Article:
Zhang, T., Jiao, Q., Zhang, Q. et al. (2024) Exploring multi-modal spatial-temporal contexts for high-performance RGB-T tracking. IEEE Transactions on Image Processing, 33. pp. 4303-4318. ISSN: 1057-7149

https://doi.org/10.1109/TIP.2024.3428316

Exploring Multi-modal Spatial-Temporal Contexts for High-performance RGB-T Tracking

Tianlu Zhang, Qiang Jiao, Qiang Zhang* and Jungong Han*, *Senior Member, IEEE*

***Abstract*—In RGB-T tracking, there exist rich spatial relationships between the target and backgrounds within multi-modal data as well as sound consistencies of spatial relationships among successive frames, which are crucial for boosting the tracking performance. However, most existing RGB-T trackers overlook such multi-modal spatial relationships and temporal consistencies within RGB-T videos, hindering them from robust tracking and practical applications in complex scenarios. In this paper, we propose a novel Multi-modal Spatial-Temporal Context (MMSTC) network for RGB-T tracking, which employs a Transformer architecture for the construction of reliable multi-modal spatial context information and the effective propagation of temporal context information. Specifically, a Multi-modal Transformer Encoder (MMTE) is designed to achieve the encoding of reliable multi-modal spatial contexts as well as the fusion of multi-modal features. Furthermore, a Quality-aware Transformer Decoder (QATD) is proposed to effectively propagate the tracking cues from historical frames to the current frame, which facilitates the object searching process. Moreover, the proposed MMSTC network can be easily extended to various tracking frameworks. New state-of-the-art results on five prevalent RGB-T tracking benchmarks demonstrate the superiorities of our proposed trackers over existing ones.**

***Index Terms*—RGB-T tracking, Multi-modal spatial context, Temporal context, Transformer**

I. Introduction

RGB-T tracking is one of the fundamental computer vision tasks, which aims to estimate the state of an arbitrary target object in each frame of an RGB-T video sequence, given only its initial appearance [1]. Due to the all-weather and all-day working capability, RGB-T tracking has attracted increasing attention. Despite the recent significant efforts in RGB-T tracking, there still exist great gaps for practical applications due to some challenging factors, such as occlusions, fast motions and appearance changes. This urges us to develop advanced RGB-T trackers with strong adaptiveness and robustness.

Since RGB and thermal information are strongly complementary to each other, most current RGB-T trackers study how to integrate cross-modality features to improve the reliability of appearance information, and tackle the task of RGB-T tracking by learning an appearance model about the target. Over the past few years, plenty of RGB-T tracking methods [2]–[9] have been developed and have shown great performance in various challenges, such as low illumination and thermal crossover. According to their types of baseline trackers, recent RGB-T tracking methods based on Convectional Neural Networks (CNNs) can be mainly divided into Siamese network based RGB-T trackers, Multi-Domain network (MDNet) based trackers and Discriminative Correlation Filter (DCF) based trackers. Especially, such Siamese network based RGB-T trackers [2]–[4] address object tracking as a similarity matching problem between the target template and the search frames in an offline trained manner. But their usually used temporal information just contains some certain motion priors (e.g., cosine window), as shown in Fig. 1 (a). Differently, as shown in Fig. 1 (b), those MDNet based trackers [5]–[7] and DCF based trackers [8]–[10] train a classifier to distinguish targets from their surrounding backgrounds in an online way. Benefiting from the utilization of historical frames, these online-trained trackers are usually more discriminative than those offline-trained ones. While these online-trained trackers update the target appearance information by simply using some previously tracked frames. Such a strategy cannot effectively capture the spatial relationships between the target and other objects as well as the temporal consistencies in the scene. Recently, several RGB-T trackers based on Transformers [11] have been introduced and show excellent tracking performance. But they [12]–[18] train their models in an offline trained manner and only rely on dynamic templates to exploit appearance information within the temporal domain.

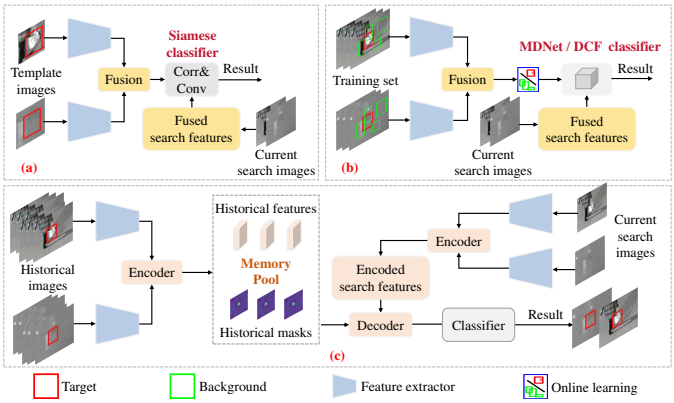


Fig. 1. Illustration of different pipelines of RGB-T trackers. (a) Siamese based RGB-T trackers. (b) MDNet or DCF based RGB-T trackers. (c) The proposed MMSTC framework.

This work was supported in part by the State Key Laboratory of Reliability and Intelligence of Electrical Equipment under Grant EERI KF2022005, in part by the Hebei University of Technology, in part by the National Natural Science Foundation of China under Grant 61803290 and Grant 61773301, in part by China Postdoctoral Science Foundation under Grant 2023M742745, and in part by the Natural Science Foundation of Shaanxi Province under Grant 2019JQ-312. (*Corresponding authors: Qiang Zhang, Jungong Han.)

Tianlu Zhang, Qiang Jiao and Qiang Zhang are with State Key Laboratory of Electromechanical Integrated Manufacturing of High-Performance Electronic Equipments, and the Center for Complex Systems, School of Mechano-Electronic Engineering, Xidian University, Xi'an, 710071, Shaanxi, China. Email: tianluzhang@stu.xidian.edu.cn, qjiao@xidian.edu.cn and qzhang@xidian.edu.cn.

Jungong Han is with Computer Science Department, University of Sheffield, S1 4DP, UK. Email: jungonghan77@gmail.com.

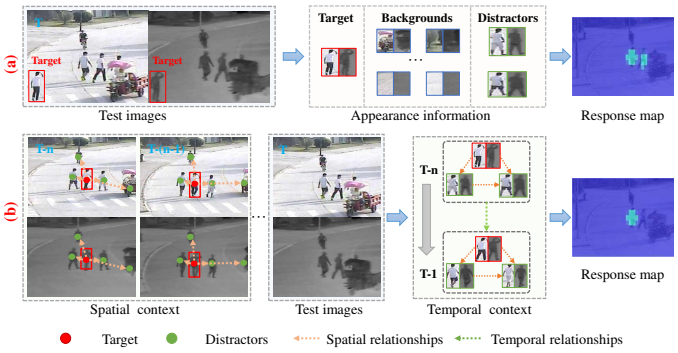


Fig. 2. Illustration of the scene information within an RGB-T video. (a) Most existing RGB-T trackers only utilize the target appearance information to track the target. However, such a strategy fails in this example. Here, the presence of distractors makes it almost impossible to correctly localize the target based on the appearance only, even if the appearance model is continuously updated by using previous frames. (b) In contrast, the scene information propagated through the sequence greatly simplifies the target localization problem, allowing us to confidently localize the target.

As a result, both those offline-trained trackers and online-trained trackers are hard to confidently locate the targets just according to their appearance information in some cases, especially with the presence of distractor objects, as shown in Fig. 2 (a).

In fact, humans usually exploit much richer scene information and maintain temporal continuity in a constantly changing environment when tracking an object. For instance, considering the example in Fig. 2(a), if we record the relationships between targets and distractors in each historical frame, and transfer such historical relationships to the current frame, we can easily detect the presence of distractors and determine the location of the target. According to the above analysis, there are two types of information that are crucial for improving the adaptive and discriminative abilities of an RGB-T tracker: multi-modal spatial contexts and temporal contexts. Especially, multi-modal spatial contexts refer to the relationships among different spatial positions within a pair of RGB-T images, which can reflect the associations between targets and backgrounds (including distractors). Meanwhile, the temporal contexts refer to the temporal consistencies of spatial relationships among different frames in an RGB-T video pair, as shown in Fig. 2 (b).

However, effectively exploiting multi-modal spatial contexts and temporal contexts within RGB-T videos for tracking is an arduous task. To accomplish this purpose, we propose a Multi-modal Spatial-Temporal Context (MMSTC) network, which employs a Transformer architecture for the construction of multi-modal spatial context information and the propagation of temporal context information, as shown in Fig. 1 (c). Especially, in MMSTC, we first propose a Multi-modal Transformer Encoder (MMTE) to obtain the encoded features of each frame, which simultaneously explores the reliable spatial context information within multi-modal data and integrate multi-modal features. Then, the encoded features of several historical frames as well as their corresponding historical tracking results will be deposited into a memory pool for the subsequent propagation of temporal information. After that, we adopt a proposed Quality-aware Transformer Decoder (QATD) to obtain the decoded features of the current frame, which enables the effective propagation of temporal contexts with the help of the memory pool constructed above. Doing so will assist the prediction of the current frame by virtue of multi-modal spatial contexts and temporal contexts. In addition, the proposed MMSTC framework does not depend on any tracking framework and can be easily integrated into different tracking frameworks, such as Siamese based trackers, DCF based trackers and Transformer based trackers.

Specifically, to exploit the spatial contexts within multi-modal data, the most straightforward way is to first model the spatial contexts of each modality independently and then aggregate them together. However, there may exist some unreliable spatial relationships within each unimodal data. Directly utilizing all of these spatial relationships within each modality data may reduce the validity of spatial context information within the entire scenario. In fact, such spatial contexts within the same scenarios across different modalities tend to be potentially consistent. It may be reasonable to introduce some reliable spatial contexts from one modality data to the other modality data to improve the reliability of spatial contexts. Therefore, in the proposed MMTE, the unimodal spatial contexts, which are modeled by using a self-attention mechanism, will be first enhanced by using some cross-modal spatial contexts to improve their effectiveness. Then, the enhanced unimodal spatial contexts will be embedded into the unimodal RGB and thermal features, respectively. After that, these unimodal features embedded with spatial contexts will be aggregated together to obtain the final encoded features.

By using the proposed MMTE module, we can obtain the encoded features of each frame. On top of that, the encoded features of several historical frames as well as their corresponding tracking results will be employed to construct a memory pool. With the assistance of the constructed memory pool, valuable temporal information across frames can be conveyed to enhance the encoded features of the current frame. However, propagating the whole temporal context information stored in the memory pool to the encoded features of the current frame is not always feasible, since there may still exist some unreliable temporal contexts in the backgrounds. For instance, when the background scenes change drastically in a video, there may exist some interference information within the backgrounds, which may reduce the effectiveness of temporal contexts and further deteriorate the tracking performance.

Considering that, in the proposed QATD, according to the locations of targets, those temporal contexts within the memory pool will be divided into target-related contexts and background-related contexts. Such target-related contexts mainly reflect the relationships between targets and backgrounds (including distractor objects), and will be fully delivered to distinguish targets and distractors. In contrast, such background-related contexts mainly reflect the relationships between backgrounds and backgrounds as well as the relationships between backgrounds and targets. Considering the existence of some interference information within the backgrounds, those background-related contexts will be selectively delivered according to their qualities to alleviate the influence of interference information.

To sum up, our work improves an RGB-T tracker dramatically because of the following four contributions:

- A novel RGB-T tracking framework based on Transformer, i.e., MMSTC, is presented to model multi-modal spatial contexts and temporal contexts for robust RGB-T tracking. Especially, the proposed MMSTC framework can be easily integrated into various tracking frameworks.
- We propose a Multi-modal Transformer Encoder (MMTE) to simultaneously explore the reliable spatial context information within multi-modal data and integrate multi-modal features by using some cross-modal spatial contexts to enhance the intra-modal spatial contexts.
- We propose a Quality-aware Transformer Decoder (QATD) to enable the effective propagation of temporal context information by using such reliable multi-modal spatial contexts stored in the memory pool to enhance the encoded features of the current frame.
- Our proposed tracker achieves new state-of-the-art results on five prevalent RGB-T tracking benchmarks, including GTOT [19], RGBT210 [20], RGBT234 [21], LasHeR [22] and VTUAV [23].

II. Related Work

RGB-T Tracking Methods based on CNNs: In the past few years, numerous RGB-T tracking algorithms based on CNNs have been proposed to boost tracking performance, which can be categorized into three main types based on the employed tracking framework: MDNet [24] based methods, Siamese network [25] based methods and DCF [26] based methods. These MDNet-based RGB-T trackers mainly improve tracking performance by mining multi-modal complementary information [6], [27]–[30] and enhancing feature representation capabilities [5], [31]. For instance, Lu *et al.* [6] proposed a duality-gated mutual condition network to exploit the discriminative information of all modalities and suppress noise interference. Zhu *et al.* [27] designed a feature aggregation network to progressively aggregate hierarchical features and eliminate redundant information through a pruning module. Li *et al.* [5] introduced a multi-adapter architecture to learn modality-shared, modality-specific and instance-aware target representations, respectively. Besides, some MDNet based methods [7], [32]–[34] try to explore attribute-based target representation for improving tracking robustness. In addition, several methods [9], [23], [35] introduced DiMP [26] as their baseline tracker and achieved promising tracking performance. Meanwhile, aiming to speed up the tracking, some works [3], [36]–[38] introduce the Siamese networks to RGB-T tracking, where their classifiers are trained in an offline manner. In contrast to those online-trained trackers, offline-trained trackers are faster by sacrificing the discriminability.

RGB-T Tracking Methods based on Transformer: Recently, with the rapid development of the Transformer architecture, more and more RGB-T tracking methods based on Transformers [11] have been introduced. Currently, there are three main paths followed by RGB-T tracking methods based on Transformer. The first type of methods [39]–[41] employed the Transformer block to match the template and search features after performing multi-modal feature fusion within the Siamese architecture. However, these methods failed to establish the global correlations between different multi-modal data. Differently, the second type of methods [12], [14], [42], [43] utilized the one-stream structure [44] for the unimodal feature extraction and designed various kinds of multi-modal feature fusion strategies based on Transformer. The third type of method [16]–[18], [45]–[47] aims to adapt the RGB tracking model to RGB-T tracking in the prompt learning manner. Although the existing methods based on Transformer achieved high accuracy, the temporal information within the RGB-T videos has not been fully studied.

Spatial Information Exploitation in Object Tracking: In the field of RGB tracking, some approaches have been proposed to model the spatial context information for object tracking. Particularly, the early methods [25] usually utilize non-local blocks to model spatial contexts. Recently, some algorithms [48] attempt to introduce the Transformer architecture into the object tracking community to explore spatial context information. For example, TransT [48] designed a feature fusion network based on Transformer to combine the template and search region features, which consists of an ego-context augment module with self-attention as well as a cross-feature augment module with cross-attention. More recently, several one-stream tracking frameworks [44], [49], [50] have been proposed to embed the feature correlation learning in the feature extraction network, which achieve promising results on multiple benchmarks. In the field of RGB-T tracking, some methods [4], [29], [34] also consider the spatial context information or temporal context information for improving tracking robustness. Especially in CMPP [51], a cross-modal pattern-propagation framework was presented to build the inter-modal pattern-propagation and the interaction across modalities within local regions. In AGMINet [52], a global mining module was proposed to explore the global spatial context information as well as the global correlation between modalities. However, these existing RGB-T trackers usually pay less attention to such unreliable spatial relationships within multi-modal spatial contexts, which may deteriorate the feature representations.

Temporal Information Exploitation in Object Tracking: The utilization of temporal information plays a crucial role in the tracking task. Numerous tracking frameworks [53]–[56] focus on adaptively updating the tracking model by leveraging the accumulated tracking results from historical frames. Additionally, several trackers [13], [57] dynamically update the target template to enhance adaptability. Meanwhile, some methods [58]–[62] attempt to propagate scene information through the temporal domain to explicitly eliminate interference from distractor objects. For instance, KYS [58] first exploited the scene information to generate some dense localized state vectors, and then propagated such valuable scene information through the sequence via a state propagation module. Those scene knowledge, along with the target appearance model, are used to predict the target state in each frame. TrDiMP [59] bridged the relationships of individual video frames and explored the temporal contexts across them via a Transformer architecture for robust object tracking. However,

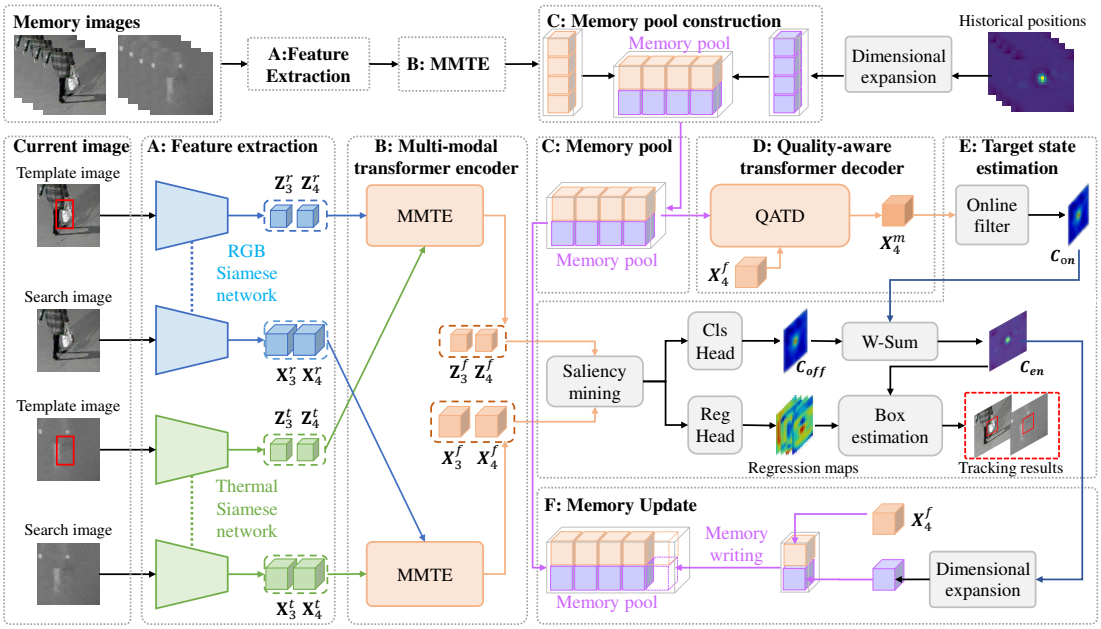


Fig. 3. The overall pipeline of the proposed Siam-MMSTC. It consists of two-stream Siamese networks, MMTE module, memory pool construction, QATD module, target state estimation and memory update. In particular, both feature extraction networks adopt the Siamese network structures.

the rich temporal information within the global scene has not been fully exploited to enhance the tracking performance in existing RGB-T trackers.

III. MMSTC FRAMEWORK

Given an RGB-T tracking sequence, the proposed Multi-modal Spatial-Temporal Context (MMSTC) network aims to model the multi-modal spatial contexts and temporal contexts based on Transformer and exploits these contexts to perform robust tracking. In this paper, we integrate the designed MMSTC into a Siamese based RGB-T tracker. Fig. 3 shows the architecture of the proposed tracker (denoted as Siam-MMSTC), which consists of six modules, including feature extraction, MMTE, memory pool construction, QATD, target state estimation (TSE) module and memory update.

In the Siam-MMSTC network, the feature extraction module extracts unimodal features of template images and search images from each RGB-T frame, respectively (see Sec. III-A). Given the unimodal (RGB and thermal) template features and search features, the MMTE module simultaneously models the spatial context information within multi-modal data as well as integrates multi-modal features to obtain the encoded template features and the encoded search features for each frame (see Sec. III-B). On top of that, the historical encoded search features and their corresponding tracking results are further combined to form a memory pool (see Sec. III-C). After that, the QATD module takes the current encoded search features and the historical encoded search features within memory pool as inputs and generates the decoded search features, which are reinforced by the temporal contexts stored in the memory pool (see Sec. III-D). Here, to make full use of the temporal context information as well as the appearance information, the encoded template features, the encoded search features and the decoded search features will be simultaneously fed into the TSE component to predict the final tracking results (see Sec. III-E). After obtaining the tracking results, we continuously update the memory pool (see Sec. III-F). In the following contents, we will describe the proposed tracking framework in detail.

A. Feature Extraction

Similar to the Siamese-based RGB-T trackers [2], our proposed Siam-MMSTC tracker also takes a pair of RGB image patches (the RGB template image and the RGB search image) and a pair of thermal image patches (the thermal template image and the thermal search image) as the inputs of the two-stream Siamese networks, i.e., an RGB Siamese network and a Thermal Siamese network, respectively. The unimodal template features are cropped from the feature maps of the template image according to its bounding box, and are pooled by a PrPool layer [26]. Here, we employ a modified ResNet-50 backbone [63] in both the RGB Siamese network and the Thermal Siamese network, yielding four levels of feature maps. Specifically, we remove the down-sampling operations and replace the traditional 3×3 convolutions with the 3×3 atrous convolutions of atrous rate 2 to increase the receptive fields in the fourth stage of ResNet-50. Meanwhile, the last stage of ResNet-50 is cut off. The two-stream Siamese networks take the third stage and the fourth stage as the final outputs of template features (RGB template features denoted as $\mathbf{Z}_3^r$ and $\mathbf{Z}_4^r$, and thermal template features denoted as $\mathbf{Z}_3^t$ and $\mathbf{Z}_4^t$) and the final outputs of search features (RGB search features denoted as $\mathbf{X}_3^r$ and $\mathbf{X}_4^r$, and thermal search features denoted as $\mathbf{X}_3^t$ and $\mathbf{X}_4^t$), respectively.

B. Multi-modal Transformer Encoder

Given unimodal template features and search features from the RGB and Thermal Siamese networks for each frame, the next step is to fully explore the reliable spatial context information within multi-modal data and obtain the encoded

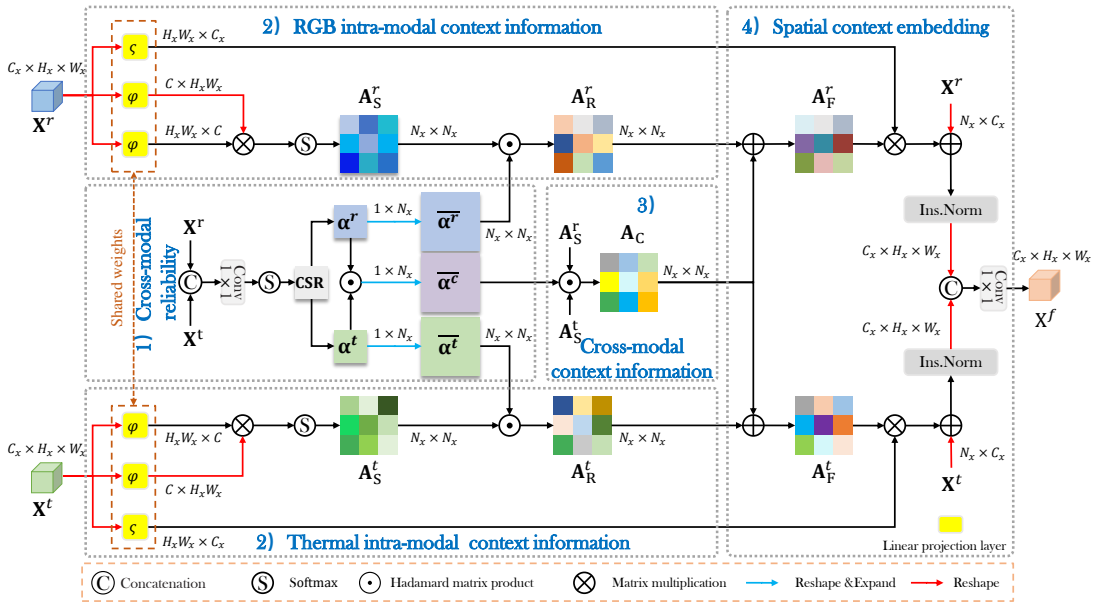


Fig. 4. The proposed multi-modal Transformer encoder.

template features and search features. For that, we design a Multi-modal Transformer Encoder (MMTE) in the template branch and the search branch, respectively.

Fig. 4 details the data-flow process of MMTE. First, considering the fact that unreliable local regions may affect the effectiveness of spatial context information, we predict the relative reliability of each spatial position between RGB features and thermal features, and determine those regions where both modalities are reliable. Secondly, we compute the intra-modal context information within RGB modality and thermal modality by the self-attention mechanism in Transformer [64], respectively, and select such reliable intra-modal contexts within each modality according to the above relative reliabilities. Thirdly, due to the fact that those spatial contexts within the same scenarios across different modalities tend to be potentially consistent, we conduct the context communication among spatial contexts of different modalities within such regions, where both modalities are reliable, to generate the cross-modal contexts. After that, such cross-modal contexts will be further employed to enhance those reliable intra-modal contexts. The final encoded features will be obtained by first embedding those enhanced contexts into the unimodal features and then integrating them together. In the following contents, we will take the computation in the search branch as an example to describe the MMTE module.

1) Cross-modal reliability: Due to the different imaging mechanisms of RGB and thermal images, their reliability degrees may be different in various tracking scenarios. Considering that, the proposed MMTE module first predicts the relative reliability of each spatial position between multi-modal images and then selects the common reliable regions.

For that, we first concatenate unimodal RGB features $\mathbf{X}^r \in \mathbb{R}^{C_x \times H_x \times W_x}$ and unimodal thermal features $\mathbf{X}^t \in \mathbb{R}^{C_x \times H_x \times W_x}$. Here, C_x, H_x and W_x denotes the channel dimension, height and width of the unimodal features. Then, we use a convolution layer of kernel size 1×1 and a softmax layer to get a two-channel cross-modal spatial reliability map $\mathbf{CSR} \in \mathbb{R}^{2 \times H_x \times W_x}$. The two-channel weight map $\mathbf{CSR}$ is split into two reliability weight maps, i.e., one weight map $\alpha^r \in \mathbb{R}^{1 \times H_x \times W_x}$ from the first channel of $\mathbf{CSR}$ for selecting the features extracted from RGB images, and the other weight map $\alpha^t \in \mathbb{R}^{1 \times H_x \times W_x}$ from the second channel of $\mathbf{CSR}$ for selecting the features extracted from thermal images. Mathematically, these steps are expressed by:

$$\alpha^r, \alpha^t = split(softmax(conv(cat(\mathbf{X}^r, \mathbf{X}^t), \theta_1))), \quad (1)$$

where $cat(*)$ denotes the concatenation operation and $conv(*, \theta_1)$ denotes the convolution layer with kernel size 1×1 and parameters θ_1. $\mathrm{split}(*)$ denotes splitting the two-channel feature map into two reliability weight maps.

Moreover, in the subsequent spatial context information modeling stage, α^r and α^t will be used to measure the reliability of spatial context information and determine those reliable spatial context information. As well, in order to facilitate subsequent matrix operations, the two reliability weight maps are first reshaped to the size of $1 \times N_x$ with $N_x = H_x W_x$, and then expanded as two matrices $\overline{\alpha}^r$ and $\overline{\alpha}^t$ of sizes $N_x \times N_x$ by copying the original matrix N_x times.

Meanwhile, with α^r and α^t available, we get the common reliable regions within the two modalities as follows:

$$\alpha^c = \alpha^r \odot \alpha^t, \quad (2)$$

where $\odot$ denotes the Hadamard matrix product. Similarly, for the subsequent matrix operations, α^c will be also reshaped and expanded as a matrix $\overline{\alpha}^c$ of sizes $N_x \times N_x$. $\overline{\alpha}^c$ reflects the reliable regions simultaneously existing in RGB modality and thermal modality.

2) Intra-modal context information: The fundamental component in a classic Transformer model is the attention mechanism. Following [64], given the query $\mathbf{Q} \in \mathbb{R}^{N_q \times C}$, key $\mathbf{K} \in \mathbb{R}^{N_k \times C}$ and value $\mathbf{V} \in \mathbb{R}^{N_k \times C}$, the attention mechanism adopts the dot-product to compute the similarity matrix $\mathbf{A}_{\mathrm{K}\to\mathrm{Q}} \in \mathbb{R}^{N_q \times N_k}$ between the query and key as follows:

$$\mathbf{A}_{\mathrm{K}\to\mathrm{Q}} = \mathbf{Atten}(\mathbf{Q}, \mathbf{K}) = softmax\left(\hat{\mathbf{Q}}\,\hat{\mathbf{K}}^{\top}/\tau\right), \quad (3)$$

where $\hat{\mathbf{Q}}$ and $\hat{\mathbf{K}}$ are ℓ_2-normalized features of $\mathbf{Q}$ and $\mathbf{K}$ across the channel dimension, respectively, and τ is a temperature parameter to control the softmax distribution. With the similarity matrix $\mathbf{A}_{\mathrm{K}\to\mathrm{Q}}$ from key to query, we can transform the value via $\mathbf{A}_{\mathrm{K}\to\mathrm{Q}}\mathbf{V} \in \mathbb{R}^{N_q\times C}$.

Inspired by the Transformer model [64], we employ the attention mechanism to model spatial context information within RGB modality and thermal modality, respectively. For that, the RGB features $\mathbf{X}^r$ and thermal features $\mathbf{X}^t$ are first reshaped to $\overline{\mathbf{X}}^r$ and $\overline{\mathbf{X}}^t \in \mathbb{R}^{N_x\times C_x}$ with $N_x = H_x W_x$. After that, we compute the self-attention maps $\mathbf{A}^r_{\mathrm{S}} = \mathbf{Atten}(\varphi(\overline{\mathbf{X}}^r), \varphi(\overline{\mathbf{X}}^r)) \in \mathbb{R}^{N_x\times N_x}$ and $\mathbf{A}^t_{\mathrm{S}} = \mathbf{Atten}(\varphi(\overline{\mathbf{X}}^t), \varphi(\overline{\mathbf{X}}^t)) \in \mathbb{R}^{N_x\times N_x}$, where $\varphi(*)$ is a linear projection layer that reduces the embedding channel from C_x to $C = C_x/4$.

The self-attention maps $\mathbf{A}^r_{\mathrm{S}}$ and $\mathbf{A}^t_{\mathrm{S}}$ reflect the spatial context information within the input RGB image and thermal image, respectively. With the reliability weight maps $\overline{\alpha}^r$ and $\overline{\alpha}^t$ obtained by Eq. 1, their corresponding reliable spatial context information is thus obtained by

$$\begin{aligned}\mathbf{A}^r_{\mathrm{R}} &= \mathbf{A}^r_{\mathrm{S}} \odot \overline{\alpha}^r,\\ \mathbf{A}^t_{\mathrm{R}} &= \mathbf{A}^t_{\mathrm{S}} \odot \overline{\alpha}^t.\end{aligned} \tag{4}$$

$\mathbf{A}^r_{\mathrm{R}} \in \mathbb{R}^{N_x\times N_x}$ reflects the reliable relationships among different spatial positions within RGB modality data. Similarly, $\mathbf{A}^t_{\mathrm{R}} \in \mathbb{R}^{N_x\times N_x}$ reflects the reliable relationships among different spatial positions within thermal modality data.

3) Cross-modal context information: The cross-modal contexts reflect the relationships among different spatial locations of different modality features, which may be simply obtained by multiplying two intra-modal contexts. The spatial contexts within the same scenarios across different modalities tend to be potentially consistent. Therefore, the intra-modal spatial contexts can be enhanced by the cross-modal contexts to improve the reliability of spatial contexts. However, we need to pay attention to the fact that the unreliable information of one modality data may cause the invalidation of the cross-modal context information. Therefore, in MMTE, we just consider the interaction of such intra-modal context information within the common reliable regions α^c determined by Eq.2. Accordingly, the cross-modal context information $\mathbf{A}_{\mathrm{C}}$ can be calculated by multiplying two intra-modal spatial relationships:

$$\mathbf{A}_{\mathrm{C}} = \overline{\alpha}^c \odot (\mathbf{A}^r_{\mathrm{S}} \odot \mathbf{A}^t_{\mathrm{S}}). \tag{5}$$

The cross-attention maps $\mathbf{A}_{\mathrm{C}} \in \mathbb{R}^{N_x\times N_x}$ reflects the reliable relationships among different spatial positions within multi-modal data, which can be further employed to enhance the intra-modal context information via,

$$\begin{aligned}\mathbf{A}^r_{\mathrm{F}} &= \mathbf{A}^r_{\mathrm{R}} + \mathbf{A}_{\mathrm{C}},\\ \mathbf{A}^t_{\mathrm{F}} &= \mathbf{A}^t_{\mathrm{R}} + \mathbf{A}_{\mathrm{C}}.\end{aligned} \tag{6}$$

Compared with the intra-modal contexts in Eq. 4, the enhanced attention maps $\mathbf{A}^r_{\mathrm{F}}$ and $\mathbf{A}^t_{\mathrm{F}}$ further improve the effectiveness of spatial contexts via the interaction of common reliable context information.

4) Spatial contexts embedding: Based on the attention maps $\mathbf{A}^r_{\mathrm{F}}$ and $\mathbf{A}^t_{\mathrm{F}}$, we transform the RGB features and thermal features through $\mathbf{A}^r_{\mathrm{F}}\varsigma(\overline{\mathbf{X}}^r)$ and $\mathbf{A}^t_{\mathrm{F}}\varsigma(\overline{\mathbf{X}}^t)$, respectively, which are then added to the original RGB features and thermal features via a residual term. Here $\varsigma(*)$ is a linear projection layer. The formulations are as follows:

$$\begin{aligned}\overline{\mathbf{X}}^r_{\mathrm{E}} &= \text{ Ins. Norm } (\mathbf{A}^r_{\mathrm{F}}\varsigma(\overline{\mathbf{X}}^r) + \overline{\mathbf{X}}^r),\\ \overline{\mathbf{X}}^t_{\mathrm{E}} &= \text{ Ins. Norm } (\mathbf{A}^t_{\mathrm{F}}\varsigma(\overline{\mathbf{X}}^t) + \overline{\mathbf{X}}^t),\end{aligned} \tag{7}$$

where Ins. Norm$(*)$ denotes the instance normalization. As in [59], the proposed MMTE module slims the classic Transformer by omitting the fully connected feed-forward layers and adopting the single-head attention to achieve a good balance between speed and performance.

The encoded unimodal RGB features $\mathbf{X}^r_{\mathrm{E}} \in \mathbb{R}^{C_x\times H_x\times W_x}$ and thermal features $\mathbf{X}^t_{\mathrm{E}} \in \mathbb{R}^{C_x\times H_x\times W_x}$ are then obtained by reshaping $\overline{\mathbf{X}}^r_{\mathrm{E}}$ and $\overline{\mathbf{X}}^t_{\mathrm{E}}$ back to their original sizes, respectively. Finally, the encoded features $\mathbf{X}^f$ are obtained by performing concatenation and convolution operations on those encoded unimodal features $\mathbf{X}^r_{\mathrm{E}}$ and $\mathbf{X}^t_{\mathrm{E}}$, i.e.,

$$\mathbf{X}^f = conv(cat(\mathbf{X}^r_{\mathrm{E}}, \mathbf{X}^t_{\mathrm{E}}), \theta_2). \tag{8}$$

Here, $conv(*, \theta_2)$ denotes the convolution layer with kernel size 1×1 and parameters θ_2.

By performing the proposed MMTE modules on the multi-modal search features from the third and fourth stages of ResNet50, we obtain two levels of encoded search features $\mathbf{X}^f_3$ and $\mathbf{X}^f_4$, respectively. Similarly, we obtain two levels of encoded template features $\mathbf{Z}^f_3$ and $\mathbf{Z}^f_4$ by using the proposed MMTE modules.

C. Memory Pool Construction

Although the above MMTE module explores the multi-modal context information within each RGB-T frame, it is still hard to locate the target confidently in case of the presence of similar objects (also called distractors), when the tracker overlooks the temporal relationships among successive frames. To bridge the spatial context information among different frames and convey such rich temporal cues across them, we need to collect the features of historical frames to build a memory feature pool for the subsequent transformation of temporal context information. It is also meaningful to record the position information of the target in the historical frames for the distinction between the target and the backgrounds. Therefore, we also collect the target position information of historical frames to build a memory position pool. The whole memory pool is thus constructed by the memory feature pool and the memory position pool.

1) Memory feature pool: In our proposed framework, we use the encoded search features $\mathbf{X}^f_4$ from the historical frames as historical encoded features $\mathbf{X}^h$ for convenience. Specifically, for n historical frames, a set of historical encoded features $\mathbf{X}^h_i, i \in (1, n)$ are concatenated to form the memory feature pool $\mathbf{MF} = cat(\mathbf{X}^h_1, ...\mathbf{X}^h_n) \in \mathbb{R}^{n\times C_x\times H_x\times W_x}$.

2) Memory position pool: According to the temporal locations of the target, we construct a series of location masks $\mathbf{m}_i \in \mathbb{R}^{H_x\times W_x}, i \in (1, n)$ from the memory features by setting a Gaussian function [26] centered at the location of the target. Similar to the memory feature pool $\mathbf{MF}$, we also concatenate these masks to form a memory position pool

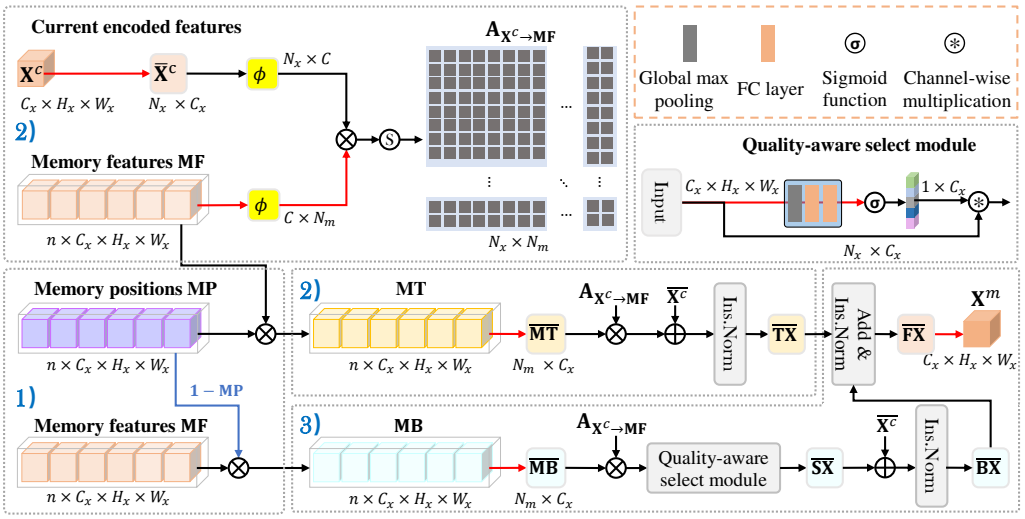


Fig. 5. The proposed quality-aware Transformer decoder.

$\mathbf{MP}^{'} = cat(\mathbf{m}_1, \cdots, \mathbf{m}_n) \in \mathbb{R}^{n \times H_x \times W_x}$, which is expanded into $\mathbf{MP} \in \mathbb{R}^{n \times C_x \times H_x \times W_x}$.

Finally, the memory pool is constructed by the memory feature pool $\mathbf{MF}$ and the memory position pool $\mathbf{MP}$ as $\mathbf{M} = \{\mathbf{MF}, \mathbf{MP}\}$.

D. Quality-aware Transformer Decoder

Thanks to the proposed MMTE module, we obtain high-quality encoded features, which simultaneously explore the effective spatial context information within multi-modal data and integrate multi-modal features. And the memory pool collects the encoded search features as well as their corresponding tracking results of the historical frames. The constructed memory pool $\mathbf{M} = \{\mathbf{MF}, \mathbf{MP}\}$ in III-C and the current encoded search feature in Sec. III-B are further fed into a decoder block to reinforce the encoded search features of the current frame. The encoded search features of each frame not only contain the spatial context relationships between targets and backgrounds (called the target-related context information here), but also the spatial context relationships between backgrounds and backgrounds as well as the relationships between backgrounds and targets (called the background-related context information here). Considering the existence of some interference information within the background-related contexts, we design a Quality-aware Transformer Decoder (QATD) in our proposed tracker, in which the target-related context information and the background-related context information are treated differently, to make full use of such valid temporal cues and reduce the introduction of those noise interferences. In QATD, we also use the encoded search features $\mathbf{X}_4^f$ of the current frame as the current encoded features $\mathbf{X}^c$. Fig. 5 details the data-flow process of QATD.

1) Determining different types of contexts: We divide the historical encoded features $\mathbf{MF}$ into the target-related context information and the background-related context information according to the target's historical positions $\mathbf{MF}$. Specifically, the target-related context information is computed via:

$$\mathbf{MT} = \mathbf{MF} \odot \mathbf{MP}. \tag{9}$$

Accordingly, the background-related context information is computed by:

$$\mathbf{MB} = \mathbf{MF} \odot (\mathbf{I} - \mathbf{MP}). \tag{10}$$

Here, $\mathbf{I}$ denotes a tensor of the same size as $\mathbf{MF}$, in which all elements are 1.

2) Target-related context transformation: Here, we will take the current encoded features $\mathbf{X}^c$ as inputs, and convey the temporal cues from such two types of context information to enhance the feature representation of current frame. Especially, to facilitate the attention computation, we reshape $\mathbf{MF}$, $\mathbf{MT}$ and $\mathbf{MB}$ to $\overline{\mathbf{MF}}$, $\overline{\mathbf{MT}}$ and $\overline{\mathbf{MB}} \in \mathbb{R}^{N_m \times C_x}$, respectively. Here, $N_m = n \times H_x \times W_x$. For the target-related context information, their cross-attention matrix is first computed by $\mathbf{A}_{\mathbf{X}^c \to \mathbf{MF}} = \mathrm{Atten}(\phi(\overline{\mathbf{X}}^c), \phi(\overline{\mathbf{MF}})) \in \mathbb{R}^{N_x \times N_m}$, where $\phi(*)$ is a linear projection layer and is similar to $\varphi(*)$. The cross-attention matrix $\mathbf{A}_{\mathbf{X}^c \to \mathbf{MF}}$ reflects the pixel-to-pixel correspondences among frames. Then, with the cross-attention matrix $\mathbf{A}_{\mathbf{X}^c \to \mathbf{MF}}$, the transformed features related to the targets are computed by $\mathbf{A}_{\mathbf{X}^c \to \mathbf{MF}}\overline{\mathbf{MT}}$ and are added to the reshaped $\overline{\mathbf{X}}^c \in \mathbb{R}^{N_x \times C_x}$ via a residual term, thus obtaining the target-related decode features $\overline{\mathbf{TX}}$, i.e.,

$$\overline{\mathbf{TX}} = \text{ Ins. Norm } \left(\mathbf{A}_{\mathbf{X}^c \to \mathbf{MF}} \overline{\mathbf{MT}} + \overline{\mathbf{X}}^c \right). \tag{11}$$

By virtue of the transformed target-related features, the decoded features $\overline{\mathbf{TX}}$ temporally aggregate diverse target representations from a series of historical features to promote themselves.

3) Background-related context transformation: Besides the target-related context information, it is also feasible to propagate those effective background-related context information from historical frames to distinguish distractors. With the cross attention map $\mathbf{A}_{\mathbf{X}^c \to \mathbf{MF}}$ obtained above, we selectively propagate those background-related context information to the encoded features of the current frame. Specifically, the transformed features related to backgrounds are computed by $\mathbf{A}_{\mathbf{X}^c \to \mathbf{MF}}\overline{\mathbf{MB}}$, and are selected via a channel attention

module to filter those unreliable background-related contexts, i.e.,

$$\begin{aligned} cw = sigmoid(fc(gmp(reshape(\mathbf{A}_{\mathbf{X}^c\rightarrow\mathbf{MF}}\overline{\mathbf{MB}})))), \\ \overline{\mathbf{SX}} = (\mathbf{A}_{\mathbf{X}^c\rightarrow\mathbf{MF}}\overline{\mathbf{MB}}) \circledast cw, \end{aligned} \tag{12}$$

where $gmp(*)$ and $fc(*)$ denote the global max pooling layer and the fully connected layer, respectively. $\circledast$ denotes channel-wise multiplication. $reshape(*)$ denotes reshaping features to their original shapes. $cw \in \mathbb{R}^{1\times C_x}$ is the learned weight vector to guide the features to focus on those effective channels. $\overline{\mathbf{SX}}$ are the selected reinforced features.

After that, these selected transformed features $\overline{\mathbf{SX}}$ are also added to $\overline{\mathbf{X}}^c$ via a residual term, obtaining background-related decoded features $\overline{\mathbf{BX}}$ by

$$\overline{\mathbf{BX}} = \text{ Ins. Norm } \left(\overline{\mathbf{SX}} + \overline{\mathbf{X}}^c\right). \tag{13}$$

Finally, we equally combine the aforementioned features $\overline{\mathbf{TX}}$ and $\overline{\mathbf{BX}}$ together, and further normalize them as follows:

$$\overline{\mathbf{FX}} = \text{ Ins. Norm } \left(\overline{\mathbf{TX}} + \overline{\mathbf{BX}}\right). \tag{14}$$

The final output feature $\overline{\mathbf{FX}} \in \mathbb{R}^{N_x\times C_x}$ is reshaped back to the original size for visual tracking. We denote the reshaped version of $\overline{\mathbf{FX}}$ as $\mathbf{X}^m \in \mathbb{R}^{C_x\times H_x\times W_x}$. By performing the proposed QATD module on the encoded search features $\mathbf{X}_4^f$ of the current frame, we obtain the decoded search features $\mathbf{X}_4^m$. The decoded search features will be fed to the TSE module to predict the target location.

E. Target State Estimation

With the encoded features (i.e., $\mathbf{Z}_3^f$, $\mathbf{Z}_4^f$, $\mathbf{X}_3^f$ and $\mathbf{X}_4^f$,) and the decoded search features $\mathbf{X}_4^m$, the tracking problem can be decomposed into a classification task and an estimation task. For the classification task, directly performing an online trained classifiers on the decoded search features $\mathbf{X}_4^m$ may usually achieve good tracking results. However, when the target object is occluded or invisible, the cross attention maps between the current frame and historical frames may be inaccurate, which will deteriorate the representation ability of the decoded features. Therefore, in addition to the online trained classifier, which utilizes the decoded features, we also design an offline trained classifier to utilize the appearance information for tacking.

Specifically, for the online-trained classier, following the end-to-end DCF optimization in DiMP [26], we also generate a discriminative CNN kernel and convolve it with the decoded search features $\mathbf{X}_4^m$ for generating the online response map cls_{on}. Meanwhile, for the offline-trained classier, as that in a typical anchor-free Siamese tracker [63], we input the encoded template features (i.e.,$\mathbf{Z}_3^f$, $\mathbf{Z}_4^f$) and the encoded search features (i.e.,$\mathbf{X}_3^f$ and $\mathbf{X}_4^f$) into the saliency mining module [63] to generate their correlation representations F_{corr}. The output correlation representations will be fed into a classification head for predicting the offline response map cls_{off}.

After that, we conduct the ensemble of the offline classification model and the online prediction model, yielding a fusion score map with high accuracy and robustness. Given the online and offline score maps cls_{on} and cls_{off}, the final score map cls_{en} can be formulated as:

$$cls_{en} = \beta cls_{on} + (1-\beta)cls_{off}, \tag{15}$$

where β is a balance weight between the online and offline score maps and is experimentally set to 0.8 in this paper.

For the regression task, the correlation representations F_{corr} are also fed into a regression head for predicting the bounding box of the target. Here, the classification head in the offline-trained classifier and the regression head in the regression task are both designed as those in SAOT [63]. After obtaining the final score map cls_{en}, the target bounding box is estimated based on the regressed box corresponding to the maximum fusion score.

F. Memory Update

During the online tracking process, in order to better exploit the temporal cues after obtaining the tracking results, we dynamically update the memory feature pool $\mathbf{MF}$ and the corresponding memory position pool $\mathbf{MP}$. To be specific, considering the small differences between densely consecutive frames, we drop the oldest memory in $\mathbf{M}$ and add the currently collected features together with their corresponding position masks to $\mathbf{MF}$ and $\mathbf{MP}$ every 5 frames, respectively. The memory pool maintains a maximal size of 20 frames. Besides, to avoid noisy information introduced by low-quality historical frames, we only update the memory pool when the classification score cls_{on} is greater than 0.7. The QATD module is leveraged in each frame, which generates per-frame decoded search features $\mathbf{X}_4^m$ by propagating the representations and attention cues from the previous memories to the current search images.

G. Implementation Details

In this section, we present the details of the proposed Siam-MMSTC, and illustrate the offline and online training of our Siam-MMSTC model and the online tracking process.

Model setting: In Sec. III-A, we used the modified ResNet-50 [63] in both RGB Siamese network and Thermal Siamese network. The search image is with an area 5^2 times that of the target and is resized to 288×288. The template features is cropped from the feature maps of the first image according to its bounding box and pooled by a PrPool [63] layer to obtain its precise representation of size 8×8. In Sec. III-C, as that in DiMP [26]the maximal size of the memory pool $\mathbf{M}$ is set to be 50 historical frames, and we update the memory pool every 5 frames by replacing the oldest one. In Sec. III-E, the online classifier employs the same structure as the DiMP classifier [26], including a convolution layer with kernel size of 1 and a convolution layer with kernel size of 4. For the offline classification and regression task, both the classification and regression heads are designed following FCOS [65], where each branch consists of 4 convolutional layers with kernel size of 3 and 1 convolutional layer with kernel size of 1.

Off-line Training: The whole Siam-MMSTC model is trained in an end-to-end manner. Specifically, we employ generalized IoU loss and binary cross-entropy loss for the

offline-trained regression and classification tasks. And the training for the online-trained classifier follows DiMP [26]. The proposed model is trained in two stages. Specifically, in the first stage, we disable the thermal branch, the MMTE modules, the memory pool and the QATD module to construct an unimodal tracking network. Here, we adopt some RGB tracking datasets, including COCO [66], GOT10k [67] and LaSOT [68], as our basic training datasets. The backbone ResNet-50 network is initialized by the pre-trained model on the ImageNet [69]. During training, the parameters in ResNet are frozen. Other parameters, except for those of the online discriminative filters, are optimized using ADAM with a learning rate decayed from 1×10^{-3} to 8×10^{-6} and a weight decay of 1×10^{-4}. The training settings of the online discriminative filers follow DiMP [26]. In the second stage, we adopt the training dataset in LasHeR [22], which contains 979 pairs of RGB-T videos, to train the whole model. As well, we fix all the parameters in the RGB feature extractor in this stage. The MMTE and QATD modules are trained with an initial learning rate of 2×10^{-4} and a decay factor of 0.2 for every 15 epochs. Other components are trained with the default learning rates collaboratively multiplied by 0.01. Our whole model is trained for 20 epochs by sampling 26,000 videos per epoch.

Online Training: For online tuning, we use the first frame to pre-train the online classifier. Similar to [26], we also perform data augmentation on the first frame with translation, rotation and blurring, yielding in total 30 initial training samples. Considering the tracking efficiency, we adopt the steepest descent method [26] for fast online optimization.

Online Tracking: During the tracking phase, the proposed Siam-MMSTC aims to predict a bounding box for the target in the current frame. Only the target object in the first frame is adopted as the template patch and is continuously matched with subsequent search images for tracking. Siam-MMSTC performs prediction in the search region to get the classification score map cls_{en} and the regression map $\mathbf{A}^{reg}$. After that, we select the bounding box with the highest classification score as the final tracking one.

IV. Extension to the Transformer Based Framework

It should be noted that the proposed framework can be quickly extended to the Transformer RGB-T tracker to achieve excellent tracking performance. In this paper, we use the one-stream structure [44] to build a Transformer based tacker, referred to as Trans-MMSTC, as shown in Fig. 6. Specifically, the input RGB and TIR search region from each RGB-T frame and template images from the initial RGB-T frame are first split and flattened as sequences of patches, and then fed into the vanilla ViT [11] for joint feature extraction and search-template matching within each modality. After that, the extracted search features from the RGB branch as well as the thermal branch will be fed into our proposed MMTE module to obtain the encoded search features. On top of that, the historical encoded search features and their corresponding tracking results are further combined to form a memory pool

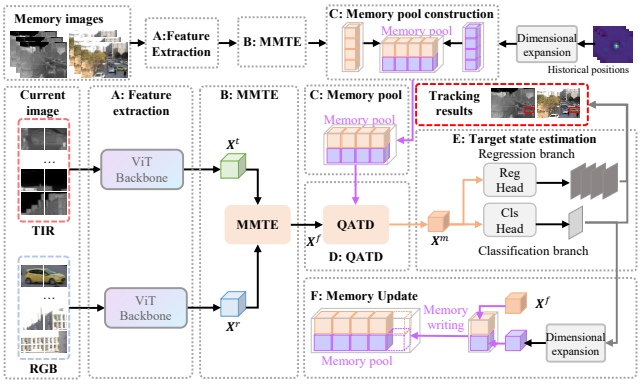


Fig. 6. An overview of the designed Trans-MMSTC.

as that in Siam-MMSTC. After that, the QATD module takes the current encoded search features and the historical encoded search features within the memory pool as inputs and generates the decoded search features, which are reinforced by the temporal contexts stored in the memory pool. Finally, the tracking head, which uses the same structure in OStrack [44], takes the decoded search features as input to predict the tracking results.

As in OStrack [44], the weighted focal loss is employed for classification, and the $L1$ Loss and the generalized IoU loss are employed for regression. We explore temporal information from 5 historical frames and employ another one frame for training, which are randomly selected from the same video. Our network is optimized by the AdamW optimizer with the weight decay of 10^{-4} for 15 epochs. The initial learning rate for the backbone and other parameters are set to 4×10^{-5} and 4×10^{-4}, respectively. The sizes of template patches and search patches are set to 127 pixels and 255 pixels, respectively.

During inference, we dynamically update the memory pool ensemble $\mathbf{M}$. Specifically, the oldest historical tokens as well as their corresponding location masks in $\mathbf{M}$ will be dropped and replaced by the current collected tokens and location masks every 5 frames. The maximal size of the memory pool is set to be 20 historical frames. In addition, in order to avoid noisy information introduced by low-quality historical frames, we only update the memory pool when the classification score is greater than 0.7. The designed QATD is leveraged in each frame to obtain the decoded tokens, which will be fed into the tracking head to obtain the classification map and regression results. Following the common practice [44], we simply multiply the classification map by the Hanning window with the same size, and select the box with the highest score after multiplication as the final tracking result.

V. Experiments

This section presents the results of our Siam-MMSTC and Trans-MMSTC on five tracking benchmark datasets, with comparisons to some state-of-the-art algorithms. Some experimental analysis is also provided to verify the effectiveness of each proposed component on RGBT234 [21]. Our tracker is implemented by using Pytorch on a personal computer with Intel-Xeon(R) 4214 CPU (2.2GHz), 64 GB RAM and Nvidia RTX-3090 GPUs.

TABLE I
QUANTITATIVE COMPARISONS OF OUR METHOD WITH SOME STATE-OF-THE-ART METHODS ON DIFFERENT BENCHMARK DATASETS. HIGHER VALUES INDICATE BETTER PERFORMANCE. THE NUMBERS WITH RED AND BLUE COLORS INDICATE THE BEST AND THE SECOND BEST RESULTS, RESPECTIVELY.

Method	Source	Baseline	Backbone	GTOT		RGBT210		RGBT234		LasHeR			VTUAV-S		VTUAV-L	
				MSR	MPR	MSR	MPR	MSR	MPR	SR	NPR	PR	MSR	MPR	MSR	MPR
FANet† [27]	TIV21	MDNet	VGG-M	72.1	90.1	-	-	53.9	79.4	34.3	42.5	48.2	-	-	-	-
ADRNet [33]	IJCV21	MDNet	VGG-M	73.9	90.4	56.5	80.3	57.1	80.9	-	-	-	46.6	62.2	17.5	23.5
MANet++ [5]	TIP21	MDNet	VGG-M	70.7	88.2	55.3	78.5	55.4	80.0	31.7	40.8	46.7	-	-	-	-
TFNet [28]	TCSVT22	MDNet	VGG-M	72.4	88.6	52.9	77.7	56.0	80.6	-	-	-	-	-	-	-
M5L [31]	TIP22	MDNet	VGG-M	71.0	89.6	-	-	54.2	79.5	-	-	-	-	-	-	-
APFNet [34]	AAAI22	MDNet	VGG-M	73.7	90.5	57.1	82.1	57.9	82.7	36.2	-	50.0	-	-	-	-
DMCNet [6]	TNNLS22	MDNet	VGG-M	73.8	90.9	55.5	79.7	59.3	83.9	35.5	43.1	49.0	-	-	-	-
DRGCNet [29]	IEEE SENS J23	MDNet	VGG-M	73.9	91.0	-	-	58.1	82.5	33.8	42.3	48.3	-	-	-	-
CAT++ [7]	TIP24	MDNet	VGG-M	73.3	91.5	56.1	82.2	59.2	84.0	35.6	44.4	50.9	-	-	-	-
JMMAC [8]	TIP21	DCF	VGG-16	73.2	90.2	57.5	79.2	57.3	79.0	-	-	-	-	-	-	-
HMFT* [23]	CVPR22	DCF	Res-50	74.9	91.2	53.5	78.6	56.8	78.8	-	-	-	62.7	75.8	35.5	41.4
MFNet† [35]	IVC22	DCF	Res-50	73.5	90.7	-	-	60.1	84.4	46.7	55.4	59.7	-	-	-	-
CMD† [9]	CVPR23	DCF	Res-18	73.4	89.2	59.3	83.4	58.4	82.4	46.4	54.6	59.0	-	-	-	-
SiamCDA* [2]	TCSVT22	Siamese	Res-50	73.2	87.7	55.3	78.5	56.9	76.0	-	-	-	-	-	-	-
SiamTDR* [3]	TICPS23	Siamese	VGG-M	71.4	88.5	-	-	55.1	77.2	-	-	-	-	-	-	-
DFAT [37]	IF23	Siamese	Res-50	72.3	89.3	55.0	75.4	55.2	75.8	-	-	-	-	-	-	-
SiamMLAA† [36]	TMM24	Siamese	Res-50	75.1	91.3	56.7	75.6	58.4	78.6	-	-	-	-	-	-	-
Siam-MMSTC†	2024	Siamese	Res-50	76.6	91.2	60.1	85.5	59.9	85.5	47.3	56.4	62.1	63.5	77.6	37.1	43.2
ViPT† [45]	CVPR23	Transformer	ViT-B	-	-	60.3	82.1	61.7	83.5	52.5	61.7	65.1	-	-	-	-
MACFT* [42]	Sensors23	Transformer	ViT-B	-	-	-	-	62.2	85.7	52.5	-	65.3	66.8	80.1	46.7	54.1
RSFNet† [70]	ISPL23	Transformer	ViT-B	75.3	92.1	-	-	62.2	86.3	52.6	-	65.9	-	-	-	-
SiamFEA† [40]	JVCIP23	Transformer	Res-50	76.6	92.0	-	-	61.7	83.7	50.9	-	64.5	-	-	-	-
SiamAFTS* [41]	SR23	Transformer	Res-50	77.7	84.9	-	-	56.4	87.3	-	-	-	-	-	-	-
TBSI† [12]	CVPR23	Transformer	ViT-B	73.4	89.1	62.5	85.3	63.7	87.1	55.6	65.7	69.2	-	-	-	-
MPLT† [47]	ArXiv23	Transformer	ViT-B	75.1	90.0	63.0	86.2	65.7	88.4	57.1	68.0	72.0	65.4	79.7	43.9	50.9
QueryTrack† [15]	TIP24	Transformer	JQF	75.9	92.6	-	-	60.0	84.1	52.0	-	66.0	-	-	-	-
GMMT† [43]	AAAI24	Transformer	ViT-B	-	-	-	-	64.7	87.9	56.6	67.0	70.7	-	-	-	-
TATrack† [13]	AAAI24	Transformer	ViT-B	-	-	61.8	85.3	64.4	87.2	56.1	66.7	70.2	-	-	-	-
BAT† [46]	AAAI24	Transformer	ViT-B	73.3	88.5	63.2	86.0	64.1	86.8	56.3	66.4	70.2	64.1	79.4	42.5	51.5
OneTrack† [16]	CVPR24	Transformer	ViT-B	-	-	-	-	-	-	53.8	-	67.2	-	-	-	-
SDSTrack† [18]	CVPR24	Transformer	ViT-B	-	-	61.4	83.7	62.5	84.8	53.1	62.7	66.5	-	-	-	-
UnTrack† [18]	CVPR24	Transformer	ViT-B	-	-	61.1	82.9	61.7	83.7	53.6	60.1	66.7	-	-	-	-
Trans-MMSTC†	2024	Transformer	ViT-B	77.9	94.1	65.7	88.6	67.3	89.8	57.4	68.6	72.3	67.7	83.9	45.5	54.4

A. Evaluation datasets and metrics

We conduct extensive experiments on five benchmark datasets, i.e., GTOT [19], RGBT210 [20], RGBT234 [21], LasHeR [22] and VTUAV [23], to verify the validity of our proposed tracker. Next, we introduce these datasets and their metrics in detail.

1) Evaluation datasets: We conduct extensive experiments on five benchmark datasets, i.e., GTOT [19], RGBT210 [20], RGBT234 [21], LasHeR [22] and VTUAV [23], to verify the validity of our proposed tracker. GTOT [19] is the first standard dataset for RGB-T tracking, including 50 RGB-T video sequences. RGBT210 [20] consists of 210 sequences with approximately 104.8K frames. RGBT234 [21] is a large-scale RGB-T tracking dataset. It contains 12 challenge attribute labels, including no occlusion (NO), partial occlusion (PO), heavy occlusion (HO), low illumination (LI), low resolution (LR), thermal crossover (TC), deformation (DEF), fast motion (FM), scale variation (SV), motion blur (MB), camera moving (CM) and background clutter (BC). VTUAV [23] is a large-scale benchmark for RGB-T UAV tracking (VTUAV), which includes 176 test sequences to evaluate short-term tracking and 74 sequences to evaluate long-term tracking. LasHeR [22] is currently the largest RGB-T tracking dataset, which consists of 1244 RGB-T videos with more than 730K frame pairs in total. Among them, 245 videos are used as the testing set, and 979 videos are used as the training set.

2) Evaluation Metrics: As that in [5], [7], to mitigate small alignment errors, we utilize two widely used metrics, i.e., maximum precision rate (MPR) and maximum success rate (MSR), to evaluate the tracking performance on GTOT [19], RGBT210 [20], RGBT234 [21] and VTUAV [23]. Specifically, precision rate (PR) is the percentage of frames whose output location is within a threshold distance of the ground truth. Success rate (SR) is the percentage of the frames whose overlap ratio between the output bounding box and the ground truth bounding box is larger than the threshold, and we calculate the representative SR score by the area under the curve. Owing to the modality-level displacement, we adopt the MPR and MSR to measure the tracker results. Differently, since LasHeR [22] employs a better alignment, it directly uses the PR and SR metrics to evaluate different trackers. As well, it adds an additional Normalized Precision Rate (NPR) metric to normalize the precision rate over the size of the ground truth bounding box.

B. Comparisons with State-of-the-art Methods

We quantitatively evaluate the proposed method on the above five benchmark datasets.

1) Comparison methods: To evaluate the superiority of our proposed method, we first compare our Siam-MMSTC with 9 RGB-T trackers based on MDNet [24] network, including FANet [27], MANet++ [5], ADRNet [33], TFNet [28], APFNet [34], M5L [31] and DMCNet [6], DRGCNet [29], and CAT++ [7], 4 RGB-T trackers based Siamese network, including SiamCDA [2], SiamMLAA [36], SiamTDR [3] and DFAT [71], and 4 RGB-T trackers based DCF, including JMMAC [8], HMFT [23], MFNet [35], and CMD [9]. Then, we compare our Trans-MMSTC with 14 recent Transformer based trackers, including ViPT [45], MACFT [42], SiamFEA [40], SiamAFTS [41], RSFNet [70], TBSI [12], MPLT [47],

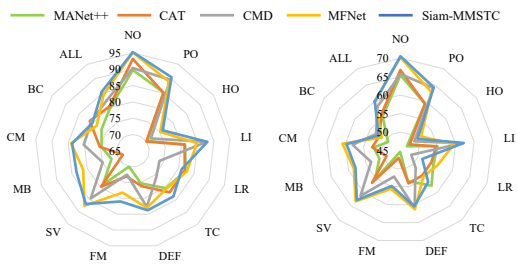
GMMT [43], TATrack [13], QueryTrack [15], SDSTrack [18], OneTrack [16], UnTrack [17] and BAT [46]. For fair comparisons, we illustrate the training data used by different methods. Here, † denotes that the model is trained on the training split of LasHeR [22]. While ∗ denotes that the model is trained on different data. For instance, SiamCDA [2] uses the generated synthetic RGB-T dataset for training. MACFT [42] is trained by using the full LasHeR dataset when testing the RGBT234 dataset. Please find more details in their corresponding papers. Besides, other unlabeled methods employ the GTOT dataset for training when tested on RGBT210 and RGBT234, and use the RGBT234 dataset for training when tested on the GTOT, VTUAV and LasHeR.

2) GTOT dataset: From Table I, we can see that our method obtains the state-of-the-art performance on GTOT dataset with 76.1% and 91.2% in MSR and MPR scores, respectively. Compared with the most recent tracker (also the second best one in this experiment), i.e., SiamMLAA [36], our algorithm achieves 1.5% improvements in MSR score. Besides, our Trans-MMSTC further improves the tracking performance and significantly outperforms all RGB-T trackers in MPR score. The exciting performance and significant promotion demonstrate the effectiveness of our proposed framework.

3) RGBT210 dataset: As shown in Table I, compared with these CNN based trackers, Siam-MMSTC achieves the best results with 60.1% and 85.5% in MSR and MPR scores, respectively. Siam-MMSTC outperforms APFNet by 3.4% and 3.0% in MPR and MSR, respectively. What's more, Trans-MMSTC achieves new state-of-the-art tracking performance with 65.7% and 88.6% in MSR and MPR scores, respectively. Compared the the second best tracker MPLT [47], Trans-MMTSC obtains 2.4%/2.7% gainsin MPR and MSR, respectively.

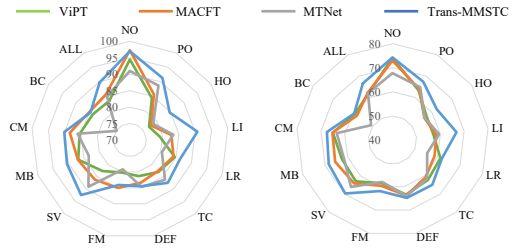
4) RGBT234 dataset: Table I reports the MPR and MSR scores of these trackers. The proposed Siam-MMSTC still achieves promising performance with the best MPR score of 85.5% and the second best MSR score of 59.9% among these CNN based trackers. In comparison with the MFNet [35], Siam-MMSTC achieves performance gains of 1.1% in MPR. Additionally, compared with recent RGB-T trackers based on Transformer, our proposed Trans-MMTSC achieves the best performance with 89.8% and 67.3% in MPR and MSR scores, respectively. Trans-MMTSC achieves 6.3%/5.6%, 2.7%/3.6% and 1.4%/1.6% improvements against ViPT [11], TBSI [12] and MPLT [47] in MPR/MSR, respectively. The favorable performance against these state-of-the-art trackers validates the effectiveness of our proposed MMSTC framework.

5) LasHeR dataset: From Table I, we can also see that our tracker achieves excellent performance on LasHeR. Siam-MMSTC's PR/SR is 3.1%/0.9% higher than those of CMD [9]. Compared with APFNet [34] and FANet [27], Siam-MMSTC advances them with 12.1%/11.1% and 13.9%/13.0% in PR/SR, respectively, which proves the huge performance advantage of our method. Compared with trackers based on Transformer, such as SiamFEA [40], RSFNet [70], ViPT [45] and TBSI [12], which have already achieved outstanding tracking performance, our method achieves performance gains of 7.8%/6.5%, 6.4%/4.8%, 7.2%/4.9% and 3.1%/1.8% in PR/SR, respectively. The excellent tracking results achieved on LasHeR also demonstrate that our proposed tracker has better generalization ability than others.



(a) The MPR scores of different attributes. (b) The MSR scores of different attributes.

Fig. 7. The MPR and MSR scores of the proposed Siam-MMSTC and other CNN based trackers under different attributes on RGBT234.



(a) The MPR scores of different attributes. (b) The MSR scores of different attributes.

Fig. 8. The MPR and MSR scores of the proposed Trans-MMSTC and other Transformer based trackers under different attributes on RGBT234.

6) VTUAV dataset: As shown in Table I, we evaluated our Siam-MMSTC and Trans-MMTC on both short-term and long-term tracking subsets of the VTUAV dataset. In the short-term subset of VTUAV, compared with HMFT, which employs the training subset in VTUAV for model training, our Siam-MMSTC still obtains the competitive results. Compared with trackers based on Transformer, such as MACFT [42], BAT [46] and MPLT [47], which have already achieved outstanding tracking performance, our method achieves performance gains of 0.9%/3.8%, 3.6%/4.5% and 2.3%/4.2% in MPR/MSR, respectively. In the long-term subset of VTUAV, without any additional re-detection mechanism, our Trans-MMSTC shows excellent tracking performance with 54.4% and 45.5% in MPR and MSR scores, respectively.

7) Attribute-based performance: To further demonstrate the effectiveness of our proposed method, we plot the attribute-based performance on RGBT234 [21], which contains 12 challenging attribute labels. As shown in Fig. 7, our Siam-MMSTC achieves the best performance on most challenges, e.g., NO, PO, HO, LI, FM, TC and DEF. Additionally, our Siam-MMSTC shows very competitive results under the challenges of SV, CM and BC. Besides, regarding the LR challenge, superior results are attained by CAT and MANe++ due to their utilization of multi-scale feature fusion. In contrast, our Siam-MMSTC relies solely on a simple concatenation operation to fuse features from the last two levels, which restricts its effectiveness in low-resolution scenarios. From Fig. 8, we can see that with a more powerful baseline tracker, our Trans-MMSTC obtains the best performance in the challenges of PO, HO, LI, LR, TC, SV, MB and CM. Under the challenges of NO, DEF, FM and BC, our proposed Trans-MMSTC obtains competitive results compared with those state-of-the-

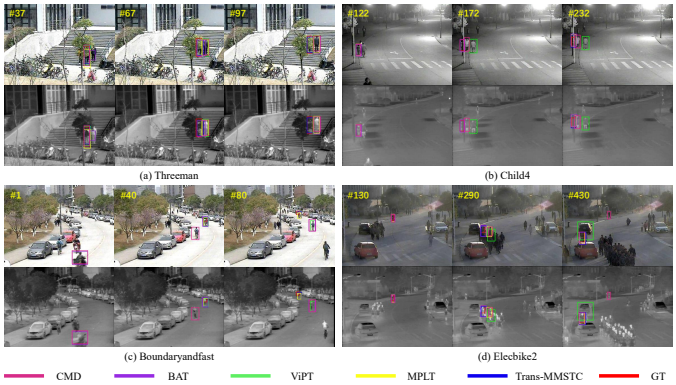


Fig. 9. Visual comparisons of our proposed tracker with another four state-of-the-art trackers.

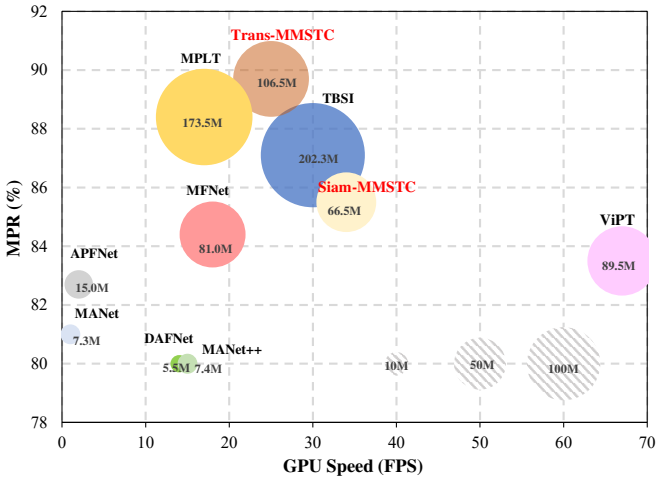


Fig. 10. The running speeds and parameters of several RGB-T tracking methods on RGBT234 dataset.

art trackers.

8) Qualitative performance: The visual comparisons between our proposed Trans-MMSTC and some state-of-the-art trackers, including ViPT [45], MPLT [47], CMD [9] and BAT [46], are illustrated in Fig. 9. Our approach performs obviously better than other methods in various complex scenarios, such as similar appearance, occlusion and motion blur. Thanks to the multi-modal spatial contexts as well as the temporal contexts, our Trans-MMSTC can track the targets accurately in these cases.

9) Tracking Speed and parameters: We compare the running speed and parameters of the proposed approach with those of state-of-the-art algorithms on GTX 3090 GPU. Fig.10 reports the running speeds and parameters of these methods. Compared with the transformer-based trackers, our Trans-MMSTC improves tracking performance while using 46% fewer parameters than TBSI [12]. Compared with the CNN-based trackers, although our Siam-MMSTC has more parameters, its running speed is still faster than most CNN-based trackers and obtains significant improvements on all datasets.

C. Ablation study

We conduct some ablation studies on RGB-T234 [21] to discuss the impacts of different components in our Siam-MMSTC tracker.

TABLE II
EXPERIMENTAL RESULTS OF DIFFERENT VARIANTS OF MMTE ON RGBT234 DATASET, SIAM-BASELINE DENOTES THE BASELINE TRACKER BASED ON SIAMESE NETWORK.

Siam-Baseline	Intra-modal Context	Cross-modal Context	MPR ↑	MSR ↑
✓			80.1	57.5
✓	✓		83.4	58.9
✓		✓	83.7	59.3
✓	✓	✓	84.0	59.5

1) Effectiveness of MMTE: Here, in order to discuss the effectiveness of MMTE, we merely utilize a Transformer encoder to promote the feature fusion of RGB features and thermal features without using the QATD module in this subsection. For that, we first take the variant as the baseline, where the intra-modal context information and the cross-modal context information are not considered, and the multi-modal features from the RGB and thermal images are directly fused by using the summation operation for tracking. Then we construct two variants that only model the intra-modal contexts or the cross-modal contexts. Especially, the former just directly calculates the reliable context information within RGB modality and thermal modality, respectively, and the latter only embeds the cross-modal context information into RGB and thermal features. Table II shows the performance of the three counterparts (the first three rows) and our proposed MMTE module (the fourth row) on the RGBT234 dataset. Compared with the baseline, the variants only modeling intra-modal or cross-modal contexts achieve performance gains of 3.3%/1.4% and 3.6%/1.8% in MPR/MSR on the RGBT234 dataset, respectively, which demonstrates that modeling the intra-modal or cross-modal contexts alone can also improve the tracking performance to some extent. In comparison with the variants only modeling the intra-modal or cross-modal contexts, our proposed MMTE module further improves the performance by 3.9%/2.0% in MPR/MSR on the RGBT234 dataset. These performance gains validate the effectiveness of jointly modeling the intra-modal and cross-modal contexts.

In addition, as shown in Fig. 11, we also obviously observe that the proposed MMTE module can enhance the effectiveness of unimodal spatial contexts by virtue of the cross-modal contexts, thus fully exploring reliable spatial context information within multi-modal data.

Fig. 11. Visualization of some attention maps for the intra-modal contexts, cross-modal contexts and enhanced unimodal contexts.

2) Comparisons with different methods for exploring multi-modal spatial contexts: To further validate the effectiveness of our proposed MMTE, we employ another 2 methods to explore multi-modal spatial contexts in our tracker, including 1) the inter-modal pattern-propagation module (IMPP) in CMPP [51]; 2) the global mining module (GMM) in AGMINet [52]. We use the two modules to replace the MMTE in

TABLE III
EXPERIMENT RESULTS OF DIFFERENT METHODS FOR EXPLORING MULTI-MODAL SPATIAL CONTEXTS, SIAM-BASELINE DENOTES THE BASELINE TRACKER BASED ON SIAMESE NETWORK.

Siam-Baseline	IMPP	GMM	MMTE	MPR ↑	MSR ↑
✓	✓			81.3	58.0
✓		✓		82.4	58.1
✓			✓	84.0	59.5

TABLE IV
EXPERIMENT RESULTS OF DIFFERENT VARIANTS OF QATD ON RGBT234 DATASET, SIAM-MMTE DENOTES THE BASELINE TRACKER BASED ON SIAMESE NETWORK EQUIPPED WITH THE PROPOSED MMTE MODULE.

Siam-MMTE	Target context	Background context	Target context + QSM	Background context + QSM	MPR ↑	MSR ↑
✓					84.0	59.5
✓	✓				84.3	59.6
✓		✓			81.2	57.0
✓	✓	✓			83.8	58.9
✓			✓	✓	85.1	59.6
✓	✓			✓	85.5	59.9

Siam-MMSTC, respectively. According to the experimental data in Table III, it can be seen that our proposed MMTE outperforms the other two fusion modules significantly. This may be attributed to the fact that the proposed MMTE can better suppress such unreliable spatial relationships within multi-modal data by using cross-modal reliability and cross-modal contexts, while IMPP and GMM ignore the influence of such unreliable spatial context information on the tracking results. By virtue of MMTE, those reliable spatial contexts within multi-modal data can be well captured for tracking.

3) Effectiveness of QATD: To analyze the effectiveness of our proposed QATD module, we evaluate two variants that only use target-related contexts or background-related contexts to propagate temporal information among different frames. Table IV shows the performance of the two counterparts (the first two rows) and our proposed QATD module (the 3rd row) on the RGBT234 dataset. The comparison between the variant using target-related context information and the baseline with the MMTE modules shows that the temporal target information can slightly improve the tracking performance. However, the variant that only propagates background-related context information temporally obtains performance drops of 2.8%/2.5% in MPR/MSR on the RGBT234 dataset. This may be due to the fact that the background scenes usually change drastically in a video, which is unreasonable to totally propagate all of the background-related contexts across frames. With the quality-aware select module (QSM) in the QATD block, the baseline method obtains a notable performance gain of 1.5%/0.4% in MPR/MSR on the RGBT234 dataset. When we transfer all of the background-related contexts, tracking performance drops significantly. And when we further select the target-related contexts, the tracking accuracy does not improve.

Additionally, as shown in Fig. 12, these background-related contexts with noisy information negatively affect the representation ability of decoded features. With the proposed quality-aware select module, the decoded features can better distinguish between foregrounds and backgrounds. This indicates that our proposed QATD module is crucial for effectively modeling the temporal context information since it mainly

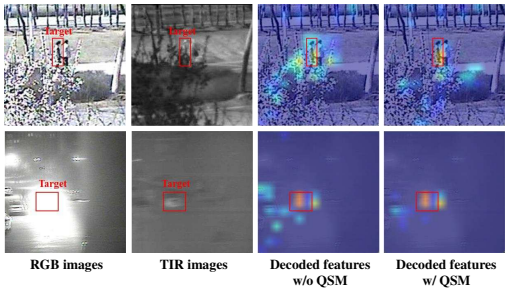


Fig. 12. Illustration of the effectiveness of the proposed quality-aware select module.

TABLE V
EXPERIMENT RESULTS OF DIFFERENT VARIANTS OF TSE ON RGBT234 DATASET.

Baseline	Online Cls	Offline Cls	MPR ↑	MSR ↑
✓	✓		84.0	58.9
✓		✓	79.8	56.8
✓	✓	✓	85.5	59.9

delivers reliable context information across frames.

4) Effectiveness of TSE: In order to evaluate the effectiveness of our proposed TSE module, we only consider the utilization of multi-modal spatial-temporal context information as well as appearance information here. For that, we construct another two variants of our proposed model that only use an online-trained classifier or an offline-trained classifier. Table V shows the results of the two variants and the Siam-MMSTC model on the RGBT234 dataset. We observe that the online-trained classifier is superior to the simple offline-trained classifier. Nevertheless, in the experiments, we show that with the help of appearance information within the offline-trained classifier, our proposed tracker is able to achieve more performance improvements.

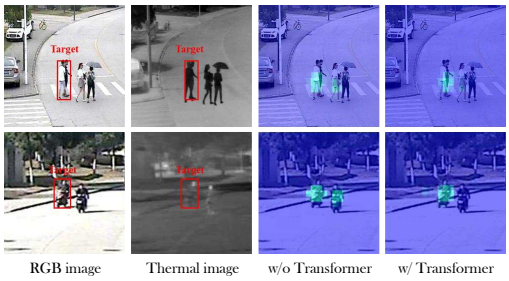


Fig. 13. Visualization of some tracking response maps of the Siam-MMSTC tracker. The 'w/o Transformer' denotes the baseline approach without MMTE and QATD modules. The 'w/ Transformer' denotes the baseline with MMTE as well as QATD modules.

5) Response Visualization: In Fig. 13, we exhibit more detailed visualization results of our tracking framework. From Fig. 13, we can observe that our baseline (the 3rd column) tends to be misled by such distracting objects in the challenging scenarios. By adopting the proposed MMTE and QATD modules (the 4th column), the target representations in the search region are effectively reinforced, which facilitates the object searching process. Therefore, the response values of background regions are largely suppressed.

6) Generality of MMSTC: To verify the generality of our proposed MMSTC framework, we further construct an RGB-T tracker based on a typical DCF tracker DiMP [26], denoted as

TABLE VI
PERFORMANCE IMPROVEMENTS OF OUR MMSTC FRAMEWORK ON THE BASELINE TRACKER DIMP.

DiMP	MMTE	QATD	MPR ↑	MSR ↑
✓			82.6	58.5
✓	✓		83.4	59.1
✓	✓	✓	84.4	59.4

DiMP-MMSTC. Following DiMP, RGB images and thermal images will be fed to two separated feature extractors (i.e. an RGB feature extractor and a Thermal feature extractor), respectively. Then, the encoded features are obtained by performing the MMTE module on the extracted RGB and thermal features. After that, the encoded features of the current frame are enhanced by using the proposed QATD module. Finally, the decoded features will be fed into the classification head of DiMP, thus obtaining the final tracking results. The experiment results on RGBT234 are shown in Table VI. MMTE and QATD consistently improve the baseline tracker in terms of MPR and MSR. Compared with some state-of-the-art trackers, DiMP-MMSTC also achieves more excellent tracking performance. This fully demonstrates the effectiveness of our proposed MMSTC framework on various types of trackers.

TABLE VII
ANALYSIS OF THE MODEL SIZE AND RUNNING SPEED OF THE PROPOSED MMSTC FRAMEWORK. SIAM-BASELINE AND TRANS-BASELINE DENOTE THE BASELINE TRACKER BASED ON SIAMESE NETWORK AND TRANSFORMER FRAMEWORK, RESPECTIVELY.

	MPR(%)	MSR(%)	FPS	Model size (MB)
Siam-Baseline	80.1	57.5	42	60.4
Siam-MMSTC	85.5	59.9	34	66.5
Trans-Baseline	85.0	63.0	36	92.5
Trans-MMSTC	89.8	67.3	27	106.5

7) Model size and running speed: As shown in Table VII, compared with the baseline method, which only employs the concatenation operation for multi-modal feature fusion, our proposed MMSTC framework introduces fewer parameters while maintaining real-time running speed.

(a) (b)

Fig. 14. Failure cases on two sequences.

8) Failure cases: Fig. 14 presents several tracking failure cases encountered by our Siam-MMSTC. In Fig. 14 (a), we noticed that our Siam-MMSTC fails to re-locate the target when it is out of view or occlusion. Since our tracker lacks a re-detection module, accurately relocating the target after occlusion or out of view poses a significant challenge. This issue is exacerbated by the presence of distractors within the scene, which often leads the proposed tracker to mistakenly select a distractor as the new target post-occlusion, thereby generating an erroneous tracking trajectory. In Fig. 14 (b), when the movement trajectories of the target and distractors intersect, our tracker struggles to accurately locate the target. In the future, we intend to delve deeper into enhancing tracking performance in these aforementioned scenarios.

VI. CONCLUSION

In this paper, we have presented a multi-modal spatial-temporal context network for RGB-T tracking, in which the encoder-decoder Transformer architecture is used for the construction of multi-modal spatial context information and the effective propagation of temporal context information. By virtue of the proposed MMTE module, we obtain high-quality encoded features, which simultaneously explore the reliable spatial context information within multi-modal data and integrate multi-modal features. Besides, by employing the proposed QATD module, the encoded search features of the current frame can be enhanced by selectively propagating the temporal context information stored in the memory pool. Such two types of contexts enhance the discriminative ability of a tracker, contributing to more accurate and robust tracking results. The comprehensive ablation studies validate the effectiveness of each component, and the favorable performance against some state-of-the-art trackers on five benchmark datasets demonstrates the effectiveness of our proposed algorithm.

In the future, we aim to advance our research on efficient multi-modal long-term tracking strategies to address the challenge of target relocation after occlusion, out-of-view and intersect of the target and distractor trajectories. Specifically, we propose that these challenges can be effectively addressed through the implementation of two additional strategies. First, the integration of a global re-detection module would facilitate the relocation of the target following occlusion. Second, the adoption of a multi-object tracking paradigm would enhance the differentiation of the movement trajectories of other distractors in the scene. The two approaches are expected to mitigate the risk of the tracker drifting towards distractors after the target has been occluded. Additionally, we plan to expand our multi-modal tracking framework to incorporate more modalities, including text, point clouds, and event data, to enhance overall tracking performance.

REFERENCES

[1] M. Feng and J. Su, "Rgbt tracking: A comprehensive review," *Information Fusion*, p. 102492, 2024.
[2] T. Zhang, X. Liu, Q. Zhang, and J. Han, "Siamcda: Complementarity- and distractor-aware rgb-t tracking based on siamese network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1403–1417, 2021.
[3] G. Wang, Q. Jiang, X. Jin, Y. Lin, Y. Wang, and W. Zhou, "Siamtdr: Time-efficient rgbt tracking via disentangled representations," *IEEE Transactions on Industrial Cyber-Physical Systems*, vol. 1, pp. 167–181, 2023.
[4] T. Zhang, X. He, Y. Luo, Q. Zhang, and J. Han, "Exploring target-related information with reliable global pixel relationships for robust RGB-T tracking," *Pattern Recognition*, p. 110707, 2024.
[5] A. Lu, C. Li, Y. Yan, J. Tang, and B. Luo, "Rgbt tracking via multi-adapter network with hierarchical divergence loss," *IEEE Transactions on Image Processing*, vol. 30, pp. 5613–5625, 2021.
[6] A. Lu, C. Qian, C. Li, J. Tang, and L. Wang, "Duality-gated mutual condition network for rgbt tracking," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[7] L. Liu, C. Li, Y. Xiao, R. Ruan, and M. Fan, "Rgbt Tracking via Challenge-Based Appearance Disentanglement and Interaction," *IEEE Transactions on Image Processing*, vol. 33, pp. 1753–1767, 2024.

[8] P. Zhang, J. Zhao, C. Bo, D. Wang, H. Lu, and X. Yang, "Jointly modeling motion and appearance cues for robust rgb-t tracking," *IEEE Transactions on Image Processing*, vol. 30, pp. 3335–3347, 2021.

[9] T. Zhang, H. Guo, Q. Jiao, Q. Zhang, and J. Han, "Efficient rgb-t tracking via cross-modality distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5404–5413.

[10] T. Zhang, X. He, Q. Jiao, Q. Zhang, and J. Han, "Amnet: Learning to Align Multi-modality for RGB-T Tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[12] T. Hui, Z. Xun, F. Peng, J. Huang, X. Wei, X. Wei, J. Dai, J. Han, and 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, pp. 13 630–13 639.

[13] H. Wang, X. Liu, Y. Li, M. Sun, D. Yuan, and J. Liu, "Temporal Adaptive RGBT Tracking with Modality Prompt." in *AAAI Conference on Artificial Intelligence (AAAI)*, 2024, pp. 5436–5444.

[14] J. Xia, D. Shi, K. Song, L. Song, X. Wang, S. Jin, L. Zhou, Y. Cheng, L. Jin, Z. Zhu *et al.*, "Unified single-stage transformer network for efficient rgb-t tracking," *arXiv preprint arXiv:2308.13764*, 2023.

[15] H. Fan, Z. Yu, Q. Wang, B. Fan, and Y. Tang, "Querytrack: Joint-Modality Query Fusion Network for RGBT Tracking," *IEEE Transactions on Image Processing*, vol. 33, pp. 3187–3199, 2024.

[16] 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, pp. 19 079–19 091.

[17] Z. Wu, J. Zheng, X. Ren, F.-A. Vasluianu, C. Ma, D. P. Paudel, L. Van Gool, and 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, pp. 19 156–196.

[18] 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, pp. 26 551–26 561.

[19] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal tracking," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5743–5756, 2016.

[20] C. Li, N. Zhao, Y. Lu, C. Zhu, and J. Tang, "Weighted sparse representation regularized graph learning for rgb-t object tracking," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1856–1864.

[21] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: Benchmark and baseline," *Pattern Recognition*, vol. 96, p. 106977, 2019.

[22] C. Li, W. Xue, Y. Jia, Z. Qu, B. Luo, J. Tang, and D. Sun, "Lasher: A large-scale high-diversity benchmark for rgbt tracking," *IEEE Transactions on Image Processing*, vol. 31, pp. 392–404, 2022.

[23] P. Zhang, J. Zhao, D. Wang, H. Lu, and X. Ruan, "Visible-thermal uav tracking: A large-scale benchmark and new baseline," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8886–8895.

[24] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 20, pp. 4293–4302.

[25] H. Tan, X. Zhang, Z. Zhang, L. Lan, W. Zhang, and Z. Luo, "Nocal-siam: Refining visual features and response with advanced non-local blocks for real-time siamese tracking," *IEEE Transactions on Image Processing*, vol. 30, pp. 2656–2668, 2021.

[26] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6182–6191.

[27] Y. Zhu, C. Li, J. Tang, and B. Luo, "Quality-aware feature aggregation network for robust rgbt tracking," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 1, pp. 121–130, 2021.

[28] Y. Zhu, C. Li, J. Tang, B. Luo, and L. Wang, "Rgbt tracking by trident fusion network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 579–592, 2021.

[29] J. Mei, D. Zhou, J. Cao, R. Nie, and K. He, "Differential reinforcement and global collaboration network for rgbt tracking," *IEEE Sensors Journal*, vol. 23, no. 7, pp. 7301–7311, 2023.

[30] Y. Cai, X. Sui, and G. Gu, "Multi-modal multi-task feature fusion for rgbt tracking," *Information Fusion*, vol. 97, p. 101816, 2023.

[31] Z. Tu, C. Lin, W. Zhao, C. Li, and J. Tang, "M5l: Multi-modal multi-margin metric learning for rgbt tracking," *IEEE Transactions on Image Processing*, vol. 31, pp. 85–98, 2022.

[32] C. Li, L. Liu, A. Lu, Q. Ji, and J. Tang, "Challenge-aware rgbt tracking," in *European Conference on Computer Vision*. Springer, 2020, pp. 222–237.

[33] P. Zhang, D. Wang, H. Lu, and X. Yang, "Learning adaptive attribute-driven representation for real-time rgb-t tracking," *International Journal of Computer Vision*, vol. 129, pp. 2714–2729, 2021.

[34] Y. Xiao, M. Yang, C. Li, L. Liu, and J. Tang, "Attribute-based progressive fusion network for rgbt tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2831–2838.

[35] Q. Zhang, X. Liu, and T. Zhang, "Rgb-t tracking by modality difference reduction and feature re-selection," *Image and Vision Computing*, vol. 127, p. 104547, 2022.

[36] M. Feng and J. Su, "Learning multi-layer attention aggregation siamese network for robust rgbt tracking," *IEEE Transactions on Multimedia*, vol. 26, pp. 3378–3391, 2024.

[37] Z. Tang, T. Xu, H. Li, X.-J. Wu, X. Zhu, and J. Kittler, "Exploring fusion strategies for accurate rgbt visual object tracking," *Information Fusion*, p. 101881, 2023.

[38] J. Peng, H. Zhao, Z. Hu, Y. Zhuang, and B. Wang, "Siamese infrared and visible light fusion network for rgb-t tracking," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 9, pp. 3281–3293, 2023.

[39] M. Feng and J. Su, "Learning reliable modal weight with transformer for robust rgbt tracking," *Knowledge-Based Systems*, vol. 249, p. 108945, 2022.

[40] L. Feng, K. Song, J. Wang, and Y. Yan, "Exploring the potential of siamese network for rgbt object tracking," *Journal of Visual Communication and Image Representation*, p. 103882, 2023.

[41] L. Fan and P. Kim, "Anchor free based siamese network tracker with transformer for rgb-t tracking," *Scientific Reports*, vol. 13, no. 1, p. 13294, 2023.

[42] Y. Luo, X. Guo, M. Dong, and J. Yu, "Learning modality complementary features with mixed attention mechanism for rgb-t tracking," *Sensors*, vol. 23, no. 14, p. 6609, 2023.

[43] Z. Tang, T. Xu, X. Wu, X. Zhu, and J. Kittler, "Generative-Based Fusion Mechanism for Multi-Modal Tracking." in *AAAI Conference on Artificial Intelligence (AAAI)*, 2024, pp. 5189–5197.

[44] B. Ye, H. Chang, B. Ma, S. Shan, and X. Chen, "Joint feature learning and relation modeling for tracking: A one-stream framework," in *European Conference on Computer Vision*. Springer, 2022, pp. 341–357.

[45] J. Zhu, S. Lai, X. Chen, D. Wang, and H. Lu, "Visual prompt multi-modal tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9516–9526.

[46] B. Cao, J. Guo, P. Zhu, and Q. Hu, "Bi-directional adapter for multi-modal tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, 2024, pp. 927–935.

[47] Y. Luo, X. Guo, H. Feng, and L. Ao, "Rgb-t tracking via multi-modal mutual prompt learning," *arXiv preprint arXiv:2308.16386*, 2023.

[48] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8126–8135.

[49] Y. Cui, C. Jiang, L. Wang, and G. Wu, "Mixformer: End-to-end tracking with iterative mixed attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 608–13 618.

[50] Y. Cui, T. Song, G. Wu, and L. Wang, "Mixformerv2: Efficient fully transformer tracking," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[51] C. Wang, C. Xu, Z. Cui, L. Zhou, T. Zhang, X. Zhang, and J. Yang, "Cross-modal pattern-propagation for rgb-t tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7064–7073.

[52] J. Mei, Y. Liu, C. Wang, D. Zhou, R. Nie, and J. Cao, "Asymmetric global–local mutual integration network for rgbt tracking," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–17, 2022.

[53] J. Zhang, S. Ma, and S. Sclaroff, "Meem: robust tracking via multiple experts using entropy minimization," in *Proceedings of the European conference on computer vision (ECCV)*. Springer, 2014, pp. 188–203.

[54] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M.-H. Yang, "Hedged deep tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4303–4311.
[55] Z. Hu, Y. Gao, D. Wang, and X. Tian, "A universal update-pacing framework for visual tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 1704–1708.
[56] C. Mayer, M. Danelljan, G. Bhat, M. Paul, D. P. Paudel, F. Yu, and L. Van Gool, "Transforming model prediction for tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8731–8740.
[57] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, "Learning spatio-temporal transformer for visual tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 448–10 457.
[58] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Know your surroundings: Exploiting scene information for object tracking," in *European Conference on Computer Vision*. Springer, 2020, pp. 205–221.
[59] N. Wang, W. Zhou, J. Wang, and H. Li, "Transformer meets tracker: Exploiting temporal context for robust visual tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1571–1580.
[60] T. Zhang, Z. Jin, K. Debattista, Q. Zhang, and J. Han, "Enhancing visual tracking with a unified temporal Transformer framework," *IEEE Transactions on Intelligent Vehicles*, pp. 1–15, 2024.
[61] Y. Bai, Z. Zhao, Y. Gong, and X. Wei, "Artrackv2: Prompting autoregressive tracker where to look and how to describe," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 048–19 057.
[62] W. Cai, Q. Liu, and Y. Wang, "Hiptrack: Visual tracking with historical prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 258–19 267.
[63] Z. Zhou, W. Pei, X. Li, H. Wang, F. Zheng, and Z. He, "Saliency-associated object tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9866–9875.
[64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[65] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9626–9635.
[66] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 740–755.
[67] L. Huang, X. Zhao, and K. Huang, "Got-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1562–1577, 2021.
[68] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling, "LaSOT: A high-quality benchmark for large-scale single object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5374–5383.
[69] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
[70] Z. Yu, H. Fan, Q. Wang, Z. Li, and Y. Tang, "Region selective fusion network for robust rgb-t tracking," *IEEE Signal Processing Letters*, vol. 30, pp. 1357–1361, 2023.
[71] M. Li, P. Zhang, M. Yan, H. Chen, and C. Wu, "Dynamic feature-memory transformer network for rgbt tracking," *IEEE Sensors Journal*, vol. 23, no. 17, pp. 19 692–19 703, 2023.

Tianlu Zhang received the B. S. degree from Xi'an Shiyou University, Xi'an, China, in 2018. He is currently pursuing the Ph.D. degree in School of Mechano-Electronic Engineering, Xidian University, China. He is also a Visiting Student at the University of Warwick. His research interests include deep learning and multi-modal image processing in computer vision.

Qiang Jiao received the B.S. and Ph.D. degrees from the Nanjing University of Science and Technology, Nanjing, China, in 2010 and 2017, respectively. He is currently working with the School of Mechano-electronic Engineering, Xidian University, Xian, China. His current research interests include deep learning, image processing and pattern recognition.

Qiang Zhang received the B.S. degree in automatic control, the M.S. degree in pattern recognition and intelligent systems, and the Ph.D. degree in circuit and system from Xidian University, China, in 2001,2004, and 2008, respectively. He is currently a professor with the Automatic Control Department, Xidian University, China. His current research interests include image processing and pattern recognition.

Jungong Han (Senior Member, IEEE) is the Chair Professor in Computer Vision at the Department of Computer Science, University of Sheffield, U.K. He also holds an Honorary Professorship with the University of Warwick, U.K. His research interests include computer vision, artificial intelligence, and machine learning. He is a Fellow of the International Association of Pattern Recognition, and serves as the Associate Editor for many prestigious journals, such as IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Multimedia, and Pattern Recognition.