.0/65.0**
BC	80.6/52.4	83.2/53.8	82.8/53.7	**83.4/57.8**

Table 6. Attribute-based Precision/Success scores on RGBT234 dataset. † denotes that the values are obtained by evaluating the authors' released raw tracking results. * means results are reproduced by us since raw results are unavailable.

4.5. Analysis and Visualization

Attribute-Based Performance. We analyze the performance of our method in various scenarios by evaluating it on different attributes of RGBT234 dataset, including no occlusion (NO), partial occlusion (PO), heavy occlusion (HO), low illumination (LI), low resolution (LR), thermal crossover (TC), deformation (DEF), fast motion (FM), scale variation (SV), motion blur (MB), camera moving (CM) and background clutter (BC). Table 6 summarizes the experimental results. Our method outperforms previous state-of-the-art trackers on most attributes. Particularly, in the scenarios of heavy occlusion, deformation, scale variation, motion blur, and camera moving, the target object is drastically deformed or even temporarily invisible, making previous methods less robust. Benefiting from the information enhancement and complement brought by the template bridging in our TBSI module and the long-range dependency modeling ability of Transformer blocks, our method could interact RGB and TIR features more comprehensively and yield better performance.

Qualitative Comparison. As shown in Figure 4, we conduct a qualitative comparison between our method and eight other deep RGB-T trackers. Four representative sequences which include various challenges such as occlusion, high illumination, deformation, scale variation, fast movement, *etc*, are selected from the LasHeR dataset to compare different methods' performances. For example, the target man in the second sequence walks from dark areas

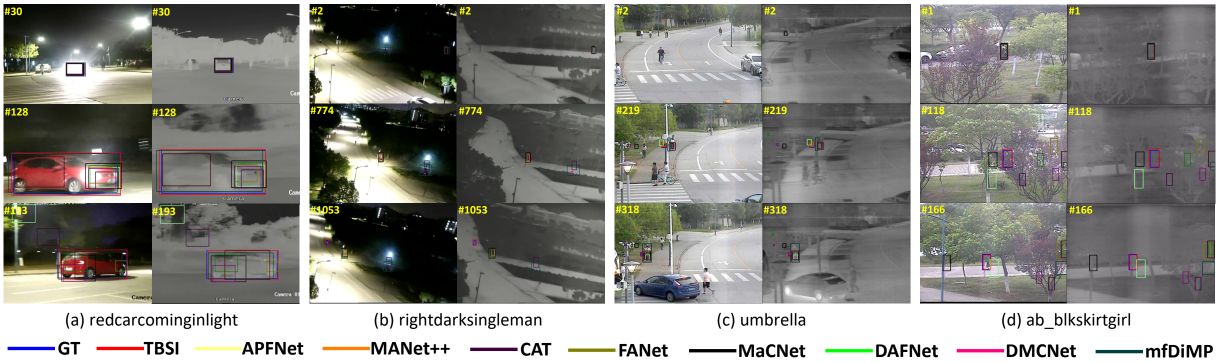
(a) redcarcominginlight
(b) rightdarksingleman
(c) umbrella
(d) ab_blkskirtgirl

Figure 4. Qualitative comparison between our method and other RGB-T trackers on four representative sequences from LasHeR dataset.

to light areas, where our method can sufficiently leverage the mutual enhancement and complement ability of RGB and TIR modalities to track the target stably. In other sequences, our method also well tackles some common challenges like scale variation, fast movement, and occlusion. These results indicate our proposed ViT tracker with TBSI module embodies a stronger discriminative ability, which is also verified in Table 6 with more attributes (challenges).

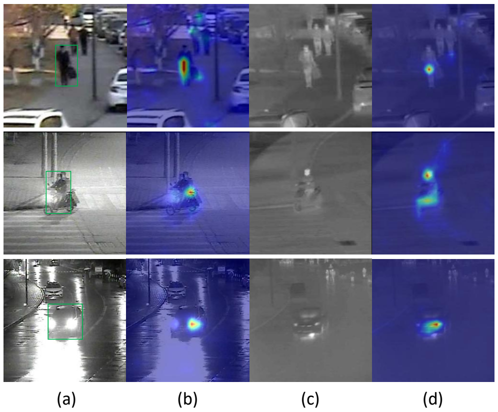

Figure 5. Visualization of attention maps between template medium tokens and search region tokens in our TBSI module. (a) RGB search region. (b) RGB attention map. (c) TIR search region. (d) TIR attention map.

Visualization of Attention Map. To understand how the template medium bridges the search region interaction between RGB and TIR modalities, we visualize the cross-attention maps between template medium tokens and search region tokens in Figure 5. We can observe that in different challenging scenarios such as high illumination, rainy weather, and dark night, the template medium can correctly attend to the target areas in both RGB and TIR search regions. Through concentrated attentions of the template medium in our TBSI module, target-relevant RGB and TIR contexts are gathered and distributed to enhance search region features of the other modality, meanwhile reducing the background noisy information to form a more adaptive and precise cross-modal interaction process.

5. Conclusion

In this paper, we explore a more effective cross-modal interaction scheme for RGB-T tracking. Most previous methods conduct simple concatenation of search frame features or various fusion operations on pairs of local candidate boxes, yielding either coarse or insufficient cross-modal interaction. To alleviate these limitations, we propose a TBSI module that bridges the RGB and TIR search region interaction using the fused template as a medium so that target-relevant contexts can be excavated to enhance search region features of both modalities, meanwhile reducing the background noises. The original templates are also updated with enriched multimodal contexts from the template medium. Extensive experiments on three RGB-T benchmarks show our method achieves state-of-the-art performances.

Limitation. Our TBSI module is implemented with the cross-modal attention mechanism since it is a common and wide-adopted practice in multimodal learning to aggregate relevant features. Thus, we also utilize it to verify the feasibility of our template-bridged interaction scheme. Though proven effective, cross-attention may not be optimally tailored for RGB-T tracking due to various task-specific challenges a tracker could encounter. In the future, we plan to explore new feature aggregation approaches in TBSI.

Acknowledgements. This research is supported in part by National Key R&D Program of China (2022ZD0115502), National Natural Science Foundation of China (Grant No. 62122010), Zhejiang Provincial Natural Science Foundation of China under Grant No. LDT23F02022F02, Key Research and Development Program of Zhejiang Province under Grant 2022C01082, and The Open Project of Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University, No. MMC202109.

References

[1] Thiemo Alldieck, Chris H Bahnsen, and Thomas B Moeslund. Context-aware fusion of rgb and thermal imagery for traffic monitoring. *Sensors*, 2016. 1

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[3] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *ECCV*, 2016. 2

[4] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *ICCV*, 2019. 2, 3

[5] Long Chen, Libo Sun, Teng Yang, Lei Fan, Kai Huang, and Zhe Xuanyuan. Rgb-t slam: A flexible slam framework by combining appearance and thermal information. In *ICRA*, 2017. 1

[6] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *CVPR*, 2021. 2

[7] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *CVPR*, 2022. 2, 3

[8] Xuerui Dai, Xue Yuan, and Xueye Wei. Tirnet: Object detection in thermal infrared images for autonomous driving. *Applied Intelligence*, 2021. 1

[9] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *CVPR*, 2019. 2

[10] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *CVPR*, 2017. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[12] 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. In *ICLR*, 2021. 2, 4

[13] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *CVPR*, 2019. 5

[14] Yuan Gao, Chenglong Li, Yabin Zhu, Jin Tang, Tao He, and Futian Wang. Deep adaptive fusion network for high performance rgbt tracking. In *ICCVW*, 2019. 2, 3, 6

[15] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *TPAMI*, 2019. 5

[16] Peng Jingchao, Zhao Haitao, Hu Zhengwei, Zhuang Yi, and Wang Bofan. Siamese infrared and visible light fusion network for rgb-t tracking. *arXiv preprint arXiv:2103.07302*, 2021. 2, 3, 6

[17] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *CVPR*, 2019. 2

[18] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, 2018. 2

[19] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: Benchmark and baseline. *Pattern Recognition*, 2019. 5

[20] Chenglong Li, Lei Liu, Andong Lu, Qing Ji, and Jin Tang. Challenge-aware rgbt tracking. In *ECCV*, 2020. 3, 6

[21] Chenglong Li, Andong Lu, Aihua Zheng, Zhengzheng Tu, and Jin Tang. Multi-adapter rgbt tracking. In *ICCVW*, 2019. 6

[22] 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. *TIP*, 2021. 5

[23] Chenglong Li, Nan Zhao, Yijuan Lu, Chengli Zhu, and Jin Tang. Weighted sparse representation regularized graph learning for rgb-t object tracking. In *ACM MM*, 2017. 5

[24] Liting Lin, Heng Fan, Yong Xu, and Haibin Ling. Swintrack: A simple and strong baseline for transformer tracking. In *NeurIPS*, 2022. 2, 3

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[27] Andong Lu, Chenglong Li, Yuqing Yan, Jin Tang, and Bin Luo. Rgbt tracking via multi-adapter network with hierarchical divergence loss. *TIP*, 2021. 2, 3, 6

[28] Andong Lu, Cun Qian, Chenglong Li, Jin Tang, and Liang Wang. Duality-gated mutual condition network for rgbt tracking. *TNNLS*, 2022. 2, 3, 6

[29] Alan Lukezic, Tomas Vojir, Luka ˇCehovin Zajc, Jiri Matas, and Matej Kristan. Discriminative correlation filter with channel and spatial reliability. In *CVPR*, 2017. 2

[30] Jiatian Mei, Dongming Zhou, Jinde Cao, Rencan Nie, and Yanbu Guo. Hdinet: hierarchical dual-sensor interaction network for rgbt tracking. *IEEE Sensors Journal*, 2021. 6

[31] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *ECCV*, 2018. 5

[32] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *CVPR*, 2016. 2, 5, 6

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019. 5

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 5

[35] Zhengzheng Tu, Chun Lin, Wei Zhao, Chenglong Li, and Jin Tang. M 5 l: Multi-modal multi-margin metric learning for rgbt tracking. *TIP*, 2021. 6

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 4

[37] Chaoqun Wang, Chunyan Xu, Zhen Cui, Ling Zhou, Tong Zhang, Xiaoya Zhang, and Jian Yang. Cross-modal pattern-propagation for rgb-t tracking. In *CVPR*, 2020. 2, 3, 6, 7

[38] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *CVPR*, 2021. 2

[39] Yun Xiao, Mengmeng Yang, Chenglong Li, Lei Liu, and Jin Tang. Attribute-based progressive fusion network for rgbt tracking. In *AAAI*, 2022. 2, 3, 5, 6, 7

[40] Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In *AAAI*, 2020. 2

[41] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *ICCV*, 2021. 2

[42] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, 2022. 2, 3

[43] Hui Zhang, Lei Zhang, Li Zhuo, and Jing Zhang. Object tracking in rgb-t videos using modal-aware attention network and competitive learning. *Sensors*, 2020. 6

[44] 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 *ICCVW*, 2019. 2, 3, 5, 6, 7

[45] Pengyu Zhang, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Learning adaptive attribute-driven representation for real-time rgb-t tracking. *IJCV*, 2021. 6

[46] Pengyu Zhang, Jie Zhao, Chunjuan Bo, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Jointly modeling motion and appearance cues for robust rgb-t tracking. *TIP*, 2021. 3, 6

[47] Tianlu Zhang, Xueru Liu, Qiang Zhang, and Jungong Han. Siamcda: Complementarity-and distractor-aware rgb-t tracking based on siamese network. *TCSVT*, 2021. 3, 6

[48] Xingchen Zhang, Ping Ye, Shengyun Peng, Jun Liu, and Gang Xiao. Dsiammft: An rgb-t fusion tracking method via dynamic siamese networks using multi-layer feature fusion. *Signal Processing: Image Communication*, 2020. 6

[49] Yabin Zhu, Chenglong Li, Bin Luo, Jin Tang, and Xiao Wang. Dense feature aggregation and pruning for rgbt tracking. In *ACM MM*, 2019. 2, 3, 6

[50] Yabin Zhu, Chenglong Li, Jin Tang, and Bin Luo. Quality-aware feature aggregation network for robust rgbt tracking. *TIV*, 2020. 6

[51] Yabin Zhu, Chenglong Li, Jin Tang, Bin Luo, and Liang Wang. Rgbt tracking by trident fusion network. *TCSVT*, 2021. 6

[52] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *ECCV*, 2018. 2