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

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

\textsuperscript{1}Intelligent Game and Decision Lab, China
\textsuperscript{2}Tianjin Artificial Intelligence Innovation Center, China
\textsuperscript{3}School of Control Science and Engineering, Shandong University, China
\textsuperscript{4}College of Computer, National University of Defense Technology, China
\textsuperscript{5}Department of Electrical and Computer Engineering, National University of Singapore, Singapore
\textsuperscript{6}Electronic Engineering Institute, Beijing University of Posts and Telecommunications, China
\textsuperscript{7}Department of Computer Science and Technology, Tsinghua University, China
\textsuperscript{8}School of Optoelectronics, Beijing Institute of Technology, China
\textsuperscript{9}Beijing Institute of Basic Medical Sciences, China
\textsuperscript{10}EVOL Lab, Institute of AI (TeleAI), China Telecom, China
\textsuperscript{11}School of Artificial Intelligence, Optics and Electronics, Northwestern Polytechnical University, China

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

\textsuperscript{*}Corresponding author: Dianxi Shi, dxshi@nudt.edu.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 US-Track, 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 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...
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. DRGCNet [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 mechanisms, and finally performs 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 OS-Track [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 $\tilde{x}_{rgb}^{rgb} \in \mathbb{R}^{H_1 \times W_1 \times 3}$, the thermal template image $\tilde{x}_{rgb}^{rgb} \in \mathbb{R}^{H_1 \times W_1 \times 3}$, the RGB search region image $\tilde{x}_{rgb}^{rgb} \in \mathbb{R}^{H_2 \times W_2 \times 3}$, and the thermal search region image $\tilde{x}_{rgb}^{rgb} \in \mathbb{R}^{H_2 \times W_2 \times 3}$. They are first split and flattened into sequences of patches $x_{rgb}, x \in \mathbb{R}^{N_s \times (3P^2)}$ and $x_{rgb}, x \in \mathbb{R}^{N_s \times (3P^2)}$, where $P \times P$ is the resolution of each patch, and $N_s = \frac{H_2 \times W_2}{P^2}$, $N_t = \frac{H_1 \times W_1}{P^2}$ are the number of patches of template and search region respectively. Then, two trainable linear layers with parameters $E_{rgb} \in \mathbb{R}^{(3P^2) \times D}$ and $E_t \in \mathbb{R}^{(3P^2) \times D}$ are used to project $z_{rgb}, x_{rgb}$ and $z_t, x_t$ into $D$ dimension latent space. The output of this projection are four patch embeddings $\tilde{z}_{rgb}$, $\tilde{x}_{rgb}$ and $\tilde{z}_t$, $\tilde{x}_t$. Learnable 1D position embeddings $P_z$ and $P_t$ are added to the template patch embeddings $\tilde{z}_{rgb}$, $\tilde{z}_t$, and search region patch embeddings $\tilde{x}_{rgb}$, $\tilde{x}_t$, separately, and Learnable 1D modality embeddings $M_{rgb}$ and $M_t$ are added to the RGB patch embeddings $\tilde{z}_{rgb}$, $\tilde{x}_{rgb}$ and thermal patch embeddings $\tilde{z}_t$, $\tilde{x}_t$ separately. The patch embeddings after adding position and modality embeddings are final features called token embeddings. The above operations can be represented as follows:

$$\tilde{z}_{rgb} = \left[ z_{rgb}^1 E_{rgb}; z_{rgb}^2 E_{rgb}; \ldots; z_{rgb}^{N_z} E_{rgb} \right] + P_z + M_{rgb},$$  
(1)

$$\tilde{z}_t = \left[ z_t^1 E_t; z_t^2 E_t; \ldots; z_t^{N_t} E_t \right] + P_z + M_t,$$  
(2)

$$\tilde{x}_{rgb} = \left[ x_{rgb}^1 E_{rgb}; x_{rgb}^2 E_{rgb}; \ldots; x_{rgb}^{N_x} E_{rgb} \right] + P_x + M_{rgb},$$  
(3)

$$\tilde{x}_t = \left[ x_t^1 E_t; x_t^2 E_t; \ldots; x_t^{N_x} E_t \right] + P_x + M_t.$$  
(4)

After passing the dual embedding layer, RGB template token embeddings $\tilde{z}_{rgb}$, thermal template token embeddings $\tilde{z}_t$, RGB search region token embeddings $\tilde{x}_{rgb}$ and thermal search region token embeddings $\tilde{x}_t$ will be input into the backbone for subsequent processing.
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 $\hat{x}_{rgb}$ 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 $\hat{x}_{rgb}$, $\hat{x}_t$, $\hat{x}_{rgb}$, $\hat{x}_t$ from dual embedding layers are concatenated as $H = [\hat{x}_{rgb}; \hat{x}_t; \hat{x}_{rgb}; \hat{x}_t] \in \mathbb{R}^{(2N_r + 2N_t) \times D}$. Then Self-attention operation is performed on $H$ as follows:

$$M = A \cdot V = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) \cdot V;$$  \hspace{1cm} (5)

$$QK^T = [Q^x_{rgb}; Q^t_1; Q^x_{rgb}; Q^t_2; K^x_{rgb}; K^t_1; K^x_{rgb}; K^t_2]^T, \hspace{1cm} (6)$$

$$V = [V^x_{rgb}; V^t_1; V^x_{rgb}; V^t_2], \hspace{1cm} (7)$$

where $M$ is the output of self-attention operation, $A$ is the attention weight, $Q$, $K$, and $V$ are query, key and value matrices, respectively. The superscripts $x$ and $t$ 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:

$$QK^T = [Q^x_{rgb}K^x_{rgb}^T, Q^x_{rgb}K^t_1, Q^x_{rgb}K^t_2; Q^t_1K^x_{rgb}, Q^t_2K^x_{rgb}; \ldots] = [W_{rgb,rgb}, W_{rgb,t}, W_{rgb,rgb}, W_{rgb,t}; \ldots], \hspace{1cm} (8)$$

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:

$$M = [W^x_{rgb,rgb}V^x_{rgb} + W^x_{rgb,t}V^t_1 + \ldots] \hspace{1cm} (9)$$

where the left part of Eq. (9) is the output corresponding to the RGB search region tokens after the self-attention operation. $W^x_{rgb,rgb}$ is responsible for aggregating the RGB search region image feature (RGB modality feature extraction). $W^x_{rgb,t}$ 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. $W^x_{rgb,t}$ 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). $W^t_{rgb,t}$ 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:

$$L_{RGB} = L_{cls_{RGB}} + \lambda_{giou} L_{giou_{RGB}} + \lambda_{L_1} L_{1_{RGB}},$$  

$$L_T = L_{cls_T} + \lambda_{giou} L_{giou_T} + \lambda_{L_1} L_{1_T},$$  

where $L_{RGB}$ and $L_T$ are the loss for each prediction head, $L_{cls_{RGB}}$ and $L_{cls_T}$ are the weighted focal loss for classification, $L_{giou_{RGB}}$ and $L_{giou_T}$ are the $L_1$ loss, $L_{giou_{RGB}}$ and $L_{giou_T}$ are the generalized IOU loss, and $\lambda_{giou} = 2$ and $\lambda_{L_1} = 5$ 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 $\lambda_{RGB}$ and $\lambda_T$ and overall loss as follows:

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

$$L_{total} = \lambda_{RGB} L_{RGB} + \lambda_T L_T,$$

where modality reliabilities $\lambda_{RGB}$ and $\lambda_T$ 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 VTVAV, RGBT234, and GTOT. GTOT and RGBT234 use success rate SR and precision rate PR as evaluation metrics. VTVAV use Maximum Precision Rate MPR and Maximum Success Rate MSR as evaluation metrics. SR measures the ratio of tracked frames, determined by the Intersection-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 $\tau$. Similar to MSR, MPR adopt the smaller center distance as the final score. $\tau$ 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 256x256 and templates are resized to 128x128. 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 VTVAV, we only use the training set of VTVAV 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, mDiM, DAFNet, DAPNet, MANet, CAT, CMP, JMMAC, MANet++, ADRNet, SiAmCDA, MSILNet, TFinet, 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 VTVAV Dataset. VTVAV dataset is the latest and largest RGB-T tracking dataset, which is currently the
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 VTUAUV. 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 VTUAUV 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 VTUAUV short-term dataset and VTUAUV 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 VTUAUV. 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 VTUAUV. 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 VTUAUV, 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 VTUAUV 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
We have improved PR and SR on RGBT234 by 3.9/4.1% respectively. Compared to the most advanced trackers ViPT and TBSI, our PR/SR scores are improved by 0.8/2.6% in performance. The test results are shown in Tab. 1. Compared to the SoTA methods CMPP and HMFT, our PR/SR scores improved by 0.8%/4.5% and 0.3%/2.1% respectively, while maintaining the fastest inference speed.

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

<table>
<thead>
<tr>
<th>Attribute</th>
<th>FSRPN</th>
<th>DAFNet</th>
<th>mfDimp</th>
<th>ADRNet</th>
<th>HMFT-LT</th>
<th>USTrack</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC</td>
<td>45.0/91</td>
<td>32.1/75</td>
<td>28.6/73</td>
<td>26.7/70</td>
<td>30.4/74</td>
<td>35.7/79</td>
</tr>
<tr>
<td>EI</td>
<td>47.6/98</td>
<td>40.0/92</td>
<td>37.3/89</td>
<td>33.7/85</td>
<td>44.8/90</td>
<td>48.4/95</td>
</tr>
<tr>
<td>PO</td>
<td>56.9/92</td>
<td>46.6/87</td>
<td>41.0/82</td>
<td>38.2/77</td>
<td>53.9/83</td>
<td>58.5/88</td>
</tr>
<tr>
<td>SV</td>
<td>63.7/98</td>
<td>60.0/94</td>
<td>51.9/88</td>
<td>48.2/82</td>
<td>64.8/90</td>
<td>69.3/95</td>
</tr>
<tr>
<td>FO</td>
<td>47.8/92</td>
<td>39.2/86</td>
<td>35.7/81</td>
<td>32.1/76</td>
<td>43.8/82</td>
<td>48.4/87</td>
</tr>
<tr>
<td>EI</td>
<td>37.8/85</td>
<td>20.8/70</td>
<td>19.1/66</td>
<td>17.4/62</td>
<td>27.6/70</td>
<td>32.2/75</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Attribute</th>
<th>FSRPN</th>
<th>DAFNet</th>
<th>mfDimp</th>
<th>ADRNet</th>
<th>HMFT-LT</th>
<th>USTrack</th>
</tr>
</thead>
<tbody>
<tr>
<td>EI</td>
<td>32.8/71</td>
<td>20.8/55</td>
<td>13.1/31</td>
<td>20.0/44</td>
<td>47.3/90</td>
<td>67.0/96</td>
</tr>
<tr>
<td>PO</td>
<td>26.2/22</td>
<td>25.4/18</td>
<td>24.6/17</td>
<td>20.0/14</td>
<td>29.0/26</td>
<td>36.2/31</td>
</tr>
<tr>
<td>PO</td>
<td>33.0/27</td>
<td>23.0/14</td>
<td>20.3/13</td>
<td>19.6/13</td>
<td>35.0/27</td>
<td>42.0/32</td>
</tr>
<tr>
<td>SV</td>
<td>37.7/32</td>
<td>32.0/27</td>
<td>21.7/20</td>
<td>19.1/16</td>
<td>35.0/27</td>
<td>42.0/32</td>
</tr>
<tr>
<td>TC</td>
<td>30.4/22</td>
<td>20.3/13</td>
<td>20.0/13</td>
<td>19.6/13</td>
<td>35.0/27</td>
<td>42.0/32</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model</th>
<th>PR</th>
<th>SR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single embedding layer</td>
<td>85.6</td>
<td>63.2</td>
</tr>
<tr>
<td>Dual embedding layer (Ours)</td>
<td><strong>87.4</strong></td>
<td><strong>65.8</strong></td>
</tr>
</tbody>
</table>

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%/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 experi-

Table 5: Comparison with different prediction head structures.

<table>
<thead>
<tr>
<th>Model</th>
<th>PR</th>
<th>SR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Based on RGB search region</td>
<td>86.2</td>
<td>64.2</td>
</tr>
<tr>
<td>Based on Thermal search region</td>
<td>86.3</td>
<td>64.7</td>
</tr>
<tr>
<td>Based on weighted concatenation</td>
<td>86.8</td>
<td>64.2</td>
</tr>
<tr>
<td>Based on weighted summation</td>
<td>86.1</td>
<td>63.9</td>
</tr>
<tr>
<td>Dual prediction head with selection (Ours)</td>
<td><strong>87.4</strong></td>
<td><strong>65.8</strong></td>
</tr>
</tbody>
</table>
ments 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 VTUA V 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


