STIFormer: RGB-T Tracking via Spatial-Temporal Interaction Transformer

Boyue Xu[a], Yaqun Fang[a], Ruichao Hou[a,*] and Tongwei Ren[a]

[a]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210008, China

ARTICLE INFO

Keywords:
RGB-T tracking
transformer
spatial-temporal
attention mechanism

ABSTRACT

Existing RGB-Thermal (RGB-T) trackers integrate the RGB and thermal modalities by using cross-attention and estimate the object position by computing the correlation between a single template and the search region. However, many trackers yield unsatisfactory performance due to their disregard for inter-frame cues between modalities and dynamic changes in the dominant modality. To address this issue, we propose a novel **S**patial-**T**emporal **I**nteraction Trans**former**, called **STIFormer**, which effectively merges multi-modal features from both spatial and temporal domains, enhancing the robustness of RGB-T tracking. In particular, a spatial-temporal feature representation module is proposed to facilitate inter-frame information exchange through token propagation, which encodes features from multi-frames and a temporal token. In addition, a token-guided mixed attention fusion module is proposed to fuse the frame features and token features from different modalities. Extensive experiments demonstrate that our proposed method achieves state-of-the-art performance on public RGB-T benchmarks. The project page is available at: https://github.com/xuboyue1999/STIFormer.

1. Introduction

Visual object tracking (VOT) aims to continuously provide the position of a specified target within a given sequence [1]. It has wide applications in fields such as autonomous driving, robotics, embodied artificial intelligence, and human-computer interaction [2–8].

To overcome the inherent limitations of single-modality sensors, RGB-Thermal (RGB-T) object tracking has emerged as a promising solution that leverages the strengths of both RGB and thermal cues to enhance all-weather tracking performance. RGB data provides rich detail and texture information, while thermal data offers heat source data from the target surface that is visible under various conditions. Combining these two modalities effectively improves the robustness of object tracking, particularly in challenging environments [9–12].

Current RGB-T trackers generally extend existing RGB trackers by incorporating an additional thermal branch and employing feature fusion methods to combine the dual-modal features. To enhance tracking stability in complex scenarios, some approaches also implement template update strategies [13, 14], as shown in Figure 1(a). Specifically, the backbone is utilized to extract features from different modalities, which are subsequently fused and employed to estimate position via the prediction head. Ultimately, the template is updated based on the prediction results. However, this approach has certain limitations. While template update strategies can improve tracking stability in complex environments to some extent, they require the design of intricate update mechanisms. Additionally, these methods often struggle to leverage temporal information, making it challenging to correlate information across consecutive frames during tracking.

rchou@nju.edu.cn (R. Hou)

To address this issue, we propose a novel RGB-T tracking framework based on the Spatial-Temporal Interaction Transformer (STIFormer), as illustrated in Figure 1(b). Building on the previous work in RGB single-modality tracking [15, 16] that redefines tracking as a token propagation task, we introduce an RGB-T tracker that incorporates multi-frame templates to capture more dynamic information from inter-frame sequences compared to traditional single-frame templates. Meanwhile, a token is propagated across frames, enabling the transmission of discriminative features to subsequent frames, thereby enhancing tracking performance. Furthermore, for RGB-T multi-modal tracking, we propose a token-guided mixed attention fusion module that leverages temporal cues to more effectively extract dual-modal information while maintaining consistency with temporal dynamics.

Specifically, we extract RGB and thermal features separately using a spatial-temporal feature representation module, where each modality consists of frame features and token features. Subsequently, we employ a token-guided mixed attention fusion module to facilitate the multi-modal fusion. Afterward, prediction results are obtained through multi-branch prediction heads, while the token information is propagated to the next frame as spatial-temporal information for predicting subsequent frames. Experimental results on RGBT234 and LasHeR datasets demonstrate that our method outperforms competitive RGB-T tracking approaches.

The main contributions of this paper are summarized as follows:

- We propose a novel RGB-T tracking framework based on spatial-temporal interaction learning, which uses the multi-frame template and token propagation to construct a comprehensive spatial-temporal feature representation.

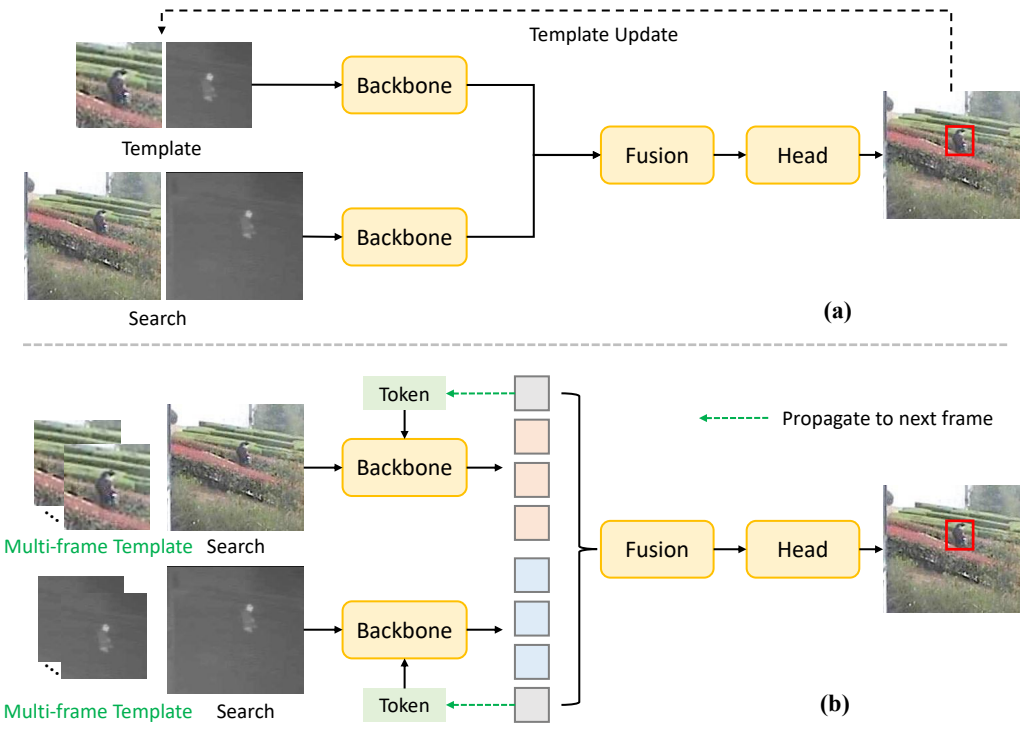

Fig. 1: Comparison of RGB-T tracking frameworks. (a) Mainstream RGB-T tracking framework with template update strategies. (b) Our framework with token propagation. The difference between the two frameworks is highlighted in green.

- We design a token-guided mixed attention fusion module that leverages spatial-temporal information to facilitate modal fusion and inter-frame interaction.

2. Related Work

2.1. RGB-T Tracking

Recently, researchers have been focusing on incorporating modal fusion methods into RGB object tracking to develop RGB-T trackers, which primarily consist of two mainstream paradigms. One category of methods is based on MDNet [17, 18], which follows the concept of tracking by detection. For instance, Lu *et al.* [19] proposed a duality-gated mutual condition network to comprehensively exploit the complementary information from dual modalities for guiding modal fusion. Hou *et al.* [20] proposed a RGB-T tracker called MIRNet, which leverages cross-attention to integrate multi-modal features and incorporates a refinement module to effectively enhance tracking accuracy. Xiao *et al.* [21] developed APFNet by generating attribute branches for different challenges in RGB-T tracking, thereby effectively improving tracking capabilities in complex scenarios. Liu *et al.* [22] explored dual-modality feature interaction and decoupling, effectively distinguishing targets in various challenging scenarios, and improving robustness.

Another category of methods employs the Siamese network [23, 24] to calculate the similarity between the template and search region. The region with the highest similarity is identified as the target, thereby enabling efficient tracking. For instance, Peng *et al.* [25] utilized the Siamese network to fuse infrared and RGB features for RGB-T tracking. Feng *et al.* [26] adjust fusion weights flexibly to enhance the feature representation and consequently bolster the robustness of the tracker. Hou *et al.* [13] utilized the Transformer to capture the global correlation and generated robust multi-modal feature representation to boost the tracking performance. Feng *et al.* [27] proposed a multi-layer attention aggregation network to achieve decision-level fusion for RGB-T tracking.

Existing RGB-T trackers have achieved satisfactory results in feature fusion and can handle various complex scenarios. However, most of them overlook the importance of temporal information in tracking tasks. Therefore, there is still room to explore cross-modal interaction and inter-frame cue mining to further enhance the robustness of RGB-T tracking.

2.2. Transformer-based Tracking

Transformer architecture has demonstrated excellent performance in handling global features, prompting researchers to incorporate the Transformer into visual object tracking to enhance the ability to process sequential global information [28]. Vaswani *et al.* [29] first proposed the Transformer, which leverages the attention mechanism to establish dependencies between each element in the sequence. Recent works have incorporated the Transformer architecture into visual object tracking. For example, Meinhardt *et al.* [30] designed the transformer-based tracking paradigm and proposed an end-to-end tracker. Lin *et al.* [31] explored the potential of the Transformer and proposed a fully-attention tracking method. Cui *et al.* [32, 33] proposed a robust Mixformer that

utilized Transformer to unify the feature extraction and aggregation, streamlining the tracking process and enhancing the tracking performance. Ye *et al.* [34] proposed OSTrack, which pioneered a one-stage one-stream transformer-based tracker. Zheng *et al.* [16] explored the importance of temporal cues in the tracking process and proposed ODTrack to correlate temporal information in adjacent frames, thus achieving outstanding performance.

Transformers have demonstrated excellent performance in visual object tracking by effectively associating global information and processing feature cues. Their powerful ability to handle global temporal information also inspires us to further explore the application of transformers in RGB-T tracking and the processing of temporal information within tracking tasks.

2.3. Temporal Modeling for Visual Object Tracking

It is widely acknowledged that temporal cues are pivotal for robust visual tracking, and strengthening context propagation over time remains a core research thread. A first line of work centers on templateupdate schemes, where the initial target template is progressively adapted. For example, Stark [35] fuses the initial template with online observations to realize adaptive updates, while Yang et al. [36] employ a multi-frame template pool to select stronger templates and alleviate the brittleness of single-frame references. Although such strategies enhance robustness, they largely retain a frame-by-frame template matching view of tracking and therefore under-utilize long-range temporal dependencies.

A second line of work explicitly propagates temporal context across frames. TCTrack [37] passes template cues between adjacent frames to refine subsequent feature extraction; ODTrack [16] injects global tokens into attention to improve propagation efficiency; ASTMT [38] extends propagation to infrared tracking; and SeqTrack [39, 40] processes entire sequences in a global manner. While these methods move beyond local template updates, many still treat temporal information as an auxiliary signal, without tightly coupling it with spatial reasoning or cross-modal fusion.

Within RGB-T tracking, a number of approaches adopt template update as the primary temporal mechanism. SDSTrack [41] selects updated templates based on confidence scores; TaTrack [14] introduces an online template that evolves with the target during tracking; and MMSTC [42] maintains a template pool, using multiple references concurrently to verify the target. STTrack [43] further incorporates position-based prediction from motion cues. These designs increase robustness by refreshing appearance, yet they still conceptualize RGB-T tracking mainly as per-frame matching, with temporal signals not deeply integrated into spatio-temporal and cross-modal interactions. STMT [44], for instance, forwards temporal cues along the sequence but treats time largely in isolation, limiting its ability to align temporal information with spatial structure and modality-specific features.

In contrast, STIFormer performs sequence-wide temporal token propagation and introduces token-guided mixed attention to inform RGBT fusion. This design jointly reasons over space, time, and modality, enabling richer temporal context to directly shape cross-modal interactions rather than merely updating templates.

3. Method

3.1. Overview

The framework of STIFormer is illustrated in Figure 2, it mainly consists of three modules: spatial-temporal feature representation, token-guided mixed attention fusion, and prediction head. Specifically, the RGB and thermal sequences are encoded separately in the spatial-temporal feature representation module, which is a typical symmetric dual-stream architecture. The encoder is utilized to extract modality features from search frames, multi-template frames, and an initially empty token. In the token-guided mixed attention fusion module, the modality features are divided into frame features and token features, which are then fused using the mixed attention fusion module to generate robust fused features. Furthermore, the temporal cues stored in the token are propagated to subsequent search frames. Finally, the output features obtained from the fusion module are subsequently fed into the prediction head to produce the tracking results, which consists of a classification branch location branch, and regression branch, and we integrate the results into the token representation.

3.2. Question Formulation

Most of the current tracking methods rely on image-pair matching to localize the target [16]. Given a pair of video frames, namely the template frame $R \in R^{3\times H_t\times W_t}$ and the search frame $S \in R^{3\times H_s\times W_s}$, the visual tracker Ψ is represented as $B \leftarrow \Psi : \{R, S\}$, where B represents the predicted bounding box coordinates of the current search window, Ψ is a Transformer-based tracker consisting of a backbone network and prediction head network. Specifically, the Transformer-based tracker takes a sequence of non-overlapping image patches (each patch has a resolution of $p \times p$) as input. A 2D template-search image pair needs to pass through a patch embedding layer to generate multiple 1D image token sequences $f_t \in R^{D\times N_t}$, $f_s \in R^{D\times N_s}$, where D is the token dimension, $N_t = \frac{H_t W_t}{p^2}$, $N_t = \frac{H_s W_s}{p^2}$. Subsequently, these 1D image tokens are concatenated and fed into an encoder with L layers of Transformers for feature extraction and relationship modeling.

We encode each modality separately and extend the input of the tracking framework from image pairs to multi-frames for temporal modeling. We introduce a temporal token for each modality to propagate temporal information related to the sequence. Then, we use the token-guided mixed attention fusion module to encode the information of different modalities and produce the prediction result. The overall process

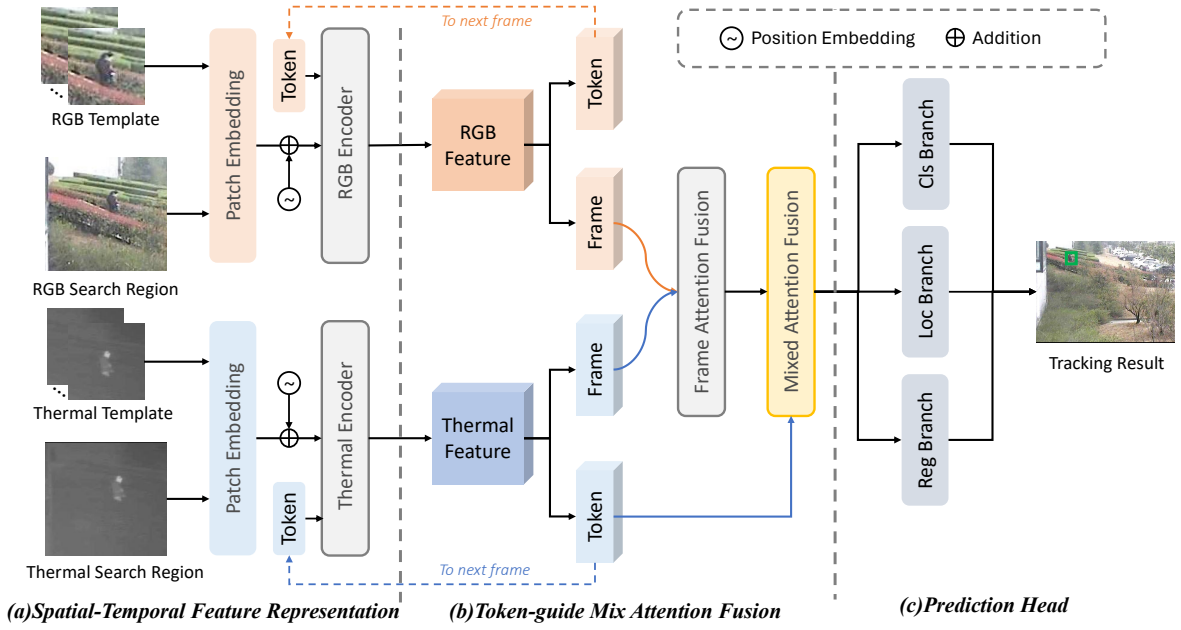

Fig. 2: The framework of STIFormer, which includes spatial-temporal feature representation, token-guided mixed attention fusion, and prediction head.

is described as follows:

$$B \leftarrow \mathcal{H} : \{\mathcal{F} : \{\Psi^r : \{R_1^r, R_2^r, ..., R_k^r, S_1^r, S_2^r, ..., S_n^r, T^r\}, \Psi^t : \{R_1^t, R_2^t, ..., R_k^t, S_1^t, S_2^t, ..., S_n^t, T^t\}\}\}, \tag{1}$$

where B represents the predicted box coordinates of the current search box, $\mathcal{H}$ represents the prediction head, $\mathcal{F}$ represents the token-guided mixed attention fusion module, Ψ^r represents the RGB encoder, Ψ^t represents the thermal encoder, $R_1^{r/t}, R_2^{r/t}, ..., R_k^{r/t}$ represent template frames of length k, $S_1^{r/t}, S_2^{r/t}, ..., S_n^{r/t}$ represent search frames of length n, $T^{r/t}$ denotes the time token.

3.3. Spatial-Temporal Feature Representation

To fully exploit the spatial-temporal feature representations, in this module, we leverage multi-frame templates and spatio-temporal token passing to propagate the spatial-temporal information across adjacent frames. We employ the Transformer to extract features separately from the RGB and thermal modalities.

Inspired by the well-designed feature encoding of RGB single modality [15, 16], we employ a shared-weighted transformer encoder to extract RGB and thermal features, respectively. For the t-th frame f_t, the computation process can be expressed as:

$$\begin{aligned} f_t &= Attn(Concat[R_1, R_2, ..., R_k, S_t, T_t]) \\ &= \sum_{st} v_{st} \cdot \frac{\exp\langle q_{st}, k_{st}\rangle}{\sum_{st} \exp\langle q_{st}, k_{st}\rangle}, \end{aligned} \tag{2}$$

where T_t is the time token sequence of the t-th frame, $Attn$ represents attention mechanism, $Concat[\cdot]$ denotes concatenation between elements, $[R_1, R_2, ..., R_k]$ is the sequence of template frames, S_t is the t-th search frame, q_{st}, k_{st}, v_{st} are linear projections of the concatenated features, respectively. Each video frame contains a token used to store the target information of the frame sequence. Once the target information is extracted via the temporal token, the token is propagated from frame t to frame $t+1$. We maintain a single temporal token $T_t \in \mathbb{R}^{1\times d}$ (with d matching the backbone hidden size), which is updated at each step and then used to condition feature extraction and cross-modal fusion at the next frame. At the start of a sequence, the token is initialized as a blank vector $T_{\text{empty}} \in \mathbb{R}^{1\times d}$. For propagation from t to $t+1$, we first form a residual seed by adding the current token to an empty token of the same shape, and then refine it with attention using the features of the $(t+1)$-th frame:

$$\begin{aligned} \tilde{T}_{t+1} &= T_t + T_{\text{empty}}, \\ f_{t+1}, T_{t+1} &= \text{Attn}\Big(\text{Concat}\big[R_1, R_2, \dots, R_k, S_{t+1}, T_{t+1}\big]\Big). \end{aligned} \tag{3}$$

where $R_1, \dots, R_k$ denote the reference/template features, and S_{t+1} denotes the search frame features at time $t+1$. f_{t+1} denotes the feature representation from the $(t+1)$-th frame.

In this way, the inference at frame $t+1$ is explicitly guided by the temporal token from frame t, leveraging accumulated historical cues to inform current predictions. The token T_{t+1} is subsequently propagated to the next step and updated recurrently until the end of the sequence.

3.4. Token-guided Mixed Attention Fusion

To fully exploit visual frame features and temporal information between frames, we propose a token-guided mixed attention fusion module that effectively integrates spatial-temporal cues for feature fusion between frames of different modalities.

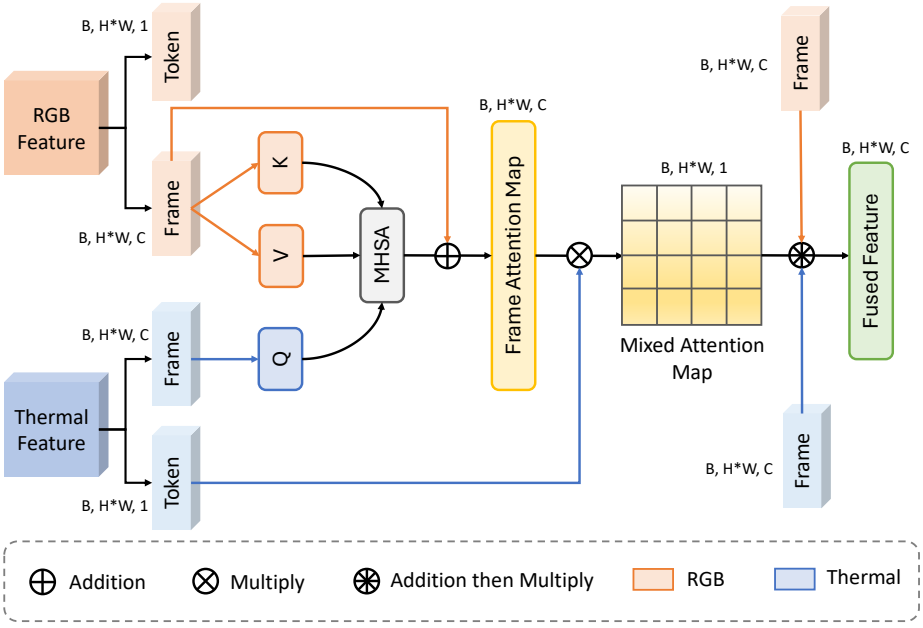

Fig. 3: The detailed design of the token-guided mixed attention fusion module.

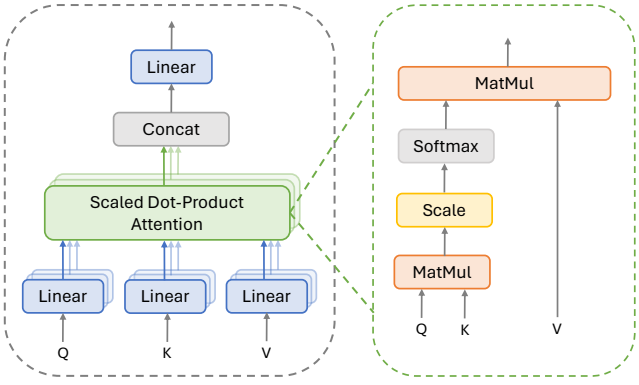

Fig. 4: Left:Multi-head Self Attention; Right:Scaled Dot-Product Attention

The detailed design of the token-guided mixed attention fusion module is illustrated in Fig. 3. The single-modality features are generated by their respective encoders. The modality features include frame features and token features, corresponding to the visual-spatial features and temporal token features from frames, which is formulated as:

$$\begin{aligned} f_{rgb} &= [f_T^r, f_V^r], \\ f_t &= [f_T^t, f_V^t], \end{aligned} \tag{4}$$

where f_{rgb} and f_t represent the features encoded by the RGB and thermal encoders, $f_T^{r/t}$, $f_V^{r/t}$ represent token features and frame features, respectively.

Firstly, we employ a multi-head self-attention mechanism [29] to capture the cross-modal correlation between RGB-T modalities. The multi-head self-attention mechanism is shown in Fig. 4. Specifically, the cross-modal attention mechanism is expressed by:

$$G_{cross} = MHSA(f_V^t, f_V^r, f_V^r), \tag{5}$$

where $MHSA(\cdot)$ represents the multi-head self-attention mechanism, Q is set to f_V^t, K and V are both set to f_V^r, respectively. The input features are reshaped and divided into n heads. The attention weights are calculated as follows:

$$G_{head}^i = \text{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) \cdot V_i, \tag{6}$$

where Q_i represents the query, K_i represents the key, V_i represents the value, i denotes the i-th head, G_{head}^i represents the attention of the i-th head. Then, the attention generated by each head is concatenated and reshaped to obtain the final attention. Equation (5) can be further formulated as:

$$G_{cross} = \text{Concat}[G_{head}^1, G_{head}^2, \ldots, G_{head}^n] \cdot \omega_0, \tag{7}$$

where G_{cross} represents the cross attention, n is the total number of heads, $Concat(\cdot)$ denotes the concatenation operation, G_{head}^i represents the attention of each head, ω_0 is the final linear transformation matrix. After the multi-head self-attention module produces its output, it is integrated back into the original RGB frame feature using a residual connection to generate the frame attention map, which can be formulated as follows:

$$G_{frame} = f_V^r + G_{cross}, \tag{8}$$

where G_{frame} means frame attention map, f_V^r represents the RGB frame feature, and G_{cross} represents the cross attention computed by equation (7). The updated frame features f_V^r

are then multiplied with the token features f_T^t from the thermal modality to obtain the mixed attention fused features, which are expressed by:

$$G_{mix} = G_{frame} \cdot f_T^t, \tag{9}$$

where G_{mix} denotes the mixed attention map, G_{frame} represents frame attention map computed by equation(8), and f_T^t represents the token feature from the thermal modality, respectively.

Finally, we add the frame features from each modality and multiply them with the mixed attention map to obtain the final fused feature F_f, which can be formulated as:

$$F_f = G_{mix} \cdot (f_V^r + f_V^t), \tag{10}$$

where F_f represents the final fused feature, f_V^r and f_V^t mean frame feature from RGB and thermal modalities, respectively.

3.5. Prediction Head and Loss Function

STIformer utilizes classification and bounding box regression heads to achieve tracking results. Specifically, triplet prediction branches are employed to obtain the classification score map, bounding box size, and offset for prediction. Our method employs focal loss as the classification loss $\mathcal{L}_{cls}$ and utilizes $\mathcal{L}_1$ loss and $GIoU$ loss as the regression losses [45]. Focal Loss effectively addresses class imbalance by adaptively scaling the loss to focus more on hard-to-classify samples, which is more suitable for the dataset with a long-tail distribution. $\mathcal{L}_1$ Loss provides a robust optimization objective less sensitive to outliers, while $GIoU$ Loss encourages better spatial alignment between the predicted and ground-truth bounding boxes. The synergistic combination of $\mathcal{L}_1$ Loss and $GIoU$ Loss leads to accurate bounding box regression and improved spatial overlap.

The classification loss is formulated as:

$$\mathcal{L}_{cls} = -\sum_t \alpha \left(1 - p_t\right)^{\gamma} log(p_t), \tag{11}$$

where t represents the $t-th$ samples, α_t is the weight coefficient, p_t indicates the probability belonging to the foreground. The regression loss is defined as:

$$\mathcal{L}_{reg} = \sum_t \left(\lambda_1 \mathcal{L}_1\left(b_t, \hat{b}_t\right) + \lambda_2 \mathcal{L}_{GIoU}\left(b_t, \hat{b}_t\right)\right), \tag{12}$$

where λ_1 and λ_2 are regularization parameters. b_t means the $t-th$ predicted bounding box, $\hat{b}_t$ means the corresponding ground truth box.

The overall loss can be expressed as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{reg}, \tag{13}$$

where $\mathcal{L}_{cls}$ represents classification loss and $\mathcal{L}_{reg}$ means regression loss.

Table 1
Comparison between the proposed method and the state-of-the-art trackers on RGB-T datasets. The best results are highlighted in **bold** and the trackers that use temporal information are marked in *. The performance is evaluated in terms of Precision Rate (PR) and Success Rate (SR).

Methods	Years	RGBT234		LasHeR		Param	FPS
		PR	SR	PR	SR		
TBSI* [46]	CVPR23	0.871	0.637	0.692	0.556	350	36
MTNet* [13]	ICME23	0.850	0.619	0.608	0.474	-	55
CAT++ [22]	TIP24	0.840	0.592	0.509	0.356	90	14
ProTrack [47]	MM23	0.795	0.599	0.538	0.420	-	30
ViPT [48]	CVPR23	0.835	0.617	0.651	0.525	93	25
GMMT [49]	AAAI24	0.879	0.647	0.707	0.566	-	-
STMT* [44]	TCSVT24	0.865	0.638	0.674	0.537	-	39
SDSTrack* [41]	CVPR24	0.848	0.625	0.665	0.531	107.8	21
MMSTC* [42]	TIP24	0.898	0.673	0.723	0.574	-	-
UN-Track [50]	CVPR24	0.837	0.618	0.667	0.536	92.1	-
OneTracker [51]	CVPR24	0.857	0.642	0.672	0.538	99.8	-
TaTrack* [14]	AAAI24	0.872	0.644	0.702	0.561	-	26
BaT [52]	AAAI24	0.868	0.641	0.702	0.563	-	-
ODTrack* [16]	AAAI24	0.867	0.648	0.702	0.555	98.3	25
IPL [53]	IJCV25	0.883	0.657	0.694	0.553	-	-
CMDTrack [54]	TPAMI25	0.859	0.618	0.688	0.566	-	67
STTrack* [43]	SJ25	0.888	0.671	0.726	0.579	92.8	32
CAFormer [55]	AAAI25	0.883	0.664	0.700	0.556	-	84
Ours	-	**0.915**	**0.683**	**0.738**	**0.583**	101.6	24

4. Experiment

4.1. Dataset and Evaluation

In this paper, we conduct comparative experiments with state-of-the-art trackers on two large-scale RGB-T benchmarks, namely RGBT234 [45] and LasHeR [56]. RGBT234 consists of 234 pairs of precisely aligned RGB-T video sequences and 12 annotated attributes, encompassing approximately 233.4K frames. LasHeR is the latest RGB-T tracking dataset, which consists of 1224 RGB-T videos captured from diverse views and scenarios with a total of 734.8K frames.

Corresponding to previous works [56], we adopt two widely used metrics, namely Precision Rate (PR) and Success Rate (SR), to evaluate the tracking performance, and the threshold of center location error is set to 20 pixels. PR represents the ratio of frames f_p with center error smaller than a threshold to the total number of frames N, and it can be expressed as:

$$PR = \frac{N_p}{N} \times 100\%, \tag{14}$$

where N_p represents the number of frames with center error smaller than a threshold; N represents the total number of the sequence. SR is defined as the ratio of frames s_p with IoU exceeding a certain threshold to the total number of frames N, and it can be calculated as:

$$SR = \frac{N_s}{N} \times 100\%, \tag{15}$$

where N_s represents the number of frames with IoU exceeding a certain threshold.

Table 2
Comparison results of our method against the state-of-the-art trackers. Attribute-based and overall performance are evaluated by PR/SR scores(%) and are produced on RGBT234 with DMCNet [19], MIRNet [20], HMFT [57], CAT++ [22], MTNet [13] and ViPT [48]. The best results are highlighted in **Bold**.

	DMCNet [19] TNNLS22	MIRNet [20] ICME22	HMFT [57] CVPR22	MTNet [13] ICME23	ViPT [48] CVPR23	CAT++ [22] TIP24	**Ours**
NO	92.3/67.1	95.4/72.4	90.9/67.4	91.0/67.8	92.4/71.0	94.9/67.6	**98.5/75.2**
PO	84.2/63.1	89.5/62.7	85.7/62.1	88.7/64.8	85.4/63.0	88.9/62.7	**93.7/70.4**
HO	74.5/52.1	71.0/49.0	66.4/46.9	78.6/56.3	77.7/56.2	74.4/52.2	**86.2/63.0**
LI	85.3/58.7	83.4/57.5	83.3/59.1	83.3/59.5	81.0/58.4	80.9/55.5	**94.4/70.0**
LR	85.4/57.9	83.9/56.3	76.3/57.1	80.4/55.4	83.0/59.3	86.0/58.6	**90.2/63.4**
TC	87.2/61.2	81.1/59.1	80.6/50.4	86.1/61.6	83.0/62.2	85.0/61.5	**91.3/67.8**
DEF	77.9/56.5	77.8/58.1	77.6/57.9	84.7/64.0	81.7/62.2	80.4/56.7	**89.1/67.9**
FM	80.0/52.4	68.3/47.1	65.9/46.9	79.2/58.0	80.2/58.5	83.3/54.9	**87.1/63.5**
SV	82.7/59.8	83.2/61.9	80.0/59.2	89.0/66.1	83.8/63.0	84.8/59.3	**92.9/70.4**
MB	77.3/55.9	74.6/54.6	70.6/50.9	83.4/61.6	83.2/62.5	76.5/55.4	**93.4/69.9**
CM	80.1/57.6	76.4/55.4	77.9/56.2	86.0/63.4	83.0/62.0	77.9/56.2	**91.1/68.5**
BC	83.8/55.9	78.9/51.7	73.8/49.8	74.9/50.8	79.6/55.6	82.0/55.3	**89.0/62.6**
ALL	83.9/59.3	81.6/58.9	78.8/56.8	85.0/61.9	83.5/61.7	84.0/59.2	**91.5/68.3**

4.2. Implementation Details

Our method utilizes the ViT-Base model [16] as the visual encoder, with its parameters initialized using the pretraining weights from MAE [58]. STIFormer takes video sequences comprising three reference frames with the size of 192×192 pixels and two search frames with the size of 384×384 pixels as the input. The AdamW optimizer is employed to optimize the network parameters, with an initial learning rate set to 1e-5 and a weight decay of 1e-4. The model is trained for 60 epochs with 10,000 samples per epoch on the LasHeR training dataset. After completing 40 epochs, we reduce the learning rate by a factor of 10.

Our model is trained on a server with two 24 GB NVIDIA RTX 3090 GPUs, and the batch size is set to 4. The training process is end-to-end, with an overall training time of approximately 30 hours and one GPU memory consumption of around 18 GB. The overall computational cost of the tracker, measured in FLOPs (Floating-Point Operations), is 147.9G. The total number of model parameters is 101.6M. When tested on a single NVIDIA RTX 3090 GPU, the average overall latency is 41.77 ms, and the achieved frame rate is 24 FPS.

4.3. Comparison with State-of-the-Arts

Quantitative Comparison. To evaluate the superiority of STIFormer, we conduct a comparative analysis with several competing trackers, including TBSI [46], MTNet [13], CAT++ [22], ProTrack [47], ViPT [48], GMMT [49], STMT [44], SDSTrack [41], MMSTC [42], UNTrack [50], OneTracker [51], TaTrack [14], BaT [52], IPL [53], ODTrack [59], STTrack [43] and CMDTrack [54]. Notably, TBSI, MTNet, STMT, SDSTrack, MMSTC, TaTrack, ODTrack, and STTrack use temporal information.

Experiments are conducted on the two most widely used RGBT tracking benchmarks, LasHeR and RGBT234 , as shown in Table 1. Our method outperforms all existing approaches on both datasets. Compared with the recent state-of-the-art CAFormer, STIFormer improves PR and SR on RGBT234 by 3.2% and 1.9%, respectively, and on LasHeR by 3.8% and 2.7%. Notably, relative to SDSTrack, which also exploits temporal information, STIFormer achieves gains of 6.8%/5.8% (PR/SR) on RGBT234 and 7.3%/5.2% on LasHeR. We attribute these improvements to our temporal modeling strategy, which aggregates sequence-wide temporal cues and uses token-guided mixed attention to inform cross-modal fusion, whereas SDSTrack primarily relies on template updating. In addition, STIFormer maintains competitive tracking speed and a parameter count comparable to prior methods, indicating that the performance gains arise from more effective modeling rather than from a heavier architecture.

Attribute Analysis. To comprehensively evaluate the performance under different challenging attributes, we further evaluate the trackers on the 12 attributes of the RGBT234 dataset, such as NO (No Occlusion), PO (Partial Occlusion), HO (Heavy Occlusion), LI (Low Illumination), TC (Thermal Crossover), LR (Low Resolution), DEF (Deformation), FM (Fast Motion), SV (Scale Variation), MB (Motion Blur), CM (Camera Moving) and BC (Background Clutter).

The attribute-based comparison results are shown in Table 2. Experimental results show that our method achieves the best tracking accuracy and robustness in all challenging scenarios. Especially in o PO and HO scenarios,the proposed method significantly outperforms the second-best approach in these cases, with improvements of 4.2% in PR and 6.9% in SR for PO, and 8.5% and 6.7% respectively for HO. This improvement is primarily due to the effective transmission of temporal information, allowing the network to rely on prior tracking cues to infer the target's position even when it is occluded. The experimental results demonstrate that STIFormer exhibits generalization and robustness when facing complex and various scenarios.

Qualitative Comparison. To provide a more intuitive qualitative comparison of STIFormer, we conduct qualitative comparisons between our method and representative trackers on the RGBT234. The qualitative comparison is

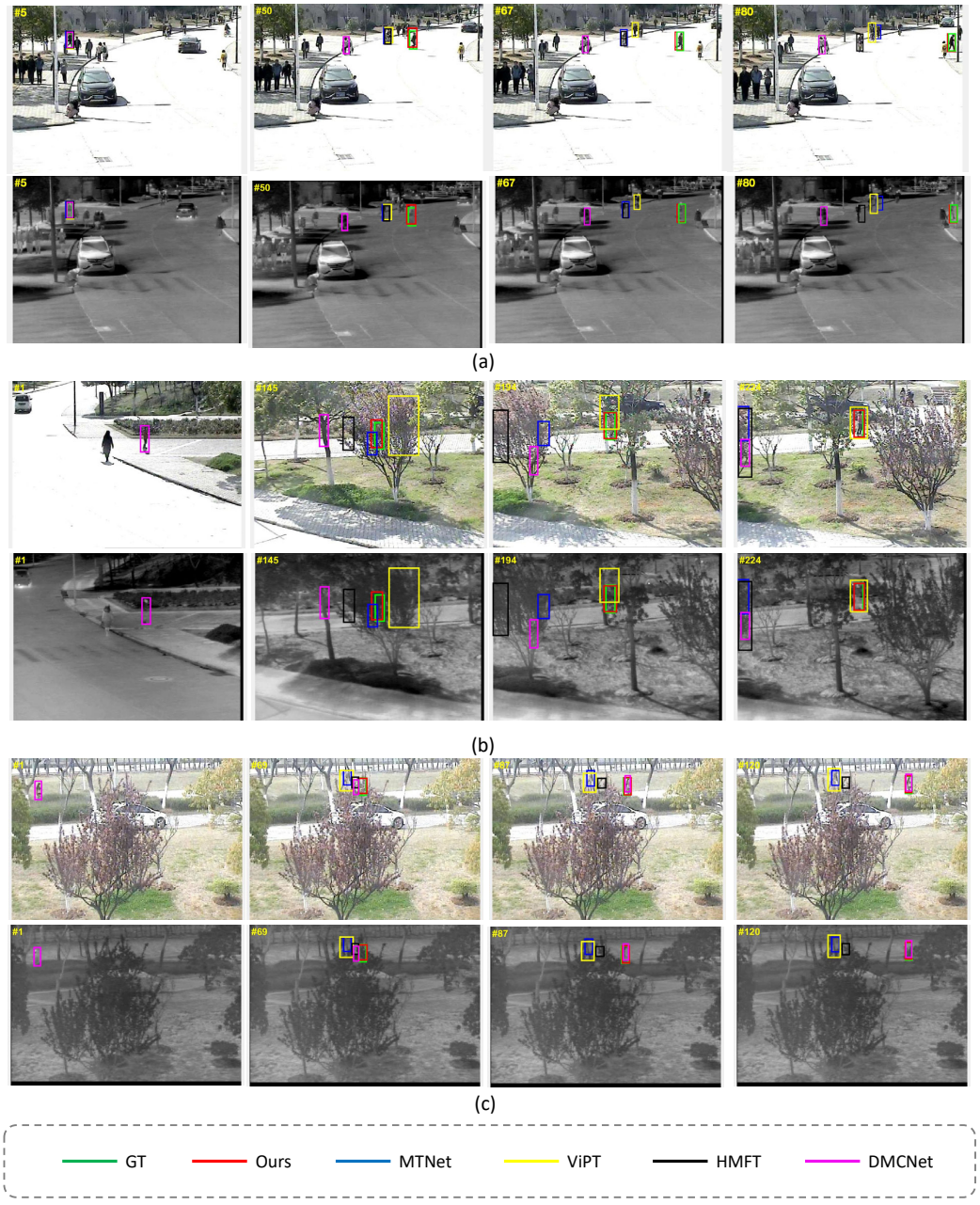

Fig. 5: Visualization results of (a)run1 sequence, (b)blancebike and (c)womanfaraway sequence, compared with MTNet [13], ViPT [48], HMFT [57] and DMCNet [19].

shown in Fig. 5. We observe that STIFormer achieves more robust tracking performance in complex scenarios.

For example, as shown in Fig. 5(a), the run1 sequence includes challenging attributes such as FM, MB and PO. Based on the visual results, we observe that even under conditions of fast motion and image blurring, the proposed method shows exceptional stability in continuously tracking a specific pedestrian, as shown in Fig. 5(b). Thanks to spatial-temporal interaction learning, our tracker performs well when encountering challenging attributes (*e.g.*, HO, FM, and CM). It can effectively utilize the correlations between adjacent frames to maintain stable tracking of the target even when the target is occluded. As shown in Fig. 5(c). In complex backgrounds, the proposed method maintains stable tracking, primarily due to the token-guided mixed attention fusion module that effectively extracts and integrates muti-modal features. This enables stable tracking using the infrared modality even in environments with complicated backgrounds. Overall, visualization results highlight the effectiveness and robustness of our STIFormer.

Feature Visualization. To further illustrate the effectiveness of the proposed token-guided mix attention fusion, we perform feature visualization on the search regions. As shown in Fig 6, before applying the Token-guided Mixed

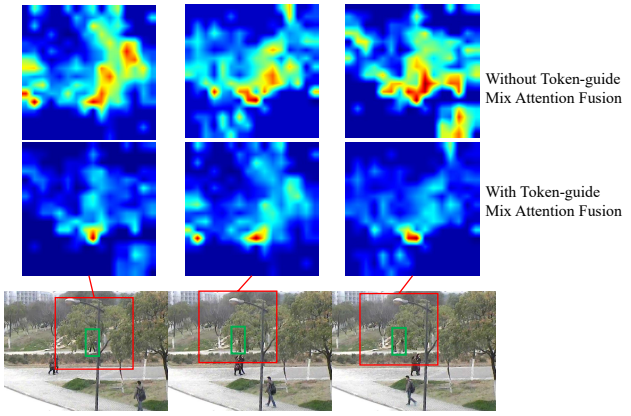

Fig. 6: The visualization of search area features: the top row shows the feature visualization before applying the Token-guided Mixed Attention Fusion, the middle row displays the features after the fusion, and the bottom row shows the original image, with the red box indicating the current search area and green box indicating the target.

Attention Fusion, the features are scattered and not focused on the target. However, the middle-row visualization results show the discriminative features become concentrated on the target by using token-guided mixed attention fusion. This improvement is mainly due to the temporal-guided feature fusion, which not only aggregates historical information across consecutive frames but also enhances the robustness of the feature representation.

Robustness Analysis. To further validate the robustness of the proposed method in complex environments, we conduct a robustness analysis, as shown in Fig. 7. The line graph below represents the deviation of the predicted bounding boxes from the ground truth during the tracking process. Generally, a deviation of less than 40 pixels is considered successful tracking. In the upper part of Fig. 7, only our method demonstrated consistent stability. Several key points are noteworthy: around frame 300, where numerous similar targets and occlusions occurred, our method experienced only brief fluctuations, whereas other methods lost the target. Thanks to the spatio-temporal information, the proposed tracker accurately maintains stable tracking among similar targets. Between the 550th and 580th frames, some methods briefly relocated the correct target but lost it again after the occlusion. In contrast, our method consistently tracked the object throughout the entire sequence. The lower part of the figure shows the robustness of the proposed method while facing occlusion. Around frame 35, the target became occluded, causing some trackers to lose the target entirely and fail to relocate it, while our method continued to track the target stably. At around frame 60, when similar targets appeared, all trackers except ours mistakenly tracked the wrong target and failed to correct this throughout the remainder of the sequence. This experiment clearly illustrates the robustness of our method in complex environments.

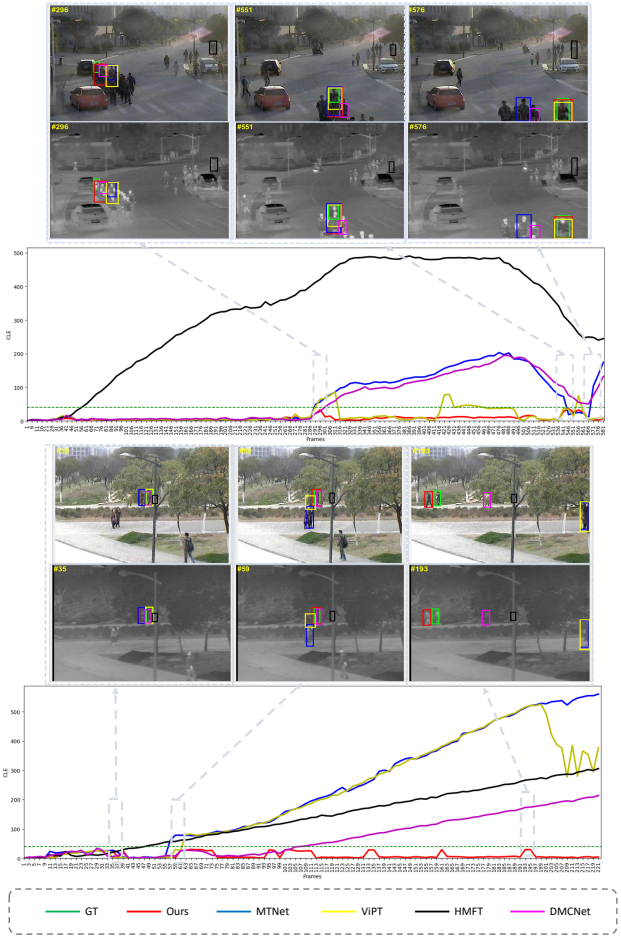

Fig. 7: Robustness analysis results of the proposed method compared with MTNet [13], ViPT [48], HMFT [57] and DMCNet [19].

Table 3
Multi-modal analysis on the RGBT234 Dataset, The best results are highlighted in **Bold**.

RGB	Thermal	PR	SR
✓	✗	81.7	60.9
✗	✓	78.4	56.4
✓	✓	**86.7**	**64.8**

4.4. Ablation Studies

Modality Analysis. To validate the benefits of complementary information in tracking performance, we conducted ablation experiments by using single-modal encoders for feature encoding on individual modalities and compared them with the method that uses feature addition with dual-modal features. The results of the modality ablation comparison on the RGBT234 dataset are shown in Table 3. It can be observed that using dual-modal features yields better tracking performance compared to using single-modal features alone.

Components Analysis. To validate the effectiveness of the proposed token-guided mixed attention fusion module, we conduct ablation experiments on different fusion

Table 4
Ablation study of mixed attention fusion module on the RGBT234 Dataset, R represents RGB and T represents thermal modality. The best results are highlighted in **Bold**.

Frame Attention	Mixed Attention	PR	SR
$[Q = R], [K, V = T]$	$[Token = R]$	82.2	60.1
$[Q = R], [K, V = T]$	$[Token = T]$	85.7	63.6
$[Q = T], [K, V = R]$	$[Token = R]$	87.3	63.2
$[Q = T], [K, V = R]$	$[Token = T]$	**91.5**	**68.3**

methods. The results of the module ablation experiment are shown in Table 4. Additionally, we perform ablation experiments on the fusion strategies of frame attention and mixed attention, respectively. To analyze the structure of frame attention fusion, we conduct experiments on different sources of Q, K and V. Moreover, we use token features from different modalities to compare the performance differences in mixed attention. The experimental results indicate that the proposed token-guided mixed attention fusion module achieves the best tracking performance, demonstrating that our method can more effectively utilize spatial-temporal information to guide multi-modal fusion. This is primarily because the thermal modality inherently contains fewer features, but it can provide additional target features to compensate for the limitations of the RGB modality. Therefore, setting the query as the thermal modality while using the key and value as the RGB modality allows for better association between the rich detail features in RGB and the complementary features in thermal. Additionally, setting the mixed token to the thermal modality can more effectively correlate temporal information across frames, helping to avoid interference from noise in the RGB modality.

Regarding model complexity and real-time performance, the proposed method comprises 101.6M parameters and achieves approximately 24 FPS on an RTX 3090 GPU. Under the same hardware and experimental settings, the baseline contains 98.3M parameters and reaches 25 FPS. Hence, our approach increases the parameter count by only +3.3M and reduces the speed by merely 1 FPS, clearly demonstrating that it preserves real-time capability with minimal computational overhead.

5. Conclusion

In this paper, we proposed STIFormer, a novel RGB-T tracking method based on spatial-temporal interaction transformer. To exploit inter-frame cues, we introduce a framework that incorporates multi-frame template and token propagation to capture comprehensive spatial-temporal feature representation. In addition, a token-guided mixed attention fusion module is proposed to effectively fuse the frame features and the token features from different modalities. Experimental results on the public RGBT234 and LasHeR datasets demonstrate the effectiveness of STIFormer, achieving the best performance across various challenging attributes.

Author statement

Funding

This work was supported by the National Natural Science Foundation of China (62072232), the Key R&D Project of Jiangsu Province (BE2022138), the Fundamental Research Funds for the Central Universities (021714380026), the Innovation Project of State Key Laboratory for Novel Software Technology Nanjing University (ZZKT2024B20), and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

Author contributions

Boyue Xu: Conceptualization, Methodology, Software, Writing original draft. Yaqun Fang: Investigation, Methodology, Validation. Ruichao Hou: Project administration, Writing original draft. Tongwei Ren: Polishing, Funding acquisition, Resource.

References

[1] R. Yao, G. Lin, S. Xia, J. Zhao, Y. Zhou, Video object segmentation and tracking: A survey, ACM Transactions on Intelligent Systems and Technology 11 (2020) 1–47.

[2] T.-X. Xu, Y.-C. Guo, Y.-K. Lai, S.-H. Zhang, Cxtrack: improving 3d point cloud tracking with contextual information, in: IEEE Conference on Computer Vision and Pattern Recognition, 2023.

[3] H. Ye, J. Zhao, Y. Pan, W. Cherr, L. He, H. Zhang, Robot person following under partial occlusion, in: IEEE International Conference on Robotics and Automation, 2023.

[4] F. Zhong, K. Wu, H. Ci, C. Wang, H. Chen, Empowering embodied visual tracking with visual foundation models and offline rl, arXiv preprint arXiv:2404.09857 (2024).

[5] L. Zhou, Z. Zhou, K. Mao, Z. He, Joint visual grounding and tracking with natural language specification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2023.

[6] B. Xu, R. Hou, J. Bei, T. Ren, G. Wu, Jointly modeling association and motion cues for robust infrared uav tracking, The Visual Computer (2024) 1–12.

[7] B. Xu, R. Hou, T. Ren, G. Wu, Rgb-d video object segmentation via enhanced multi-store feature memory, in: Proceedings of the International Conference on Multimedia Retrieval, 2024, pp. 1016–1024.

[8] Z. Ding, H. Li, R. Hou, Y. Liu, S. Xie, X modality assisting rgbt object tracking, Applied Intelligence 55 (2025) 775.

[9] Z. Yu, H. Fan, Q. Wang, Z. Li, Y. Tang, Region selective fusion network for robust rgb-t tracking, IEEE Signal Processing Letters (2023).

[10] M. Li, P. Zhang, M. Yan, H. Chen, C. Wu, Dynamic feature-memory transformer network for rgbt tracking, IEEE Sensors Journal (2023).

[11] M. Feng, J. Su, Rgbt image fusion tracking via sparse trifurcate transformer aggregation network, IEEE Transactions on Instrumentation and Measurement 73 (2024) 1–10.

[12] Y. Zhu, C. Li, X. Wang, J. Tang, Z. Huang, Rgbt tracking via progressive fusion transformer with dynamically guided learning, IEEE Transactions on Circuits and Systems for Video Technology (2024).

[13] R. Hou, B. Xu, T. Ren, G. Wu, Mtnet: learning modality-aware representation with transformer for rgbt tracking, in: IEEE International Conference on Multimedia and Expo, 2023, pp. 1163–1168.

[14] H. Wang, X. Liu, Y. Li, M. Sun, D. Yuan, J. Liu, Temporal adaptive rgbt tracking with modality prompt, in: Proceedings of the AAAI Conference on Artificial Intelligence Conference on Artificial Intelligence, 2024.

[15] X. Chen, H. Peng, D. Wang, H. Lu, H. Hu, Seqtrack: Sequence to sequence learning for visual object tracking, in: IEEE Conference on Computer Vision and Pattern Recognition, 2023, pp. 14572–14581.
[16] Y. Zheng, B. Zhong, Q. Liang, Z. Mo, S. Zhang, X. Li, Odtrack: Online dense temporal token learning for visual tracking, in: Association for the Advance of Artificial Intelligence, 2024.
[17] I. Jung, J. Son, M. Baek, B. Han, Real-time mdnet, in: Proceedings of the European conference on computer vision (European Conference on Computer Vision), 2018, pp. 83–98.
[18] H. Nam, B. Han, Learning multi-domain convolutional neural networks for visual tracking, in: IEEE conference on Computer Vision and Pattern Recognition, 2016, pp. 4293–4302.
[19] A. Lu, C. Qian, C. Li, J. Tang, L. Wang, Duality-gated mutual condition network for rgbt tracking, IEEE Transactions on Neural Networks and Learning Systems (2022) 1–14.
[20] R. Hou, T. Ren, G. Wu, Mirnet: A robust rgbt tracking jointly with multi-modal interaction and refinement, in: IEEE International Conference on Multimedia and Expo, 2022, pp. 1–6.
[21] Y. Xiao, M. Yang, C. Li, L. Liu, J. Tang, Attribute-based progressive fusion network for rgbt tracking, in: Association for the Advancement of Artificial Intelligence, 2022, pp. 2831–2838.
[22] L. Liu, C. Li, Y. Xiao, R. Ruan, M. Fan, Rgbt tracking via challenge-based appearance disentanglement and interaction, IEEE Transactions on Image Processing (2024).
[23] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, P. H. Torr, Fully-convolutional siamese networks for object tracking, in: European Conference on Computer Vision, 2016, pp. 850–865.
[24] M. Zhou, X. Zhao, F. Luo, J. Luo, H. Pu, T. Xiang, Robust rgb-t tracking via adaptive modality weight correlation filters and cross-modality learning, ACM Transactions on Multimedia Computing, Communications and Applications 20 (2023) 1–20.
[25] J. Peng, H. Zhao, Z. Hu, Y. Zhuang, B. Wang, Siamese infrared and visible light fusion network for rgb-t tracking, International Journal of Machine Learning and Cybernetics 14 (2023) 3281–3293.
[26] M. Feng, J. Su, Learning reliable modal weight with transformer for robust rgbt tracking, Knowledge-based systems 249 (2022) 108945.
[27] M. Feng, J. Su, Learning multi-layer attention aggregation siamese network for robust rgbt tracking, IEEE Transactions on Multimedia 26 (2024) 3378–3391.
[28] X. Lin, S. Sun, W. Huang, B. Sheng, P. Li, D. D. Feng, EAPT: efficient attention pyramid transformer for image processing, IEEE Transactions on Multimedia 25 (2021) 50–61.
[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017.
[30] T. Meinhardt, A. Kirillov, L. Leal-Taixe, C. Feichtenhofer, Trackformer: Multi-object tracking with transformers, in: IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 8844–8854.
[31] L. Lin, H. Fan, Z. Zhang, Y. Xu, H. Ling, Swintrack: A simple and strong baseline for transformer tracking, in: Advances in Neural Information Processing Systems, 2022, pp. 16743–16754.
[32] Y. Cui, T. Song, G. Wu, L. Wang, Mixformerv2: Efficient fully transformer tracking, in: Advances in Neural Information Processing Systems, 2024.
[33] Y. Cui, C. Jiang, L. Wang, G. Wu, Mixformer: End-to-end tracking with iterative mixed attention, in: IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 13608–13618.
[34] B. Ye, H. Chang, B. Ma, S. Shan, X. Chen, Joint feature learning and relation modeling for tracking: A one-stream framework, in: European Conference on Computer Vision, 2022.
[35] B. Yan, H. Peng, J. Fu, D. Wang, H. Lu, Learning spatio–temporal transformer for visual tracking, in: Proceedings of IEEE International Conference on Computer Vision, 2021.
[36] T. Yang, A. B. Chan, Visual tracking via dynamic memory networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (2021) 360–374.
[37] Z. Cao, Z. Huang, L. Pan, S. Zhang, Z. Liu, C. Fu, Tctrack: Temporal contexts for aerial tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.
[38] D. Yuan, X. Shu, Q. Liu, Z. He, Aligned spatial–temporal memory network for thermal infrared target tracking, IEEE Transactions on Circuits and Systems 70 (2023) 1224–1228.
[39] X. Chen, H. Peng, D. Wang, H. Lu, H. Hu, Seqtrack: Sequence-to-sequence learning for visual object tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.
[40] X. Chen, B. Kang, J. Zhu, D. Wang, H. Peng, H. Lu, Unified sequence-to-sequence learning for single- and multi-modal visual object tracking, arXiv:2404.00000 (2024).
[41] X. Hou, J. Xing, Y. Qian, Y. Guo, S. Xin, J. Chen, K. Tang, M. Wang, Z. Jiang, L. Liu, et al., Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.
[42] T. Zhang, Q. Jiao, Q. Zhang, J. Han, Exploring multi-modal spatialtemporal contexts for high-performance rgb-t tracking, IEEE Transactions on Image Processing 33 (2024) 4303–4318.
[43] D. Yuan, H. Zhang, Q. Liu, X. Chang, Z. He, Transformer-based rgbt tracking with spatio-temporal information fusion, IEEE Sensors Journal 25 (2025) 25386–25396.
[44] D. Sun, Y. Pan, A. Lu, C. Li, B. Luo, Transformer rgb-t tracking with spatio-temporal multimodal tokens, IEEE Transactions on Circuits and Systems for Video Technology 34 (2024) 12059–12072.
[45] C. Li, X. Liang, Y. Lu, N. Zhao, J. Tang, Rgb-t object tracking: Benchmark and baseline, Pattern Recognition 96 (2019) 106977.
[46] T. Hui, Z. Xun, F. Peng, J. Huang, X. Wei, X. Wei, J. Dai, J. Han, S. Liu, Bridging search region interaction with template for rgb-t tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.
[47] J. Yang, Z. Li, F. Zheng, A. Leonardis, J. Song, Prompting for multi-modal tracking, in: Proceedings of the ACM International Conference on Multimedia, 2022.
[48] J. Zhu, S. Lai, X. Chen, D. Wang, H. Lu, Visual prompt multi-modal tracking, in: IEEE Conference on Computer Vision and Pattern Recognition, 2023, pp. 9516–9526.
[49] Z. Tang, T. Xu, X. Wu, X.-F. Zhu, J. Kittler, Generative-based fusion mechanism for multi-modal tracking, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024.
[50] Z. Wu, J. Zheng, X. Ren, F.-A. Vasluianu, C. Ma, D. P. Paudel, L. Van Gool, R. Timofte, Single-model and any-modality for video object tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.
[51] L. Hong, S. Yan, R. Zhang, W. Li, X. Zhou, P. Guo, K. Jiang, Y. Chen, J. Li, Z. Chen, et al., Onetracker: Unifying visual object tracking with foundation models and efficient tuning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.
[52] B. Cao, J. Guo, P. Zhu, Q. Hu, Bi-directional adapter for multimodal tracking, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024.
[53] A. Lu, C. Li, J. Zhao, J. Tang, B. Luo, Modality-missing rgb-t tracking: Invertible prompt learning and high-quality benchmarks, International Journal of Computer Vision 133 (2025) 2599–2619.
[54] T. Zhang, Q. Zhang, K. Debattista, J. Han, Cross-modality distillation for multi-modal tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 47 (2025) 5847–5865.
[55] Y. Xiao, J. Zhao, A. Lu, C. Li, B. Yin, Y. Lin, C. Liu, Cross-modulated attention transformer for rgb-t tracking, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, 2025, pp. 8682–8690.
[56] C. Li, W. Xue, Y. Jia, Z. Qu, B. Luo, J. Tang, D. Sun, Lasher: A large-scale high-diversity benchmark for rgbt tracking, IEEE Transactions on Image Processing 31 (2021) 392–404.
[57] P. Zhang, J. Zhao, D. Wang, H. Lu, X. Ruan, Visible-thermal uav tracking: A large-scale benchmark and new baseline, in: IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp.

8886–8895.

[58] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.

[59] Y. Zheng, B. Zhong, Q. Liang, Z. Mo, S. Zhang, X. Li, Odtrack: Online dense temporal token learning for visual tracking, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024.