Unified Single-Stage Transformer Network for Efficient RGB-T Tracking

Jianqiang Xia[1,2] , **Dianxi Shi**[1*] , **Ke Song**[3] , **Linna Song**[4] , **Xiaolei Wang**[1] , **Songchang Jin**[1] , **Chenran Zhao**[4] , **Yu Cheng**[5] , **Lei Jin**[6] , **Zheng Zhu**[7] , **Jianan Li**[8] , **Gang Wang**[9] , **Junliang Xing**[7] and **Jian Zhao**[10,11]

[1]Intelligent Game and Decision Lab, China
[2]Tianjin Artificial Intelligence Innovation Center, China
[3]School of Control Science and Engineering, Shandong University, China
[4]College of Computer, National University of Defense Technology, China
[5]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[6]Electronic Engineering Institute, Beijing University of Posts and Telecommunications, China
[7]Department of Computer Science and Technology, Tsinghua University, China
[8]School of Optoelectronics, Beijing Institute of Technology, China
[9]Beijing Institute of Basic Medical Sciences, China
[10]EVOL Lab, Institute of AI (TeleAI), China Telecom, China
[11]School of Artificial Intelligence, Optics and Electronics, Northwestern Polytechnical University, China
jianqiang.xia@foxmail.com, dxshi@nudt.edu.cn, zhengzhu@ieee.org, zhaoj90@chinatelecom.cn

Abstract

Most existing RGB-T tracking networks extract modality features in a separate manner, which lacks interaction and mutual guidance between modalities. This limits the network's ability to adapt to the diverse dual-modality appearances of targets and the dynamic relationships between the modalities. Additionally, the three-stage fusion tracking paradigm followed by these networks significantly restricts the tracking speed. To overcome these problems, we propose a unified single-stage Transformer RGB-T tracking network, namely USTrack, which unifies the above three stages into a single ViT (Vision Transformer) backbone through joint feature extraction, fusion and relation modeling. With this structure, the network can not only extract the fusion features of templates and search regions under the interaction of modalities, but also significantly improve tracking speed through the single-stage fusion tracking paradigm. Furthermore, we introduce a novel feature selection mechanism based on modality reliability to mitigate the influence of invalid modalities for final prediction. Extensive experiments on three mainstream RGB-T tracking benchmarks show that our method achieves the new state-of-the-art while achieving the fastest tracking speed of 84.2FPS. Code is available at https://github.com/xiajianqiang/USTrack.

1 Introduction

Visible-Thermal (RGB-T) tracking greatly expands the application scenarios of visual object tracking by using both RGB

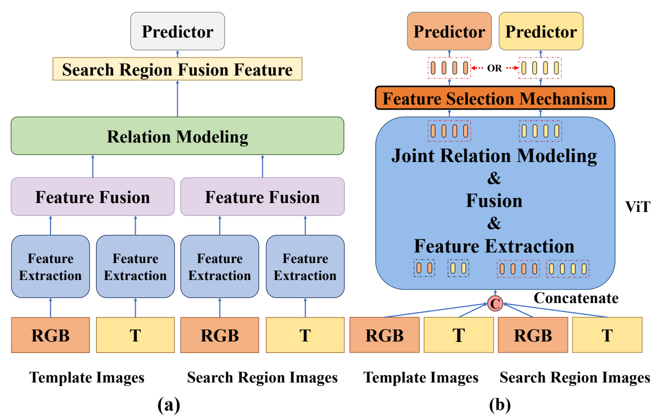


Figure 1: (a) The three-stage tracking paradigm followed by existing RGB-T tracking networks. Tracking networks performs modalities feature extraction, fusion, and relation modeling operations in three stages, respectively. (b) Our RGB-T tracking framework performs joint feature extraction, fusion and relation modeling by unifying the above three parts into a single ViT backbone. In addition, we designed a feature selection mechanism to help us discard features from invalid modalities.

and thermal information, significantly improving the tracking performance under challenging conditions such as illumination variation, occlusion, and extreme weather. Therefore, RGB-T tracking has become a research focus in recent years.

Most existing RGB-T tracking methods follow a three-stage fusion tracking paradigm which can be shown in Fig. 1(a). These networks separately employ two shallow CNN [He *et al.*, 2015] or Transformer [Dosovitskiy *et al.*, 2020] subnetworks to extract RGB and thermal features from the template and search region. These features are then fused

*Corresponding author: Dianxi Shi, dxshi@nudt.edu.cn

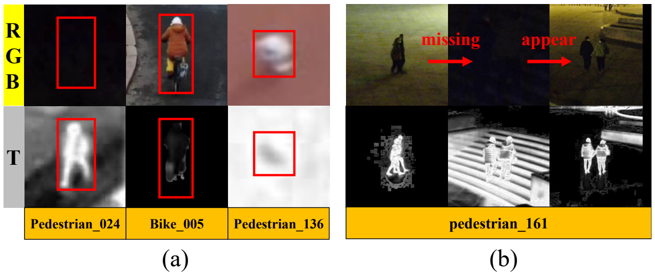


Figure 2: (a) Diverse dual-modality appearances of targets sampled in VTUAV dataset. (b) During the dynamic tracking process, the state of the RGB appearance of the target is constantly changing, resulting in dynamic relationships between modality appearances.

using additional customized modality feature fusion modules to obtain the fusion features. Subsequently, a relation modeling operation between the fusion features of the template and search region, such as online-training [Nam and Han, 2015], cross-correlation [Bertinetto *et al.*, 2016], discriminative correlation [Bhat *et al.*, 2019] and cross-attention mechanism [Hui *et al.*, 2023], will be performed. After relation modeling, the processed search region fusion features are used for prediction. However, the separate subnetworks lead to the lack of interaction between two modalities during the feature extraction stage. As a consequence, the network can only extract regular features from each modality, rather than the dynamic features based on the state of modalities. As shown in Fig. 2, such pattern is not fit to RGB-T tracking especially in complex environments, because different targets have diverse dual-modality appearances, and the appearances of both modalities can change continuously with the environment. Temporary changing or missing appearances in the corresponding modality frequently happened due to the factors like occlusion, illumination variation, or thermal, which leads to the regions covered by the appearances of both modalities are not always consistent. In addition, three-stage fusion tracking paradigm greatly limits the speed improvement.

We propose a unified single-stage Transformer RGB-T tracking network USTrack to solve the above problems. As shown in Fig. 1(b), the core of USTrack is to unify feature extraction, feature fusion, and relation modeling into a single ViT [Dosovitskiy *et al.*, 2020] backbone for simultaneous execution, efficiently obtaining search region fusion features used for prediction. Specifically, we first map the image patches from two modalities to appropriate latent spaces through a dual embedding layer to align the patterns and mitigate the impact of intrinsic heterogeneity to feature fusion. Then, within the attention layers of the ViT backbone, we directly concatenate the token sequences of the four images from the template and search region, upon which we then apply the self-attention operation to the concatenated features. In this self-attention operation, the attention weights between the features of the same image are responsible for extracting modality features, while the weights between the features of images from different modalities are responsible for fusing complementary modality information. The attention weights between the template images and the search region images are responsible for relation modeling. Therefore, we can conveniently unify the three functional stages of RGB-T tracking through the self-attention mechanism for simultaneous execution. This unification of feature extraction and feature fusion alleviates the lack of modality interaction during the feature extraction phase in traditional three-stage RGB-T tracking frameworks, allowing us to directly extract fused features from the template and search region under the modalities interaction. The further unification of fusion feature extraction and relation modeling helps us achieve joint feature extraction, fusion and relation modeling for the first time in the RGB-T tracking networks without designing any complex customized fusion modules, greatly simplifying the current network architecture of RGB-T tracking. The high parallelism of the self-attention also help USTrack achieve a speed more than twice that of existing SoTA methods.

In challenging scenarios, invalid modalities often provide a large amount of noise information. At present, most Transformer-based networks [Xiao *et al.*, 2022; Hou *et al.*, 2022; Hui *et al.*, 2023] directly concatenate or weighted sum the fusion features from search regions of two modalities for final prediction, which inevitably introduces noise information for the final prediction. In order to reduce the impact of noise information, unlike them, we propose a feature selection mechanism based on modality reliability. This mechanism reduces the impact of noise information on prediction by discarding fusion features from invalid modalities. Our contributions are summarized as follows:

- We propose joint feature extraction, fusion, and relation modeling method. It can extract the fusion features of templates and search regions under the interaction of modalities, and simultaneously perform the relation modeling. For the first time, an efficient and concise single-stage fusion tracking paradigm has been provided for RGB-T tracking without the need for designing any customized and specialized feature fusion modules.
- We propose the feature selection mechanism based on modality reliability, which can discard fusion features of invalid modalities according to the modality reliabilities of different tracking environments, thereby reducing the impact of noise information on final prediction and further improving tracking performance.
- USTrack exhibits new state-of-the-art performance on benchmark GTOT [Li *et al.*, 2016], RGBT234 [Li *et al.*, 2019], and VTUAV [Zhang *et al.*, 2022a] while creating the fastest inference speed at 84.2 FPS. In particular, MPR/MSR on the short-term and long-term subsets of VTUAV increased by 11.1%/11.7% and 11.3%/9.7%.

2 Related Work

Due to the lack of global perception ability in CNN networks, complementary information cannot be directly aggregated across modalities. So almost all CNN-Based RGB-T tracking networks are designed under the three-stage fusion tracking framework. In this part, we briefly summarize the Transformer-based RGB-T tracking methods, which are the most relevant works to us. With the introduction of Transformer into RGB-T tracking, the attention mechanism was initially only used in the feature fusion stage. DRGC-Net [Mei *et al.*, 2023] and MIRNet [Hou *et al.*, 2022] use cross-attention to enhance discriminative features from one modality to another, and assign adaptive weights to features of two modalities through gating mechanism to filter redundant and noise information. APFNet [Xiao *et al.*, 2022] proposes an attribute-based progressive fusion network, which enhances the discriminative information specific to challenging attributes through cross-attention. However, the aforementioned Transformer-based RGB-T tracking methods are designed within a detection-based tracking framework[Nam and Han, 2015]. On one hand, during the feature extraction stage, the modality features lack interaction due to the limited global context modeling capability of convolutional neural networks. On the other hand, although some RGB-T tracking networks [Hou *et al.*, 2022] based on RT-MDNet [Jung *et al.*, 2018] have almost achieved real-time inference speed, they still follow a three-stage tracking paradigm, which first extracts modality features separately, then fuses features through various attention mechanism, and finally perform the relation modeling operation between the template and search region through online training and continuous fine-tuning, resulting in significant speed bottlenecks for these RGB-T tracking networks.

The latest works TBSI [Hui *et al.*, 2023] and ViPT [Zhu *et al.*, 2023] adopt the powerful RGB tracking network OSTrack [Ye *et al.*, 2022] as their base network architecture, achieving the unification of feature extraction and relation modeling. However, influenced by the three-stage fusion tracking paradigm, they still design the fusion module as a separate component, which is inserted between two Transformer encoders to obtain the fusion features of templates and search regions. It is worth noting that TBSI [Hui *et al.*, 2023] achieves feature fusion by inserting a complex cross-attention fusion module between Transformer encoders. This approach alleviates the lack of interaction between modalities during the feature extraction stage and significantly improves the performance. However, due to the extra complex cross-attention fusion module, TBSI has only just achieved real-time performance. Unlike ViPT and TBSI, in order to achieve more efficient and concise interaction between modalities during the feature extraction stage, and to enable the fusion tracking network to achieve faster inference speed, despite being inspired by joint feature extraction and relation modeling method [Ye *et al.*, 2022], we completely freed ourselves from the influence of the three-stage fusion tracking paradigm and proposed joint feature extraction, fusion and relation modeling for the first time. We further attempt to unify feature extraction, feature fusion, and relation modeling directly through the self-attention layer within the ViT. It not only can effectively alleviate the problem of lack of interaction between modalities in the feature extraction stage to improve tracking performance, but also fully utilizes the high parallelism of self attention operations, simplifying and accelerating the RGB-T tracking network.

3 Unified Single-Stage RGB-T Tracking

Overview. As shown in Fig. 3, the overall architecture of USTrack consists of three components: a dual embedding layer, a single ViT backbone and the dual prediction heads with feature selection mechanism based on modality reliability. The dual embedding layer uses two learnable embedding layers to map inputs belonging to different modalities to a latent space that is conducive to fusion, reducing the impact of intrinsic heterogeneity of modalities on feature fusion which is based on attention similarity weights. We choose single ViT as our backbone network to achieve joint feature extraction, fusion, and relationship modeling, unifying the three functional stages of RGB-T tracking and providing an efficient single-stage network for RGB-T tracking. Feature selection mechanism based on modality reliability includes two prediction heads and two reliability evaluation modules. It will help the network select search region fusion features generated by the modalities that are more suitable for the current tracking scene for final prediction, reducing the impact of noise caused by invalid modalities on the final results.

3.1 Dual Embedding Layer

The input of USTrack is a pair of target template images and a pair of search region images, containing four images, namely, the RGB template image $\boldsymbol{z}_{image}^{rgb} \in \mathbb{R}^{H_z \times W_z \times 3}$, the RGB search region image $\boldsymbol{x}_{image}^{rgb} \in \mathbb{R}^{H_x \times W_x \times 3}$, the thermal template image $\boldsymbol{z}_{image}^{t} \in \mathbb{R}^{H_z \times W_z \times 3}$ and the thermal search region image $\boldsymbol{x}_{image}^{t} \in \mathbb{R}^{H_x \times W_x \times 3}$. They are first split and flattened into sequences of patches $\boldsymbol{z}_{rgb}, \boldsymbol{z}_t \in \mathbb{R}^{N_z \times (3P^2)}$ and $\boldsymbol{x}_{rgb}, \boldsymbol{x}_t \in \mathbb{R}^{N_x \times (3P^2)}$, where $P \times P$ is the resolution of each patch, and $N_z = \frac{H_z W_z}{P^2}$, $N_x = \frac{H_x W_x}{P^2}$ are the number of patches of template and search region respectively. Then, two trainable linear layers with parameters $\boldsymbol{E}_{rgb} \in \mathbb{R}^{(3P^2) \times D}$ and $\boldsymbol{E}_t \in \mathbb{R}^{(3P^2) \times D}$ are used to project $\boldsymbol{z}_{rgb}$, $\boldsymbol{x}_{rgb}$ and $\boldsymbol{z}_t$, $\boldsymbol{x}_t$ into D dimension latent space. The output of this projection are four patch embeddings $\hat{\boldsymbol{z}}_{rgb}$, $\hat{\boldsymbol{x}}_{rgb}$ and $\hat{\boldsymbol{z}}_t$, $\hat{\boldsymbol{x}}_t$. Learnable $1D$ position embeddings $\boldsymbol{P}_z$ and $\boldsymbol{P}_x$ are added to the template patch embeddings $\hat{\boldsymbol{z}}_{rgb}$, $\hat{\boldsymbol{z}}_t$ and search region patch embeddings $\hat{\boldsymbol{x}}_{rgb}$, $\hat{\boldsymbol{x}}_t$ separately, and Learnable $1D$ modality embeddings $\boldsymbol{M}_{rgb}$ and $\boldsymbol{M}_t$ are added to the RGB patch embeddings $\hat{\boldsymbol{z}}_{rgb}$, $\hat{\boldsymbol{x}}_{rgb}$ and thermal patch embeddings $\hat{\boldsymbol{z}}_t$, $\hat{\boldsymbol{x}}_t$ separately. The patch embeddings after adding position and modality embeddings are final features called token embeddings. The above operations can be represented as follows:

$$\hat{\boldsymbol{z}}_{rgb} = \left[\boldsymbol{z}_{rgb}^1 \boldsymbol{E}_{rgb}; \boldsymbol{z}_{rgb}^2 \boldsymbol{E}_{rgb}; ...; \boldsymbol{z}_{rgb}^{N_z} \boldsymbol{E}_{rgb}\right] + \boldsymbol{P}_z + \boldsymbol{M}_{rgb}, \tag{1}$$

$$\hat{\boldsymbol{z}}_t = \left[\boldsymbol{z}_t^1 \boldsymbol{E}_t; \boldsymbol{z}_t^2 \boldsymbol{E}_t; ...; \boldsymbol{z}_t^{N_z} \boldsymbol{E}_t\right] + \boldsymbol{P}_z + \boldsymbol{M}_t, \tag{2}$$

$$\hat{\boldsymbol{x}}_{rgb} = \left[\boldsymbol{x}_{rgb}^1 \boldsymbol{E}_{rgb}; \boldsymbol{x}_{rgb}^2 \boldsymbol{E}_{rgb}; ...; \boldsymbol{x}_{rgb}^{N_x} \boldsymbol{E}_{rgb}\right] + \boldsymbol{P}_x + \boldsymbol{M}_{rgb}, \tag{3}$$

$$\hat{\boldsymbol{x}}_t = \left[\boldsymbol{x}_t^1 \boldsymbol{E}_t; \boldsymbol{x}_t^2 \boldsymbol{E}_t; ...; \boldsymbol{x}_t^{N_x} \boldsymbol{E}_t\right] + \boldsymbol{P}_x + \boldsymbol{M}_t. \tag{4}$$

After passing the dual embedding layer, RGB template token embeddings $\hat{\boldsymbol{z}}_{rgb}$, thermal template token embeddings $\hat{\boldsymbol{z}}_t$, RGB search region token embeddings $\hat{\boldsymbol{x}}_{rgb}$ and thermal search region token embeddings $\hat{\boldsymbol{x}}_t$ will be input into the backbone for subsequent processing.

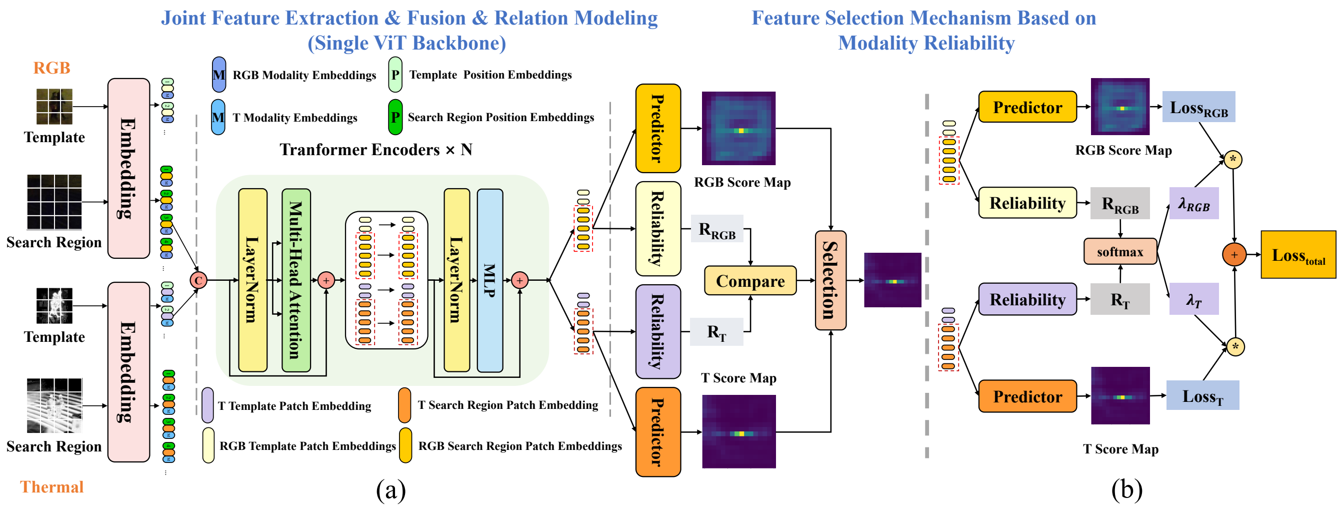


Figure 3: (a) The overall architecture of USTrack. The template and search region are split, flattened, and linear projected through the dual embedding layer. Image embeddings are then concatenated and fed into Transformer encoder layers for joint feature extraction, fusion and relation modeling. The feature selection mechanism is responsible for selecting fusion features with higher reliability for result prediction. (b) The training of the feature selection mechanism based on modality reliability.

3.2 Joint Feature Extraction & Fusion & Relation Modeling

The self-attention mechanism is the core component of the ViT, and it is also the key to performing joint feature extraction, feature fusion and relation modeling in a single ViT backbone. From the perspective of the self-attention mechanism, we take the RGB search region token embeddings $\hat{\boldsymbol{x}}_{rgb}$ as an example to further analyze the intrinsic reasons why the proposed network is able to realize simultaneous feature extraction, feature fusion and relation modeling.

In the attention layer, the token sequences $\hat{\boldsymbol{x}}_{rgb}$, $\hat{\boldsymbol{x}}_{t}$, $\hat{\boldsymbol{z}}_{rgb}$, $\hat{\boldsymbol{z}}_{t}$ from dual embedding layers are concatenated as $\boldsymbol{H} = \left[\widehat{\boldsymbol{x}}_{rgb}; \widehat{\boldsymbol{x}}_{t}; \widehat{\boldsymbol{z}}_{rgb}; \widehat{\boldsymbol{z}}_{t}\right] \in \mathbb{R}^{(2N_x+2N_t)\times D}$. Then Self-attention operation is performed on $\boldsymbol{H}$ as follows:

$$\boldsymbol{M} = \boldsymbol{A}\cdot\boldsymbol{V} = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\cdot\boldsymbol{V}, \tag{5}$$

$$\boldsymbol{Q}\boldsymbol{K}^T = [\boldsymbol{Q}^x_{rgb}; \boldsymbol{Q}^x_{t}; \boldsymbol{Q}^z_{rgb}; \boldsymbol{Q}^z_{t}][\boldsymbol{K}^x_{rgb}; \boldsymbol{K}^x_{t}; \boldsymbol{K}^z_{rgb}; \boldsymbol{K}^z_{t}]^T, \tag{6}$$

$$\boldsymbol{V} = \left[\boldsymbol{V}^x_{rgb}; \boldsymbol{V}^x_{t}; \boldsymbol{V}^z_{rgb}; \boldsymbol{V}^z_{t}\right], \tag{7}$$

where $\boldsymbol{M}$ is the output of self-attention operation. $\boldsymbol{A}$ is the attention weight. $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are query, key and value matrices separately. The superscripts z and x denote matrix items belonging to the template and search region. The subscripts rgb and t denote matrix items belonging to the RGB modality and thermal modality. The calculation of attention weights in Eq. (6) can be expanded to follows:

$$\begin{aligned}\boldsymbol{Q}\boldsymbol{K}^T &= [\boldsymbol{Q}^x_{rgb}{\boldsymbol{K}^x_{rgb}}^T, \boldsymbol{Q}^x_{rgb}{\boldsymbol{K}^x_{t}}^T, \boldsymbol{Q}^x_{rgb}{\boldsymbol{K}^z_{rgb}}^T, \boldsymbol{Q}^x_{rgb}{\boldsymbol{K}^z_{t}}^T; ...]\\ &= [\boldsymbol{W}^{x_{rgb}}_{x_{rgb}}, \boldsymbol{W}^{x_{rgb}}_{x_t}, \boldsymbol{W}^{x_{rgb}}_{z_{rgb}}, \boldsymbol{W}^{x_{rgb}}_{z_t}; ...],\end{aligned} \tag{8}$$

where the left part of Eq. (8) represents the calculation and the output of attention weights between the RGB search region tokens and the other inputs. the output of self-attention operation can be further written as follows:

$$\begin{aligned}\boldsymbol{M} =& [\boldsymbol{W}^{x_{rgb}}_{x_{rgb}}\boldsymbol{V}^x_{rgb} + \boldsymbol{W}^{x_{rgb}}_{x_t}\boldsymbol{V}^x_{t}\\ &+\boldsymbol{W}^{x_{rgb}}_{z_{rgb}}\boldsymbol{V}^z_{t} + \boldsymbol{W}^{x_{rgb}}_{z_t}\boldsymbol{V}^z_{t}; ...],\end{aligned} \tag{9}$$

where the left part of Eq. (9) is the output corresponding to the RGB search region tokens after the self-attention operation. $\boldsymbol{W}^{x_{rgb}}_{x_{rgb}}\boldsymbol{V}^x_{rgb}$ is responsible for aggregating the RGB search region image feature (RGB modality feature extraction). $\boldsymbol{W}^{x_{rgb}}_{x_t}\boldsymbol{V}^x_{t}$ is responsible for aggregating the thermal modality-specific information based on semantic similarity between two modalities features (feature fusion and modality features interaction). The attention weights can intuitively measure the semantic similarity between modalities. Network can model modality-sharing information based on this similarity. The aggregation of complementary information enables the network to promptly adjust the subsequent extraction of features in RGB search region image. $\boldsymbol{W}^{x_{rgb}}_{z_{rgb}}\boldsymbol{V}^z_{t}$ is responsible for aggregating RGB template image feature to further obtain the relation information between the RGB template and the RGB search region (relation modeling based on modality-specific information). $\boldsymbol{W}^{x_{rgb}}_{z_t}\boldsymbol{V}^z_{t}$ is responsible for aggregating thermal template image feature to further obtain the relation information between the thermal template and the RGB search region (relation modeling based on modality-sharing information). The RGB search region fusion features, which contains relation information, can be used for prediction. Therefore, with the global perception ability of the self-attention, we seamlessly unify feature extraction, feature fusion, and relation modeling into a single ViT backbone. The network can directly extract fusion features of the template and search region under the mutual interaction of modalities, and simultaneously performs relation modeling between fusion features of the template and search region. This alleviates the lack of interaction and guidance between modalities during the feature extraction stage, as well as the problem of additional fusion modules significantly affecting the inference speed of the RGB-T tracking network. Additionally, by inheriting the advantages of relation modeling which is performed by the self-attention, the network can extract more target-specific search region fusion features for prediction under the guidance of two templates.

3.3 Feature Selection Mechanism Based on Modality Reliability

After passing the ViT backbone, two search region fusion features can be obtained for final prediction: Thermal-assisted RGB fusion features based on the RGB search region image, and RGB-assisted thermal fusion features based on the thermal search region image. Both fusion features contain the fusion information of modalities and the relation information between the template and the search region, which can be directly used for target position prediction. To avoid the impact of interference information from invalid modalities on the final prediction. Unlike other networks that obtain fused features based on attention operations, we do not directly concatenate the two fused features or perform weighted sum operations. Instead, we directly discard invalid modality features that are not suitable for the current tracking scene.

As shown in Fig. 3(b), during the training phase, we equip each fusion feature with a prediction head and a reliability evaluation module. We set the same loss for each prediction head and let each reliability evaluation module output a adaptive weight as the modality reliability for the loss of the corresponding prediction head. Then, they are combined into a final total loss function for end-to-end training. This method allows the modality that is not suitable for the current scene to produce inferior results, which will result in larger losses, and then uses the difference between two predictions to guide the modality reliability evaluation module to assign smaller weights to the larger loss by minimizing the overall loss function. Conversely, for fusion features that are more suitable for the current scene, it assigns larger weights. During the testing phase, the network will simultaneously output two results and evaluate the reliability of both modalities. Based on the reliability R_{RGB} and R_T, we select the predicted results with higher reliability scores as the final output.

We adopt the prediction head of OSTrack [Ye *et al.*, 2022] directly as our prediction head. The detailed information and corresponding settings can be found in OSTrack. The loss corresponding to the two prediction heads are set as follows:

$$\mathcal{L}_{RGB} = \mathcal{L}_{cls_{RGB}} + \lambda_{giou}\mathcal{L}_{giou_{RGB}} + \lambda_{L_1}\mathcal{L}_{1_{RGB}}, \quad (10)$$

$$\mathcal{L}_{T} = \mathcal{L}_{cls_{T}} + \lambda_{giou}\mathcal{L}_{giou_{T}} + \lambda_{L_1}\mathcal{L}_{1_{T}}, \quad (11)$$

where $\mathcal{L}_{RGB}$ and $\mathcal{L}_T$ are the loss for each prediction head, $\mathcal{L}_{cls_{RGB}}$ and $\mathcal{L}_{cls_T}$ are the weighted focal loss for classification, $\mathcal{L}_{1_{RGB}}$ and $\mathcal{L}_{1_T}$ are the $\mathcal{L}_1$ loss, $\mathcal{L}_{giou_{RGB}}$ and $\mathcal{L}_{giou_T}$ are the generalized IoU loss, and λ_{giou} = 2 and λ_{L_1} = 5 are the regularization parameters. On the basis, a modality reliability evaluation module is added to each search region fusion features. The evaluation module is a fully convolutional neural network, which consists of several stacked Conv-BN-ReLU layers. Two modality reliability evaluation modules will output the reliability scores $R_{RGB}, R_T \in \mathbb{R}$ respectively. In order to prevent the model from directly making the weight R_{RGB} and R_T zero to minimize the overall loss during the training process, we softmax the reliability scores to obtain the adaptive weight λ_{RGB} and λ_T and overall loss as follows:

$$\lambda_{RGB}, \lambda_T = softmax(R_{RGB}, R_T), \quad (12)$$

$$\mathcal{L}_{total} = \lambda_{RGB}\mathcal{L}_{RGB} + \lambda_T\mathcal{L}_T, \quad (13)$$

where modality reliabilities λ_{RGB} and λ_T are used as the adaptive weights of the loss of the two prediction heads, and the two losses with adaptive weights are combined together as the overall loss to train the model.

4 Experiment

4.1 Experiment Settings

We compare our method with previous state-of-the-art RGB-T tracking methods on three benchmarks including VTUAV, RGBT234, and GTOT. GTOT and RGBT234 use success rate SR and precision rate PR as evaluation metrics. VTUAV use Maximum Precision Rate MPR and Maximum Success Rate MSR as evaluation metrics. SR measures the ratio of tracked frames, determined by the Interaction-over-Union (IoU) between tracking result and ground truth. With different overlap thresholds, a success plot (SP) can be obtained, and SR is calculated as the area under curve of SP. MSR adopts the maximum overlap in frame level as the final score. PR measures the percentage of frames whose distance between the predicted position and the ground truth is less than a certain threshold τ. Similar to MSR, MPR adopt the smaller center distance as the final score. τ is set to 20 in our experiment.

Our model is implemented based on Python 3.8, PyTorch 2.0.0. All experiments are conducted on one NVIDIA RTX3090 GPU. We adopt AdamW as the optimizer with 1e-4 weight decay. The learning rate is set as 4e-5 for the backbone and 4e-4 for other parameters. The search regions are resized to 256×256 and templates are resized to 128×128. Each batch size is set to 24, and each epoch contains 30k image pairs. In order to fairly compare our method with other SoTA methods, we aligned our experimental conditions with other methods. We pretrained the network on RGB tracking datasets such as COCO [Lin *et al.*, 2014], LaSOT [Fan *et al.*, 2018], GOT-10k [Huang *et al.*, 2018], and TrackingNet [Muller *et al.*, 2018]. When testing on the GTOT and RGBT234, we only used LasHeR as the training set. When testing on the short-term and long-term testing sets of VTUAV, we only use the training set of VTUAV for training.

4.2 Comparison with SoTA Methods and Analysis

We test our network USTrack on three popular RGB-T tracking benchmarks, comparing performance and speed with the SoTA trackers, such as FSRPN, mfDimp, DAFNet , DAPNet, MANet, CAT, CMPP, JMMAC, MANet++, ADRNet, SiamCDA, M5LNet, TFNet, DMCNet, MFGNet, APFNet, HMFT, MIRNet, ECMD, ViPT and TBSI, to validate the effectiveness of our method. The test results on three datasets show that our method has achieved significant improvements in both performance and inference speed.

Evaluation on VTUAV Dataset. VTUAV dataset is the latest and largest RGB-T tracking dataset, which is currently the

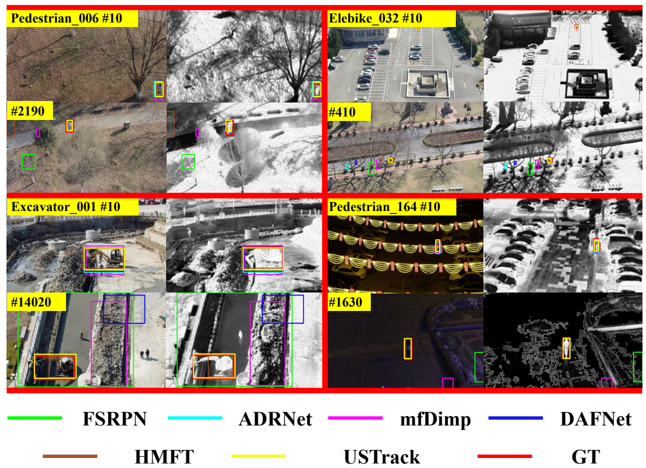
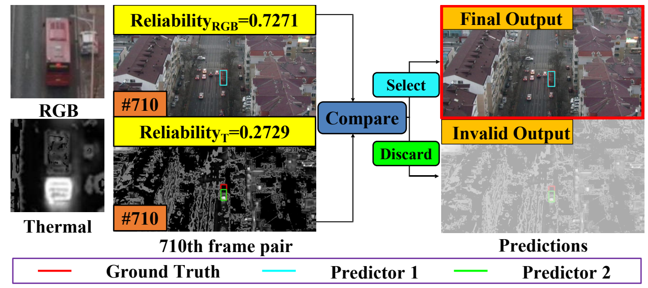
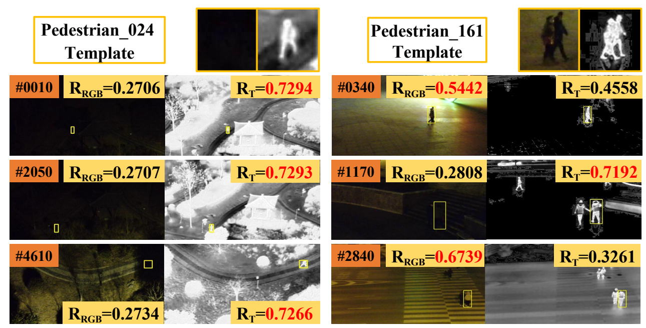


Figure 4: Visualization between our method and other RGB-T trackers on four representative sequences which include multiple challenge attributes from VTUAV dataset.



(a) Predictor1 represents the prediction result based on the fusion features of RGB search region image, while Predictor2 represents the prediction result based on the fusion features of thermal search region image. Reliability$_{RGB}$ and Reliability$_{T}$ represent the reliability evaluations of the two modalities respectively.



(b) $\boldsymbol{R_{RGB}}$ and $\boldsymbol{R_T}$ represent the reliability evaluations of the two modalities respectively.

Figure 5: (a) This flowchart illustrates the inference process of the feature selection mechanism based on modality reliability on frame 710 of the VTUAV bus_007 video sequence.(b) This figure shows that the feature reliability evaluation module can provide real-time reliability evaluation for each modality in continuous video frames.

only dataset that provides a test for long-term tracking performance of RGB-T tracking methods. Long-term sequences can effectively demonstrate the differences in the target appearances of different modalities, and the persistent changing relationships of two appearances during the tracking process. As shown in Tab. 1 and Fig. 4, despite USTrack being a short-term tracking network with no template updating or local-to-global strategies for long-term tracking, we still conduct tests on the short-term and long-term subsets of VTUAV to verify the performance of USTrack. The results were very satisfactory. Compared to the SoTA methods HMFT and HMFT-LT, we achieved 11.1%/11.7% and 11.3%/9.7% increases in MPR/MSR on the VTUAV short-term dataset and VTUAV long-term dataset, respectively. Our speed was also 2.78 times and 10.4 times faster than the SoTA method HMFT and HMFT-LT [Zhang *et al.*, 2022a], significantly surpassing the baseline methods of VTUAV. To our knowledge, HMFT-LT is currently the only long-term RGB-T tracking method. In comparison, we have achieved significantly better performance, with a tracking speed that is ten times faster than it.

In order to validate that USTrack can adapt well to the diverse dual-modality appearances of targets and the dynamic relationships between modalities, we analyzed the performance of USTrack across all challenging attributes on both short-term and long-term subsets of VTUAV. As shown in Tab. 2 and Tab. 3, in terms of the evaluation metrics MPR/MSR scores on the short-term and long-term datasets of VTUAV, USTrack achieved the highest performance improvements of 23.8%/22.6%&10.5%/10.1%, 39.8%/33.4%&19.1%/16.3%, 20.1%/18.9%&12.2%/10.8%, 11.1%/11.9%&14.9%/13.5%, 12.3%/12.6%&10.6%/9.2% and 5.7%/6.4%&19.7%/16.4% on the challenge attributes deformation (DEF), scale variation (SV), full occlusion (FO), partial occlusion (PO), thermal crossover (TC), and extreme illumination (EI), respectively. For more experimental results of USTrack on VTUAV attributes, please refer to the **Appendix**. In particular, the DEF and SV attributes effectively demonstrate the differences in the dual-modality appearances of the targets. The FO, PO, TC, and EI attributes can cause the appearance of the corresponding modality to change or disappear, effectively demonstrating the dynamic relationship between two appearances of the target during the tracking process. The most significant performance improvement achieved by USTrack in these attributes effectively proves that the joint feature extraction, fusion and relation modeling method can adapt to diverse dual-modality appearances of targets and the dynamic relationships between modalities by alleviating the lack of modality interaction in the modality feature extraction stage of the three-stage fusion tracking paradigm. Moreover, with the help of the unified single-stage fusion tracking paradigm, USTrack, through a simple network structure and the high parallelism of self-attention operations, has created the fastest inference speed 84.2FPS for RGB-T tracking to date.

Evaluation on RGBT234 Dataset. RGBT234 is currently the most widely used large-scale RGB-T tracking benchmark dataset, consisting of 234 highly aligned videos with about 234K image pairs in total. As shown in Tab.1, Unlike ViPT and TBSI, which also use pure ViT as the backbone but still design customized complex fusion modules under the three-stage fusion tracking paradigm, we provide a novel single-stage fusion tracking paradigm that achieves joint feature

Method	Pub	GTOT		RGBT234		VTUAV-short		VTUAV-long		Speed
		PR	SR	PR	SR	MPR	MSR	MPR	MSR	FPS
FSRPN [Kristan *et al.*, 2019]	ICCVW'19	89.0	69.5	71.9	52.5	65.3	54.4	36.6	31.4	36.8
mfDimp [Zhang *et al.*, 2019]	ICCVW'19	83.6	69.7	84.6	59.1	67.3	55.4	31.5	27.2	34.6
DAFNet [Gao *et al.*, 2019]	ICCVW'19	89.1	71.6	79.6	54.4	62.0	45.8	25.3	18.8	20.5
DAPNet [Zhu *et al.*, 2019]	ACM MM'19	88.2	70.7	76.6	53.7	-	-	-	-	-
MANet [Lu *et al.*, 2020a]	TIP'20	89.4	72.4	77.7	53.9	-	-	-	-	2.1
CAT [Li *et al.*, 2020]	ECCV'20	88.9	71.7	80.4	56.1	-	-	-	-	-
CMPP [Wang *et al.*, 2020]	CVPR'20	92.6	73.8	82.3	57.5	-	-	-	-	-
JMMAC [Zhang *et al.*, 2020]	TIP'21	90.2	73.2	79.0	57.3	-	-	-	-	-
MANet++ [Lu *et al.*, 2020b]	TIP'21	88.2	70.7	79.5	55.9	-	-	-	-	25.4
ADRNet [Zhang *et al.*, 2021]	IJCV'21	90.4	73.9	80.7	57.1	62.2	46.6	23.5	17.5	25.0
SiamCDA [Zhang *et al.*, 2022b]	TCSVT'21	87.7	73.2	79.5	54.2	-	-	-	-	24.0
M5LNet [Tu *et al.*, 2021]	TIP'22	89.6	71.0	79.5	54.2	-	-	-	-	9.0
TFNet [Zhu *et al.*, 2022]	TCSVT'22	88.6	72.9	80.6	56.0	-	-	-	-	-
DMCNet [Lu *et al.*, 2020c]	TNNLS'22	-	-	83.9	59.3	-	-	-	-	-
MFGNet [Wang *et al.*, 2021]	TMM'22	88.9	70.7	78.3	53.5	-	-	-	-	3.0
APFNet [Xiao *et al.*, 2022]	AAAI'22	90.5	73.7	82.7	57.9	-	-	-	-	1.9
HMFT [Zhang *et al.*, 2022a]	CVPR'22	91.2	74.9	78.8	56.8	75.8	62.7	41.4	35.5	30.2
HMFT_LT [Zhang *et al.*, 2022a]	CVPR'22	-	-	-	-	-	-	53.6	46.1	8.1
MIRNet [Hou *et al.*, 2022]	ICME'23	90.9	74.4	81.6	58.9	-	-	-	-	30.0
ECMD [Zhang *et al.*, 2023]	CVPR'23	90.7	73.5	84.4	60.1	-	-	-	-	18.0
ViPT [Zhu *et al.*, 2023]	CVPR'23	-	-	83.5	61.7	-	-	-	-	-
TBSI [Hui *et al.*, 2023]	CVPR'23	-	-	87.1	63.7	-	-	-	-	36.2
USTrack (Ours)	-	**93.4**	**78.3**	**87.4**	**65.8**	**86.9**	**74.4**	**64.9**	**55.8**	**84.2**

Table 1: Comparison with state-of-the-art methods on GTOT, RGBT234, VTUAV short-term subset and long-term subset.

Attribute	FSRPN	DAFNet	mfDimp	ADRNet	HMFT	USTrack
DEF	53.8/47.3	43.7/35.1	60.0/52.1	45.1/31.7	68.4/59.7	**92.2/82.3**
EI	60.1/47.7	67.6/49.4	63.3/49.4	66.4/48.5	74.7/58.8	**80.4/65.2**
FO	45.4/39.1	32.1/27.5	38.5/32.3	36.2/26.9	37.8/32.1	**85.2/72.5**
PO	54.2/46.1	51.4/37.9	54.9/46.2	48.9/36.5	64.8/53.7	**84.9/72.6**
SV	66.8/65.9	53.2/39.6	65.9/55.8	53.0/39.6	72.9/61.8	**85.2/74.4**
TC	48.2/40.4	41.9/34.5	52.6/47.4	45.5/38.1	56.4/48.7	**67.5/60.6**

Table 2: Top six attributes with improvement on VTUAV short-term.

Attribute	FSRPN	DAFNet	mfDimp	ADRNet	HMFT-LT	USTrack
DEF	35.1/29.4	24.7/20.3	40.8/34.8	18.7/13.7	67.1/58.2	**77.6/68.3**
EI	32.8/27.1	20.8/15.5	23.1/21.1	20.0/14.9	47.3/40.3	**67.0/56.7**
FO	26.3/22.1	25.4/18.4	29.4/24.0	25.1/17.2	43.8/37.1	**62.9/53.4**
PO	33.0/27.3	20.3/14.8	26.6/22.8	19.6/13.7	55.8/47.2	**68.0/58.0**
SV	37.7/32.3	23.0/17.7	31.7/27.9	21.3/16.2	53.9/46.7	**64.5/55.9**
TC	30.4/22.4	21.0/13.9	20.1/13.5	20.6/13.0	45.9/35.6	**60.8/49.1**

Table 3: Top six attributes with improvement on VTUAV long-term.

extraction, fusion, and relation modeling for the first time without the need for any additional feature fusion modules, greatly simplifying the RGB-T tracking network architecture while bringing speed and performance improvements. Compared to the most advanced trackers ViPT and TBSI. We have improved PR/SR on RGBT234 by 3.9%/4.1% and 0.3%/2.1% respectively. TBSI has a speed of 36FPS, while our speed has reached 84FPS. Both performance and speed can prove the effectiveness and efficiency of our method.

Evaluation on GTOT Dataset. GTOT is the first standard dataset in the field of RGB-T tracking. It contains 50 RGB-T video sequences and 7 challenge attributes. We also conducted testing on this dataset and achieved SoTA performance. The test results are shown in Tab. 1. compared with the SoTA methods CMPP and HMFT, our PR/SR scores improved by 0.8%/4.5% and 2.2%/3.4% respectively, while maintaining the fastest inference speed.

Model	PR	SR
Single embedding layer	85.6	63.2
Dual embedding layer (Ours)	**87.4**	**65.8**

Table 4: Results of the ablation of dual embedding layer.

Model	PR	SR
Based on RGB search region	86.2	64.2
Based on Thermal search region	86.3	64.7
Based on weighted concatenation	86.8	64.2
Based on weighted summation	86.1	63.9
Dual prediction head with selection (Ours)	**87.4**	**65.8**

Table 5: Comparison with different prediction head structures.

4.3 Ablation Experiment and Analysis

Ablation of Dual Embedding Layer. To verify the effectiveness of the dual embedding layer structure, we conducted ablation experiments on the RGBT234 dataset. As a comparison, we have all inputs use the same embedding layer. The results are shown in Tab. 2. The single embedded layer structure resulted in a performance decrease of 1.8% and 2.6% in PR and SR scores. The results show that the use of two independent embedding layers can map the features of two modalities into the latent space conducive to fusion, which can alleviate the impact of the intrinsic heterogeneity of modalities on feature fusion based on attention weight.

Ablation of Feature Selection Mechanism. In order to verify the effectiveness of the feature selection mechanism based on modality reliability, we conducted comparative experiments between our dual prediction head structure with feature selection mechanism and several common prediction head structures on RGBT234. We set up the following comparative experiments, namely, single prediction head based on the fusion features from a single modality search region, single prediction head based on the weighted concatenated fusion features from two search regions, and single prediction head structure based on weighted summation of fusion features from two search regions. All prediction heads have the same structure. To ensure the fairness of the comparative experiment, the weight module used for weighted summation or concatenation of fusion features has the same structure as the modal reliability evaluation module. The experimental results are shown in Tab. 3. Compared to other prediction heads, our dual prediction head structure with feature selection mechanism based on modality reliability performs better. As shown in Fig. 5, we also visualized the actual test sequence, and the visualization showed that our modality reliability had a good correspondence with the real scene, which intuitively reflects the reliability of each modality in the current tracking scene. USTrack will select the fusion features from the search region with high reliability scores to output better prediction results.

5 Conclusion

In this paper, we propose an efficient unified single-stage transformer RGB-T tracking network, USTrack. The core of USTrack is the introduction of the joint feature extraction, fusion and relation modeling approach to address the lack of modality interaction during the feature extraction phase in traditional three-stage fusion tracking paradigms, thereby enhancing the adaptability to diverse dual-modality appearances of targets and the dynamic relationships between modalities. Furthermore, we introduce the feature selection mechanism based on modality reliability. This mechanism discards fusion features generated from ineffective modalities, thereby reducing the impact of noise information on the final prediction to achieve better performance. USTrack has achieved SoTA performance on three mainstream datasets and set a new record for the fastest RGB-T tracking inference speed at 84.2 FPS. Notably, on the VTUAV dataset, which is currently the largest RGB-T tracking dataset, evaluation metrics MPR/MSR has increased by 11.1%/11.7% and 11.3%/9.7%.

Acknowledgments

This work was supported by the Integrated Program of National Natural Science Foundation of China under Grant No.91948303.

References

[Bertinetto *et al.*, 2016] Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H. S. Torr. Fully-convolutional siamese networks for object tracking. *ArXiv*, abs/1606.09549, 2016.

[Bhat *et al.*, 2019] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. *ICCV*, pages 6181–6190, 2019.

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

[Fan *et al.*, 2018] 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. *CVPR*, pages 5369–5378, 2018.

[Gao *et al.*, 2019] Yuan Gao, Chenglong Li, Yabin Zhu, Jin Tang, Tao He, and Futian Wang. Deep adaptive fusion network for high performance rgbt tracking. *ICCVW*, pages 91–99, 2019.

[He *et al.*, 2015] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, pages 770–778, 2015.

[Hou *et al.*, 2022] Ruichao Hou, Tongwei Ren, and Gangshan Wu. Mirnet: A robust rgbt tracking jointly with multi-modal interaction and refinement. *ICME*, pages 1–6, 2022.

[Huang *et al.*, 2018] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *TPAMI*, 43:1562–1577, 2018.

[Hui *et al.*, 2023] Tianrui Hui, Zizheng Xun, Fengguang Peng, Junshi Huang, Xiaoming Wei, Xiaolin Wei, Jiao Dai, Jizhong Han, and Si Liu. Bridging search region interaction with template for rgb-t tracking. In *CVPR*, pages 13630–13639, 2023.

[Jung *et al.*, 2018] Ilchae Jung, Jeany Son, Mooyeol Baek, and Bohyung Han. Real-time mdnet. In *ECCV*, 2018.

[Kristan *et al.*, 2019] Matej Kristan, Jiri Matas, et al. The seventh visual object tracking vot2019 challenge results. *ICCVW*, pages 2206–2241, 2019.

[Li *et al.*, 2016] Chenglong Li, Hui Cheng, Shiyi Hu, Xiaobai Liu, Jin Tang, and Liang Lin. Learning collaborative sparse representation for grayscale-thermal tracking. *TIP*, 25:5743–5756, 2016.

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

[Li *et al.*, 2020] Chenglong Li, Lei Liu, Andong Lu, Qing Ji, and Jin Tang. Challenge-aware rgbt tracking. In *ECCV*, pages 222–237. Springer, 2020.

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

[Lu *et al.*, 2020a] Andong Lu, Chenglong Li, Y. Yan, Jin Tang, and Bin Luo. Rgbt tracking via multi-adapter network with hierarchical divergence loss. *TIP*, 30:5613–5625, 2020.

[Lu *et al.*, 2020b] Andong Lu, Chenglong Li, Y. Yan, Jin Tang, and Bin Luo. Rgbt tracking via multi-adapter network with hierarchical divergence loss. *TIP*, 30:5613–5625, 2020.

[Lu *et al.*, 2020c] Andong Lu, Cun Qian, Chenglong Li, Jin Tang, and Liang Wang. Duality-gated mutual condition network for rgbt tracking. *TNNLS*, PP, 2020.

[Mei *et al.*, 2023] Jiatian Mei, Dongming Zhou, Jinde Cao, Rencan Nie, and Kangjian He. Differential reinforcement and global collaboration network for rgbt tracking. *Sensors*, 23:7301–7311, 2023.

[Muller *et al.*, 2018] 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*, pages 300–317, 2018.

[Nam and Han, 2015] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. *CVPR*, pages 4293–4302, 2015.

[Tu *et al.*, 2021] Zhengzheng Tu, Chun Lin, Wei Zhao, Chenglong Li, and Jin Tang. M5l: Multi-modal multi-margin metric learning for rgbt tracking. *TIP*, 31:85–98, 2021.

[Wang *et al.*, 2020] Chaoqun Wang, Chunyan Xu, Zhen Cui, Lingli Zhou, Tong Zhang, Xiaoya Zhang, and Jian Yang. Cross-modal pattern-propagation for rgb-t tracking. *CVPR*, pages 7062–7071, 2020.

[Wang *et al.*, 2021] Xiao Wang, Xiu Shu, Shiliang Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Mfgnet: Dynamic modality-aware filter generation for rgb-t tracking. *ArXiv*, abs/2107.10433, 2021.

[Xiao *et al.*, 2022] Yun Xiao, Mengmeng Yang, Chenglong Li, Lei Liu, and Jin Tang. Attribute-based progressive fusion network for rgbt tracking. In *AAAI*, 2022.

[Ye *et al.*, 2022] 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*, pages 341–357. Springer, 2022.

[Zhang *et al.*, 2019] Lichao Zhang, Martin Danelljan, Abel Gonzalez-Garcia, Joost van de Weijer, and Fahad Shahbaz Khan. Multi-modal fusion for end-to-end rgb-t tracking. *ICCVW*, pages 2252–2261, 2019.

[Zhang *et al.*, 2020] 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*, 30:3335–3347, 2020.

[Zhang *et al.*, 2021] Pengyu Zhang, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Learning adaptive attribute-driven representation for real-time rgb-t tracking. *IJCV*, 129:2714 – 2729, 2021.

[Zhang *et al.*, 2022a] Pengyu Zhang, Jie Zhao, D. Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. *CVPR*, pages 8876–8885, 2022.

[Zhang *et al.*, 2022b] Tianlu Zhang, Xueru Liu, Qiang Zhang, and Jungong Han. Siamcda: Complementarity- and distractor-aware rgb-t tracking based on siamese network. *TCSVT*, 32:1403–1417, 2022.

[Zhang *et al.*, 2023] Tianlu Zhang, Hongyuan Guo, Qiang Jiao, Qiang Zhang, and Jungong Han. Efficient rgb-t tracking via cross-modality distillation. In *CVPR*, pages 5404–5413, 2023.

[Zhu *et al.*, 2019] Yabin Zhu, Chenglong Li, Bin Luo, Jin Tang, and Xiao Wang. Dense feature aggregation and pruning for rgbt tracking. *ACM MM*, 2019.

[Zhu *et al.*, 2022] Yabin Zhu, Chenglong Li, Jin Tang, Bin Luo, and Liang Wang. Rgbt tracking by trident fusion network. *TCSVT*, 32:579–592, 2022.

[Zhu *et al.*, 2023] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *CVPR*, pages 9516–9526, 2023.