Cross-modulated Attention Transformer for RGBT Tracking

Yun Xiao[1,4,5], **Jiacong Zhao**[1], **Andong Lu**[2], **Chenglong Li**[1,4,5*], **Bing Yin**[3], **Yin Lin**[3], **Cong Liu**[3]

[1] School of Artificial Intelligence, Anhui University, Hefei, China
[2] School of Computer Science and Technology, Anhui University, Hefei, China
[3] iFLYTEK CO.LTD., Hefei, China
[4] Anhui Provincial Key Laboratory of Security Artificial Intelligence, Hefei, China
[5] Information Materials and Intelligent Sensing Laboratory of Anhui Province, Hefei, China
xiaoyun@ahu.edu.cn, JiacongZhao2022@163.com, adlu_ah@foxmail.com, lcl1314@foxmail.com, bingyin@iflytek.com, yinlin@iflytek.com, congliu2@iflytek.com

Abstract

Existing Transformer-based RGBT trackers achieve remarkable performance benefits by leveraging self-attention to extract uni-modal features and cross-attention to enhance multi-modal feature interaction and search-template correlation. Nevertheless, the independent search-template correlation calculations are prone to be affected by low-quality data, which might result in contradictory and ambiguous correlation weights. It not only limits the intra-modal feature representation, but also harms the robustness of cross-attention for multi-modal feature interaction and search-template correlation computation. To address these issues, we propose a novel approach called Cross-modulated Attention Transformer (CAFormer), which innovatively integrates inter-modality interaction into the search-template correlation computation within typical attention mechanism, for RGBT tracking. In particular, we first independently generate correlation maps for each modality and feed them into the designed correlation modulated enhancement module, which can modify inaccurate correlation weights by seeking the consensus between modalities. Such kind of design unifies self-attention and cross-attention schemes, which not only alleviates inaccurate attention weight computation in self-attention but also eliminates redundant computation introduced by extra cross-attention scheme. In addition, we design a collaborative token elimination strategy to further improve tracking inference efficiency and accuracy. Experiments on five public RGBT tracking benchmarks show the outstanding performance of the proposed CAFormer against state-of-the-art methods.

Code — https://github.com/opacity-black/CAFormer

Introduction

RGBT tracking (Li et al. 2020, 2021; Lu et al. 2022; Hui et al. 2023; Hou et al. 2024), which fuses information from visible and thermal infrared (TIR) modalities for visual tracking, has become an active research field in the computer vision community. Recently, with the success of Transformers in visual object tracking (Chen et al. 2021; Ye et al. 2022; Chen et al. 2022), Transformer-based RGBT trackers have gradually gained advantages in both speed and performance.

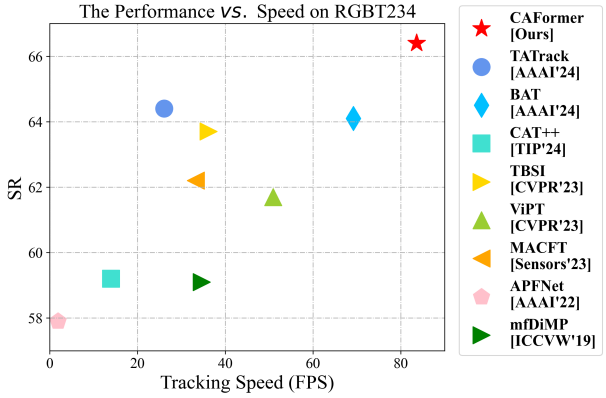


Figure 1: Comparison of performance and speed for state-of-the-art tracking methods on RGBT234 (Li et al. 2019a). We visualize the Success Rate (SR) against Frames Per Second (FPS). Closer to the top indicates higher performance, and closer to the right indicates faster speed. CAFormer ranks first in SR while running at 83.6 FPS.

The Transformer is successfully applied in RGBT tracking due to its attention mechanism, which allows it to selectively focus on relevant information and ignore irrelevant information. Existing Transformer-based RGBT trackers (Hui et al. 2023; Hou et al. 2024; Hou, Ren, and Wu 2022) achieve remarkable performance by leveraging self-attention to extract uni-modal features and cross-attention to enhance multi-modal feature interaction. However, we observe that the calculation of correlations in self-attention is sensitive to low-quality data. Even state-of-the-art trackers (Hui et al. 2023; Luo et al. 2023) that use feature fusion suffer from the same problem, resulting in ambiguous and inappropriate correlation weights. This phenomenon is illustrated in Figure 2. Importantly, existing works (Ye et al. 2022; Gao et al. 2022; Fu et al. 2022) suggest that proper correlation is crucial for tracking. Therefore, we believe there are limitations in the modality self-attention independent modeling strategy widely adopted in existing methods. This limitation not only impairs intra-modality feature rep-

*Chenglong Li is the corresponding author.

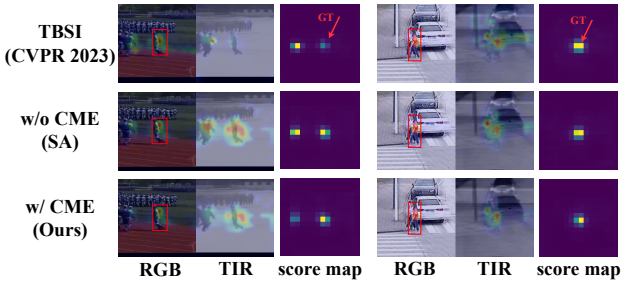


Figure 2: Illustration of correlation maps with different fusion methods under different modal quality inputs. The score map is the output of the location branch in the tracking head. "GT" denotes ground truth response position.

resentation but also affects subsequent multi-modal feature interactions and the robustness of template and search cross-correlation. Moreover, the individual computation of self-attention and cross-attention in existing methods introduces redundant computation, limiting the speed of current RGBT trackers.

To address these issues, we propose a novel approach called the Cross-modulated Attention Transformer (CAFormer) for RGBT tracking, which performs intra-modality feature extraction and inter-modality feature interaction in a unified attention model. Visible and infrared image pairs in RGBT tracking are highly spatio-temporally aligned, so their correlations between search frames and target templates should also be consistent. Consequently, different modality self-correlations exhibit similar interaction properties with multi-modal image features. Therefore, we propose an intuitive idea of enhancing and correcting low-quality modal correlations through high-quality modal correlations. To adapt to the dynamic changes in modal quality during RGBT tracking, we design a Cross-Modulated Attention (CMA) mechanism in both directions to achieve adaptive correlation modulation.

In particular, we first compute the correlation maps for each modality independently and then feed them into the designed Correlation Modulation Enhancement (CME) module for cross-correlation modeling to achieve correlation agreement between the two modalities. This approach corrects inaccurate correlation relationships from previous self-attention mechanisms, as shown in the third row of Figure 2. Additionally, CMA is more efficient in fusion. For example, in the ViT-base (Dosovitskiy et al. 2021) feature fusion module, the dimension size of input features is 768. In contrast, CMA only processes the search-template part of the correlation map, with correlation vector dimensions typically related to the number of template tokens, usually 64. By avoiding the computation of higher-dimensional features, CAFormer significantly outperforms existing feature fusion methods in terms of efficiency.

In summary, the proposed CMA unifies self-attention and cross-attention schemes, mitigating inaccurate correlations in self-attention and avoiding the computational burden of additional cross-attention. Inspired by the candidate elimination method in OSTrack (Ye et al. 2022), we propose a collaborative token elimination strategy to further enhance tracking inference efficiency and accuracy. Specifically, within the search region, we consider each token as a potential candidate for the target and treat each template token as part of the target object. Using prior knowledge about the similarity between the target and each candidate from individual modality branches, we sum the similarities of both modalities to obtain the overall similarity, then eliminate tokens with lower similarity. This approach coordinates initial elimination results from both modalities to improve background elimination precision. Consequently, our module not only enhances tracking efficiency but also maintains robust performance. Figure 1 compares CAFormer with existing state-of-the-art methods in terms of tracking accuracy and speed, demonstrating CAFormer's superiority in both metrics. The contributions of this paper can be summarized as follows:

- We propose a novel cross-modulated attention Transformer for accurate and efficient RGBT tracking. To the best of our knowledge, it is the first work to perform intra-modality self-correlation, inter-modality feature interaction, and search-template correlation computation in a unified attention for RGBT tracking.
- Through taking both the attention weights from two modalities into account, we introduce a collaborative token elimination strategy to improve the inference efficiency with further performance enhancement.
- The proposed method achieves state-of-the-art results on all mainstream RGBT tracking datasets with an impressive speed of 83.6 FPS.

Related Work

Attention Mechanism

Attention mechanisms have been widely used in computer vision tasks over the past decade (Vaswani et al. 2017; Hu, Shen, and Sun 2018; Wang et al. 2018; Guo et al. 2022). Among these, the Transformer (Vaswani et al. 2017) is favored for its powerful representation of self-attention and cross-attention. Existing attention studies can be broadly classified into two categories. One category focuses on lightweight attention mechanisms (Zhu et al. 2020a; Liu et al. 2021; Schlatt, Fröbe, and Hagen 2024; Zhou et al. 2024). For example, the Swin Transformer (Liu et al. 2021) reduces the computational complexity of attention by introducing local windows into self-attention. Schlatt et al. (Schlatt, Fröbe, and Hagen 2024) design a sparse interaction strategy between query and key tokens, improving the efficiency of cross-attention. However, these methods may compromise performance by removing global relationship modeling or not supporting cross-attention. The other category (Gao et al. 2022; Xu et al. 2023) focuses on improving the quality of attention maps. For example, AiA (Gao et al. 2022) refines the original attention mechanism by constructing a second-order relation matrix of the attention map. Xu et al. (Xu et al. 2023) propose a self-calibrated cross-attention to enhance discrimination between foreground and

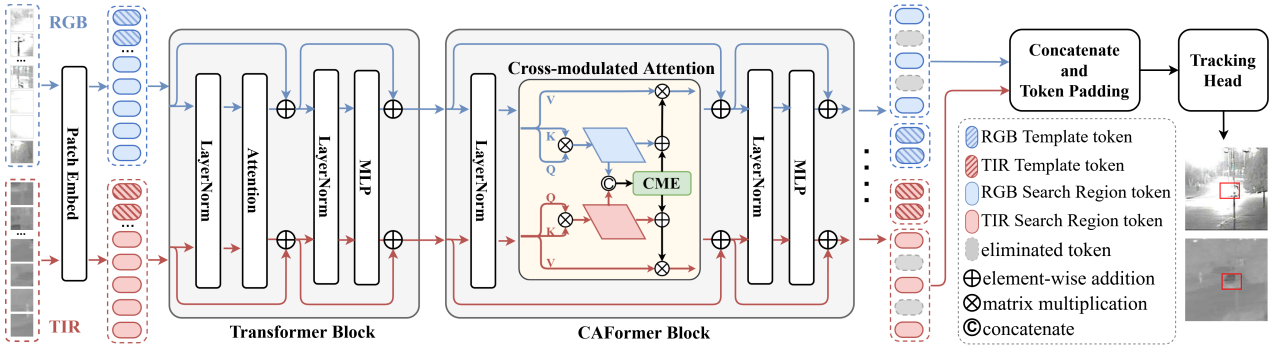


Figure 3: Overall framework of Cross-modulated Attention Transformer (CAFormer) for RGBT tracking. RGB and TIR images are transformed as tokens, and fed into the Encoder that consist of Transformer Block and CAFormer Block. In the proposed Cross-modulated Attention, the correlation maps from both modalities are fed into the CME module for interaction. Finally, we compensate for the eliminated tokens and concatenate search region tokens as the input of tracking head to get the final result.

background images. However, these schemes struggle to accurately model attention weights when encountering low-quality data inputs. In contrast, this paper proposes a multi-modal cross-modulated attention mechanism for the first time, which enhances the attention quality of each modality by establishing a strong association between the attention of RGB and thermal modalities.

RGBT Tracking

Due to the highly complementary nature of RGB and TIR modalities, using TIR as an additional modality can effectively improve the robustness of tracking. Consequently, RGBT tracking has been proposed and has attracted wide attention. With the publication of large-scale RGBT datasets (Li et al. 2021; Pengyu et al. 2022), the Transformer has been widely used in RGBT tracking. For example, Xiao et al. (Xiao et al. 2022) design attribute-specific fusion branches and utilize Transformers to enhance attribute aggregation features and modality-specific features. Hui et al. (Hui et al. 2023) extend ViT (Dosovitskiy et al. 2021) to a multi-modal backbone and propose using fusion templates as a medium for modal interactions to enhance feature fusion with target-related contexts. Luo et al. (Luo et al. 2023) employ three distinct Transformer backbones to extract both modality-specific and modality-shared features. Some works (Zhu et al. 2023; Hong et al. 2024) explore the application of prompt learning to multi-modal tracking. However, the correlation calculation for each modality in these methods is performed independently, making it challenging to avoid inaccurate correlations for low-quality inputs, thus limiting further performance improvement. Moreover, existing fusion modules are typically designed for high-dimensional modal features, requiring significant computational resources, which is not conducive to achieving efficient tracking.

Method

Overview

The proposed approach, named Cross-modulated Attention Transformer (CAFormer), is designed to address the challenges of RGBT tracking by performing intra-modality self-correlation and inter-modality feature interaction in a unified attention model. As illustrated in Figure 3, the framework consists of a backbone network comprising Transformer and CAFormer blocks that process flattened and embedded tokens of RGB and TIR image pairs. The cross-modulated attention mechanism employs correlation maps from both modalities to enhance interaction in the Correlation Modulated Enhancement (CME) module. Furthermore, to filter out non-target tokens, we employ the Collaborative Token Elimination (CTE) strategy in certain layers, which improves the reliability by adding correlation maps. Subsequently, we complete the RGB and TIR tokens belonging to the search region using a token padding scheme, and then concatenate them in the channel and feed them into the tracking head for target state prediction.

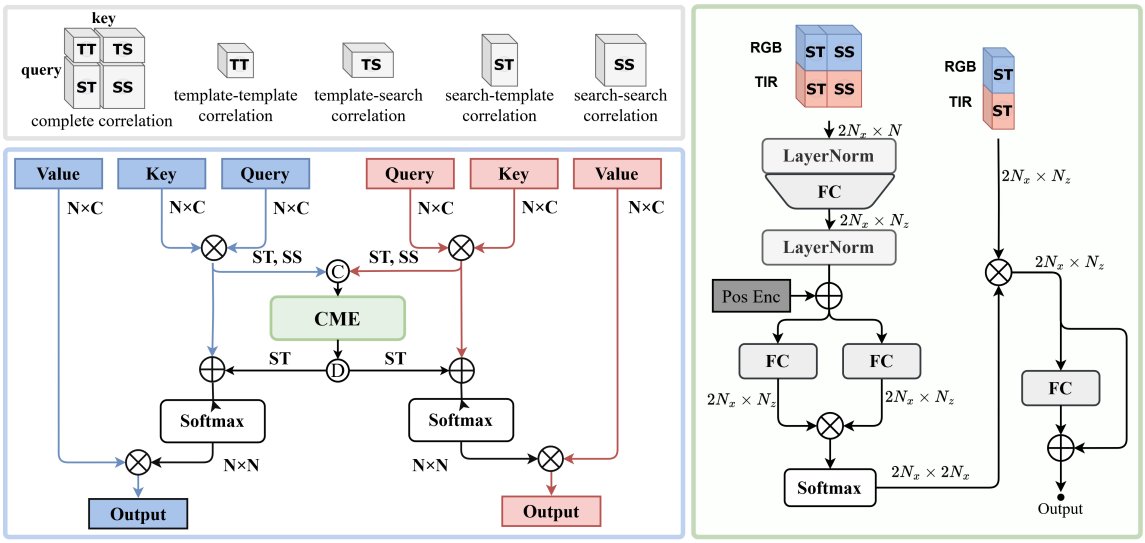
RGBT Baseline Tracker

We adopt a similar approach to recent single object tracking (SOT) methods (Ye et al. 2022; Chen et al. 2022) by concatenating the template frames and search frames together into the Transformer backbone, and then extending it to be the multi-modal backbone of our tracker.

Specifically, given the input RGB and TIR template image pair $\boldsymbol{I}_r^z, \boldsymbol{I}_t^z \in \mathbb{R}^{H_z \times W_z \times 3}$, and search region image pair $\boldsymbol{I}_r^x, \boldsymbol{I}_t^x \in \mathbb{R}^{H_x \times W_x \times 3}$ respectively, we first divide these images into patches of size $P \times P$ and then flatten them to obtain sequences of patches $\boldsymbol{P}_r^z, \boldsymbol{P}_t^z \in \mathbb{R}^{N_z \times (3P^2)}$ and $\boldsymbol{P}_r^x, \boldsymbol{P}_t^x \in \mathbb{R}^{N_x \times (3P^2)}$, where $N_z = H_z W_z / P^2$ and $N_x = H_x W_x / P^2$ denote the number of patches for the template and search frames, respectively. A patch embedding layer, with parameter $\boldsymbol{W}^0 \in \mathbb{R}^{(3P^2) \times C}$ and learnable positional encoding $\boldsymbol{E}_z \in \mathbb{R}^{N_z \times C}$ and $\boldsymbol{E}_x \in \mathbb{R}^{N_x \times C}$, is then applied to obtain template features $\boldsymbol{Z}_r^0$, $\boldsymbol{Z}_t^0$ and search region features $\boldsymbol{X}_r^0$, $\boldsymbol{X}_t^0$ as follows:

$$\begin{aligned} \boldsymbol{Z}_r^0 &= \boldsymbol{P}_r^z \boldsymbol{W}^0 + \boldsymbol{E}_z, \ \boldsymbol{X}_r^0 = \boldsymbol{P}_r^x \boldsymbol{W}^0 + \boldsymbol{E}_x; \\ \boldsymbol{Z}_t^0 &= \boldsymbol{P}_t^z \boldsymbol{W}^0 + \boldsymbol{E}_z, \ \boldsymbol{X}_t^0 = \boldsymbol{P}_t^x \boldsymbol{W}^0 + \boldsymbol{E}_x. \end{aligned} \tag{1}$$

Subsequently, concatenating these features yields token sequences $\boldsymbol{F}_r^0 = [\boldsymbol{Z}_r^0; X_r^0]$, $\boldsymbol{F}_t^0 = [\boldsymbol{Z}_t^0; X_t^0]$, then we feed



(a) Cross-modulated Attention (b) Correlation Modulated Enhancement Module

Figure 4: The proposed Cross-modulated Attention with the Correlation Modulated Enhancement (CME) module. Ⓓ denotes dividing the features of two modalities, ⊗ denotes matrix multiplication, and ⊕ denotes element-wise addition. The numbers beside the arrows are feature dimensions that do not include the batch size. Linear projections in (a) and matrix transpose operations are omitted for brevity. N, N_x, and N_z represent all token numbers, search region token numbers, and template region token numbers, respectively.

them into the l-layer ($l = 1, 2, ..., L$) Transformer block T, whose structure is shown in Figure 3. For simplicity, we use a tracking head consistent with OSTrack (Ye et al. 2022) and denote it as ϕ. The forward propagation process is formulated as follows:

$$\boldsymbol{F}_r^l = T^l(\boldsymbol{F}_r^{l-1}),\ \boldsymbol{F}_t^l = T^l(\boldsymbol{F}_t^{l-1}),\ l = 1, 2, ..., L, \quad (2)$$

where $\boldsymbol{F}_r^L, \boldsymbol{F}_t^L$ are outputs of the last Transformer block. We merge these features along the channel dimension and feed them into the tracking head ϕ to derive the final predicted bounding box $\boldsymbol{B} = \phi(\boldsymbol{F}_r^L, \boldsymbol{F}_t^L)$. At this point, we have a basic multi-modal tracker composed of two branches that share parameters and process different modalities independently.

Cross-modulated Attention

Attention mechanism is a key component of the Transformer tracker (Cui et al. 2022; Song et al. 2023), and the correlation map is an intermediate result of the Transformer attention, which measures the similarity between the tokens (Ye et al. 2022). To avoid the low-quality data affecting the correlation calculation in self-attention, we use high-quality modal correlations to achieve enhancement and correction of low-quality modal correlations. Considering the dynamic changes in the quality of modal correlations, a bidirectional cross-modulated strategy is used to achieve an adaptive correlation modulated process. We design a cross-modulated attention mechanism employs correlation maps from both modalities to enhance interaction in the Correlation Modulated Enhancement (CME) module.

Recalling the backbone in our base tracker, the inputs to layer l are $\boldsymbol{F}_r^l = [\boldsymbol{Z}_r^l; \boldsymbol{X}_r^l]$ and $\boldsymbol{F}_t^l = [\boldsymbol{Z}_t^l; \boldsymbol{X}_t^l]$, here we omit the superscript l and use $\boldsymbol{F}_r$, $\boldsymbol{F}_t$ for simplicity. $\boldsymbol{Q}_r = \boldsymbol{F}_r\boldsymbol{W}_q = [\boldsymbol{Z}_r; \boldsymbol{X}_r]\boldsymbol{W}_q = [\boldsymbol{Q}_r^z; \boldsymbol{Q}_r^x]$, $\boldsymbol{K}_r = \boldsymbol{F}_r\boldsymbol{W}_k = [\boldsymbol{Z}_r; \boldsymbol{X}_r]\boldsymbol{W}_k = [\boldsymbol{K}_r^z; \boldsymbol{K}_r^x]$ denote query and key matrix from RGB modality, and $\boldsymbol{W}_q$, $\boldsymbol{W}_k$ denote the linear projection weights for queries and keys, respectively. For the RGB branch, the correlation map $\boldsymbol{M}_r \in \mathbb{R}^{N\times N}$ is produced as,

$$\begin{aligned}\boldsymbol{M}_r &= \boldsymbol{Q}_r\boldsymbol{K}_r^\top = [\boldsymbol{Q}_r^z; \boldsymbol{Q}_r^x][\boldsymbol{K}_r^z; \boldsymbol{K}_r^x]^\top \\ &= [\boldsymbol{Q}_r^z\boldsymbol{K}_r^{z\top}, \boldsymbol{Q}_r^z\boldsymbol{K}_r^{x\top}; \boldsymbol{Q}_r^x\boldsymbol{K}_r^{z\top}, \boldsymbol{Q}_r^x\boldsymbol{K}_r^{x\top}] \\ &= [\boldsymbol{M}_r^{zz}, \boldsymbol{M}_r^{zx}; \boldsymbol{M}_r^{xz}, \boldsymbol{M}_r^{xx}],\end{aligned} \quad (3)$$

note that $\boldsymbol{M}_r$ needs to undergo softmax and scale to be attention map in the usual meaning. For $\boldsymbol{Q}_t$ and $\boldsymbol{K}_t$ from TIR modality, the processing of RGB features is symmetric to TIR features, we can get correlation map $\boldsymbol{M}_t \in \mathbb{R}^{N\times N}$ in the same way. Known from Eq. 3, $\boldsymbol{M}_r, \boldsymbol{M}_t$ can all be partitioned into four parts $\boldsymbol{M}_r^{zz}, \boldsymbol{M}_r^{zx}, \boldsymbol{M}_r^{xz}, \boldsymbol{M}_r^{xx}$ and $\boldsymbol{M}_t^{zz}, \boldsymbol{M}_t^{zx}, \boldsymbol{M}_t^{xz}, \boldsymbol{M}_t^{xx}$ with different roles in tracking, as proposed by CTTrack (Song et al. 2023). To simplify the description, we rename each part as $\boldsymbol{TT}, \boldsymbol{TS}, \boldsymbol{ST}, \boldsymbol{SS}$ based on the query-key pairs used to calculate the correlation. Among them, $\boldsymbol{ST}$ is a special part, it controls the info stream from template to search frame. Specifically, in most Transformer trackers (Ye et al. 2022; Chen et al. 2022), the tracking head accepts features from the search region, but actually it relies heavily on the template features to output results. Thus, the effect of $\boldsymbol{ST}$ on tracking results is significant. And importantly, due to the spatio-temporally aligned multi-modal image pairs, $\boldsymbol{ST}$ within different branches has remarkable associations.

Existing methods (Luo et al. 2023; Hui et al. 2023) perform separate calculations for correlation in modality, which

ignore the crucial cross-modality associations. To achieve an adaptive correlation modulated process, we design a cross-modulated attention mechanism to employ correlation maps from both modalities to enhance interaction in CME module. The purpose of CME is to modulate $\boldsymbol{ST}$, but we need to take $\boldsymbol{SS}$ into account as well to modulate the final attention map. Specifically, we obtain the aggregated information $\boldsymbol{U}$ for two horizontally adjacent parts as follows:

$$\begin{aligned}\boldsymbol{U} &= LN(LN([\boldsymbol{ST_r}, \boldsymbol{SS_r}; \boldsymbol{ST_t}, \boldsymbol{SS_t}])\boldsymbol{W}_e) \\ &\triangleq [\boldsymbol{U}_r; \boldsymbol{U}_t],\end{aligned} \tag{4}$$

LN denotes the LayerNorm (Ba, Kiros, and Hinton 2016) layer, and $\boldsymbol{W}_e$ is a learnable linear projection weight for embedding two correlation parts. Then we perform an attention operation on $\boldsymbol{U}$ to obtain the modulated correlation map $\boldsymbol{M}^{'}$.

$$\boldsymbol{M}^{'} = Softmax(\frac{(\boldsymbol{U}\boldsymbol{W}_q^{'})(\boldsymbol{U}\boldsymbol{W}_k^{'})^{\top}}{\sqrt{N_z}})[\boldsymbol{ST_r}; \boldsymbol{ST_t}], \tag{5}$$

where N_z is the template tokens number, $\boldsymbol{W}_q^{'}$ and $\boldsymbol{W}_k^{'}$ denote linear projection weights for queries and keys in CME module. Next, we separate $\boldsymbol{ST_r^{'}}$ and $\boldsymbol{ST_t^{'}}$ from the initial modulated correlation map $\boldsymbol{M}_r^{'}$ and $\boldsymbol{M}_t^{'}$, respectively.

$$CME(\boldsymbol{M}_r; \boldsymbol{M}_t) = \boldsymbol{M}^{'}(1 + \boldsymbol{W}^{'}) \quad = [\boldsymbol{ST_r^{'}}; \boldsymbol{ST_t^{'}}], \tag{6}$$

$$\boldsymbol{M}_r^{'} = [0 \cdot \boldsymbol{TT_r}, 0 \cdot \boldsymbol{TS_r}; \boldsymbol{ST_r^{'}}, 0 \cdot \boldsymbol{SS_r}], \tag{7}$$

where $\boldsymbol{W}^{'}$ is a learnable linear projection.

Finally, we add the obtained $\boldsymbol{M}_r^{'}$ to the original correlation map $\boldsymbol{M}_r$ to get the final modulated correlation map. The process of yielding the final RGB attention map $\boldsymbol{A}_r$ can be described as follows:

$$\boldsymbol{A}_r = Softmax(\frac{\boldsymbol{M}_r^{'} + \boldsymbol{M}_r}{\sqrt{C}}), \tag{8}$$

where C denotes the dimension size of the token.

In addition, as illustrated in Figure 4 (a), the proposed cross-modulated attention is a symmetric structure and the parameters are shared that are at the corresponding positions on the left and right sides of the figure. For a multi-head attention block, we share the parameters of the CME module between the parallel multi-heads. It is worth noting that our CME module can be easily applied to other parts in the attention map.

Collaborative Token Elimination

Efficiency is an important metric for evaluating tracking methods (Yan et al. 2021; Cui et al. 2024). An early candidate elimination strategy (Ye et al. 2022) is employed to speed up the inference process in some blocks. This mechanism requires constructing accurate attention weights between the target and each candidate, but it is difficult to achieve from low-quality modalities. To solve the above problem, we propose a Collaborative Token Elimination (CTE) strategy that combines the attention weights from two modalities to make judgments.

Given the query vector $\boldsymbol{q}_r^z$ from $\boldsymbol{Q}_r^z$ and $\boldsymbol{q}_t^z$ from $\boldsymbol{Q}_t^z$, we choose the token in the center of the template as OSTrack (Ye et al. 2022), each search region token at absolute position i can be given a scalar h_i:

$$\boldsymbol{h} = softmax(\boldsymbol{q}_r^z \boldsymbol{K}_r^x) + softmax(\boldsymbol{q}_t^z \boldsymbol{K}_t^x), \tag{9}$$

where $\boldsymbol{K}_r^x$ and $\boldsymbol{K}_t^x$ are the key vectors of search region tokens. After that, we use h_i to sort the search region tokens and keep the top-k tokens. Our method enhances the stability of token elimination, specifically in cases where the quality of one modality declines. It accelerates the network's inference speed while maintaining robustness.

Experiments

Implementation Details

To get a more concrete understanding of the proposed method, here we present details of the implementation. In our method, the proposed CAFormer block is integrated into the last 3 layers of the backbone, and the CTE strategy is adopted at layers 3,6 and 9. The search regions are resized to 256 × 256, while the templates are resized to 128 × 128. For the training process, CAFormer is trained on 2 NVIDIA 2080ti GPUs with a global batch size of 32. We set the learning rates of the backbone network and other parameters to 5e-6 and 5e-5, respectively. The optimization algorithm employed is AdamW (Loshchilov and Hutter 2017) with a weight decay of 1e-4. We train our model for 10 epochs on the training set of LasHeR (Li et al. 2021), and each epoch consists of 60K image pairs. For GTOT (Li et al. 2016), RGBT210 (Li et al. 2017), and RGBT234 (Li et al. 2019a), we directly evaluate our model without any further fine-tuning. For VTUAV (Pengyu et al. 2022) dataset, we adopt the VTUAV training set for our training process, and adjust the number of training epochs to 5. Following previous work (Hui et al. 2023), all experiments in this paper are loaded with pre-trained weights from the public SOT method (Ye et al. 2022). Additionally, we complete the speed test on a device with an Nvidia RTX 3080ti GPU, which is the same as other trackers.

Evaluation on Public Datasets

Our experiments are conducted on five public datasets: GTOT (Li et al. 2016), RGBT210 (Li et al. 2017), RGBT234 (Li et al. 2019a), LasHeR (Li et al. 2021), and VTUAV (Pengyu et al. 2022). For evaluation, we use the commonly adopted Precision Rate (PR) and Success Rate (SR) metrics. Following previous work (Li et al. 2021), we also use the Normalized Precision Rate (NPR) (Muller et al. 2018) metric for LasHeR. Additionally, the GTOT (Li et al. 2016), RGBT234 (Li et al. 2019a), and VTUAV (Pengyu et al. 2022) datasets provide ground truth for both modalities. As in prior works (Li et al. 2019a; Pengyu et al. 2022; Hou et al. 2024), we use the best results from both modalities to address small alignment errors.

GTOT contains 50 video sequence pairs. As shown in Table 1, compared to previous state-of-the-art trackers, our method outperforms HMFT (Pengyu et al. 2022) and the SR score is the best.

Method	backbone	Pub. Info.	GTOT		RGBT210		RGBT234		LasHeR			VTUAV		FPS ↑
			PR ↑	SR ↑	PR ↑	SR ↑	PR ↑	SR ↑	PR ↑	NPR ↑	SR ↑	PR ↑	SR ↑	
DAPNet (Zhu et al. 2019)	VGG-M	ACM MM'19	88.2	70.7	-	-	76.6	53.7	43.1	38.3	31.4	-	-	-
MANet (Li et al. 2019b)	VGG-M	ICCVW'19	89.4	72.4	-	-	77.7	53.9	45.5	-	32.6	-	-	2.1
DAFNet (Gao et al. 2019)	VGG-M	ICCVW'19	89.1	71.6	-	-	79.6	54.4	44.8	39.0	31.1	62.0	45.8	20.5
mfDiMP (Zhang et al. 2019)	ResNet-50	ICCVW'19	83.6	69.7	84.9	59.3	84.2	59.1	59.9	-	46.7	67.3	55.4	34.6
CAT (Li et al. 2020)	VGG-M	ECCV'20	88.9	71.7	79.2	53.3	80.4	56.1	45.0	39.5	31.4	-	-	-
MaCNet (Zhang et al. 2020)	VGG-M	Sensors'20	-	-	-	-	79.0	55.4	48.2	42.0	35.0	-	-	1.6
CMPP (Wang et al. 2020)	VGG-M	CVPR'20	**92.6**	73.8	-	-	82.3	57.5	-	-	-	-	-	-
FANet (Zhu et al. 2020b)	VGG-M	TIV'21	-	-	-	-	78.7	55.3	44.1	38.4	30.9	-	-	-
MANet++ (Lu et al. 2021)	VGG-M	TIP'21	88.2	70.7	-	-	80.0	55.4	46.7	40.4	31.4	-	-	-
SiamCDA (Zhang et al. 2021)	ResNet-50	TCSVT'21	87.7	73.2	-	-	76.0	56.9	-	-	-	-	-	-
DMCNet (Lu et al. 2022)	VGG-M	TNNLS'22	-	-	79.7	55.5	83.9	59.3	49.0	43.1	35.5	-	-	-
APFNet (Xiao et al. 2022)	VGG-M	AAAI'22	90.5	73.7	-	-	82.7	57.9	50.0	43.9	36.2		-	1.9
MIRNet (Hou, Ren, and Wu 2022)	VGG-M	ICME'22	90.9	74.4	-	-	81.6	58.9	-	-	-	-	-	-
TFNet (Zhu et al. 2021)	VGG-M	TCSVT'22	88.6	72.9	77.7	52.9	80.6	56.0	-	-		-	- -	
HMFT (Pengyu et al. 2022)	ResNet-50	CVPR'22	91.2	74.9	-	-	78.8	56.8	-	-	-	75.8	62.7	30.2
ViPT (Zhu et al. 2023)	ViT-B	CVPR'23	-	-	-	-	83.5	61.7	65.1	-	52.5	-	-	-
MACFT (Luo et al. 2023)	ViT-B	Sensors'23	90.0	72.7	-	-	85.7	62.2	65.3	-	51.4	80.1	66.8	33.3
TBSI (Hui et al. 2023)	ViT-B	CVPR'23	-	-	85.3	62.5	87.1	63.7	69.2	65.7	55.6	-	-	36.2
CMD (Zhang et al. 2023)	ResNet-50	CVPR'23	90.7	73.5	-	-	84.4	60.1	59.7	55.4	46.7	-	-	18
CAT++ (Liu et al. 2024)	VGG-M	TIP'24	91.5	73.3	82.2	56.1	84.0	59.2	50.9	44.4	35.6	-	-	14
OneTracker (Hong et al. 2024)	ViT-B	CVPR'24	-	-	-	-	85.7	64.2	67.2	-	53.8	-	-	-
Un-Track (Wu et al. 2024)	ViT-B	CVPR'24	-	-	-	-	84.2	62.5	66.7	-	53.6	-	-	-
SDSTrack (Hou et al. 2024)	ViT-B	CVPR'24	-	-	-	-	84.8	62.5	66.5	-	53.1	-	-	-
TATrack (Wang et al. 2024)	ViT-B	AAAI'24	-	-	85.3	61.8	87.2	64.4	**70.2**	-	56.1	-	-	26.1
BAT (Bing et al. 2024)	ViT-B	AAAI'24	90.7	76.5	84.9	61.2	86.8	64.1	**70.2**	-	**56.3**	-	-	69.2
CAFormer	ViT-B	-	91.8	**76.9**	**85.6**	**63.2**	**88.3**	**66.4**	70.0	**66.1**	55.6	**88.6**	**76.2**	**83.6**

Table 1: Comparison with state-of-the-art methods. The best results are highlighted with bold font, respectively.

RGBT210 is a challenging RGBT dataset, which contains 210 video sequence pairs, 210K frames, and 12 tracking challenge attributes. In the evaluation of the RGBT210 dataset, our method gets the best PR/SR score with 85.6%/63.2%. Compared with TBSI (Hui et al. 2023), there is a minor improvement of 0.3%/0.7% on PR/SR, but in terms of efficiency, the proposed method is twice as efficient as TBSI. In addition, our method has a significant advantage over other methods, outperforming CAT (Li et al. 2020), TFNet (Zhu et al. 2021) 6.4%/9.9% and 7.9%/10.3% in terms of PR/SR, respectively.

RGBT234 is extended from RGBT210, which contains 12 challenge attributes, and 234 video sequence pairs. As shown in Table 1, we compare our method with 22 advanced RGBT trackers and achieve the best result. TBSI is the state-of-the-art method and it uses feature fusion. Our method outperforms TBSI by a significant margin of 1.2%/2.7% on PR/SR and obtains the best performance. For other trackers, our method outperformed mfDiMP (Zhang et al. 2019) and ViPT (Zhu et al. 2023) in PR and SR scores by 4.1%/7.3% and 4.8%/4.7%, respectively.

LasHeR contains 19 challenge attributes, 734.8K frames, and 1224 video sequence pairs. We compare with recently proposed RGBT trackers and achieve the best result. Specifically, our tracker significantly outperforms the mfDiMP and ViPT, i.e. 10.1%/8.9% and 4.9%/3.1% respectively in PR/SR. Although compared with TBSI, our method only has the performance advantage of 0.8%/0.4% in PR/NPR metrics, TBSI obviously lags behind our method in tracking efficiency because of its bulky multi-level feature interaction.

VTUAV stands out as a large-scale RGBT dataset specifically designed for UAV perspectives. VTUAV contains 500 video sequence pairs having 1.7M image pairs with 1920 × 1080 resolution. It can be seen that our method outperforms all previous methods. Specifically, compared to MACFT (Luo et al. 2023), which is the previous state-of-the-art method, our method leads by 8.5%/9.4% in PR/SR. This indicates that the proposed method is equally applicable to UAV scenarios and its efficiency is suitable for the needs of UAV scenarios.

Method	RGBT234		LasHeR	
	PR	SR	PR	SR
RGBT baseline	86.4	64.5	67.8	54.0
w/o $\boldsymbol{SS}$	87.6	65.6	69.2	55.1
w/o Cross-modal	87.5	65.9	68.3	54.3
Full model (CAFormer)	**88.3**	**66.4**	**70.0**	**55.6**

Table 2: Evaluation results for different structures.

Ablation Study

Component Analysis As shown in Table 2, we compare different designs for the proposed CMA module.

w/o $\boldsymbol{SS}$**.** Since the softmax operation can span parts of two horizontally neighboring correlation maps, they affect each other. When processing $\boldsymbol{ST}$, we also take $\boldsymbol{SS}$ into account. Removing $\boldsymbol{SS}$ from the model results in a 0.7%/0.8% decrease in PR/SR scores on RGBT234 and a 0.8%/0.5% decrease on LasHeR, compared to using both $\boldsymbol{ST}$ and $\boldsymbol{SS}$. The results, shown in Table 2, demonstrate that $\boldsymbol{SS}$ plays an important role in adjusting the weights of $\boldsymbol{ST}$.

w/o Cross-modal. The proposed CME module is aim to exploit the association of correlations between modalities. When we remove this mechanism, it means that the correlation weights of the two modalities only perform self-interaction. As shown in the Table 2, there is a significant decrease of 1.7% and 1.3% in PR and SR compared to the full model on LasHeR, respectively. It suggests that the main

part	RGBT234		LasHeR	
	PR	SR	PR	SR
$\boldsymbol{ST}$	**88.3**	**66.4**	**70.0**	**55.6**
$\boldsymbol{SS}$	87.8	65.8	69.1	55.0
$\boldsymbol{TS}$	87.7	65.7	68.2	54.4
($\boldsymbol{ST}$,$\boldsymbol{SS}$)	86.4	64.4	68.9	54.7

Table 3: Modulating different parts of the correlation map. "($\boldsymbol{ST}$, $\boldsymbol{SS}$)" implies that the two parts are treated as a whole in the interaction.

Layers	RGBT234		LasHeR	
	PR	SR	PR	SR
last 1 layer	87.5	65.6	69.2	55.1
last 3 layers	88.3	**66.4**	**70.0**	**55.6**
last 6 layers	86.6	65.0	68.1	54.2
4,7,10 layers	**88.5**	65.8	69.6	55.1

Table 4: Apply layers of the proposed CAFormer block.

Method		LasHeR		FPS	MACs (G)
CE	CTE	PR	SR		
		69.3	55.1	76.4	58.43
✓		69.5	55.2	83.6	42.91
	✓	**70.0**	**55.6**	83.6	42.91

Table 5: Different candidate elimination strategies.

performance increase of our method comes from the cross-modal interaction of correlation weights.

CMA Utilization in Different Parts Besides interacting with $\boldsymbol{ST}$, we attempt to deploy the CME module in other parts of the correlation map. The results on RGBT234 and LasHeR are summarized in Table 3. We can observe that the best result is obtained when applied to the $\boldsymbol{ST}$. It confirms that the part $\boldsymbol{ST}$ has a stronger cross-modal correlation and is critical for tracking. When we do not distinguish different parts, i.e. "($\boldsymbol{ST}$, $\boldsymbol{SS}$)", it leads to 1.9%/2.0% decrease on RGBT234 and 1.1%/0.9% decrease on LasHeR in PR/SR scores. This shows that it is necessary to distinguish different parts of the correlation map.

CMA Insertion in Different Layers We insert the CAFormer block to different layers and summarize the experimental results on RGBT234 and LasHeR in Table 4. When we employ CAFormer block at last layer, we observe a notable enhancement. It shows the necessity of correlation fusion. As we increasing to last 3 layer, the boosting effect is weakened, and when continuing to increase to 6, worse results are obtained. This suggests that less a priori information can lead to difficulties in distinguishing potentially correct correlation weights, thus yielding erroneous interactions and resulting in performance degradation. Finally, we choose the last 3 layers as the final solution.

Effectiveness of Different Token Elimination Schemes To verify the effectiveness of the proposed collaborative token elimination (CTE) strategy, we compare it with the candidate elimination (CE) strategy in OSTrack (Ye et al. 2022). As shown in Table 5, CTE not only helps to improve the inference speed, but also significantly enhances

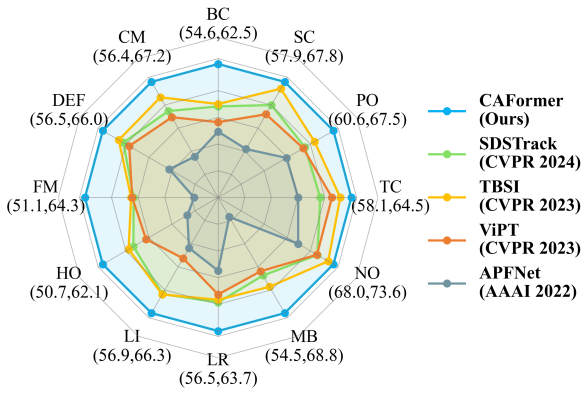


Figure 5: Attribute-based evaluation on RGBT234 dataset. In parentheses, the value on the left indicates the minimum success rate, and on the right the maximum success rate.

performance, whereas the CE strategy primarily improves efficiency. Specifically, adding CTE or CE policy improves the tracking speed by 9.4%, and decreases the MACs by 26.6%. In terms of tracking performance, CTE obtains a 0.7%/0.5% improvement in PR/SR. It is significantly larger than that of CE, which is 0.2%/0.1%. The results show that the CTE module is more compatible with the CMA module as it ensures interaction between weights at corresponding positions across modalities.

Attribute-based Performance

We evaluate the performance of our proposed method in various scenarios by conducting experiments on different challenge attribute subsets of the RGBT234 dataset, including no occlusion (NO), partial occlusion (PO), heavy occlusion (HO), low illumination (LI), low resolution (LR), thermal crossover (TC), deformation (DEF), fast motion (FM), scale change (SC), motion blur (MB), camera moving (CM) and background clutter (BC). All results are summarized in Figure 5. We can observe the proposed method outperforms existing methods (Xiao et al. 2022; Hui et al. 2023; Zhu et al. 2023) under all challenges. In particular, CAFormer makes significant improvements on LI, FM, MB and TC challenges. This demonstrates the advantages of the proposed correlation fusion scheme.

Conclusion

In this work, we propose a novel Cross-modulated Attention Transformer (CMA) for RGBT tracking, which is the first reveal the consistency of visible and infrared modalities in search frame and template frame correlations in RGBT tracking. We also present a novel correlation fusion insight for multi-modal tracking, which provides clear advantages in both performance and efficiency over existing mainstream feature fusion. Additionally, a Collaborative Token Elimination strategy is proposed that enhances the distinction between foreground and background and further improves efficiency and performance. In the future, we plan to combine correlation fusion and feature fusion to further improve tracking performance.

Acknowledgments

This work was supported the Anhui Provincial Natural Science Foundation (No. 2408085MF153), the National Natural Science Foundation of China (No. 62406002) and the China Postdoctoral Science Foundation (2024M760011).

References

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bing, C.; Junliang, G.; Pengfei, Z.; and Qinghua, H. 2024. Bi-directional Adapter for Multimodal Tracking. In *AAAI Conference on Artificial Intelligence*.

Chen, B.; Li, P.; Bai, L.; Qiao, L.; Shen, Q.; Li, B.; Gan, W.; Wu, W.; and Ouyang, W. 2022. Backbone is all your need: A simplified architecture for visual object tracking. In *Proceedings of the European Conference on Computer Vision*, 375–392.

Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; and Lu, H. 2021. Transformer tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8126–8135.

Cui, Y.; Jiang, C.; Wang, L.; and Wu, G. 2022. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13608–13618.

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

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Fu, Z.; Fu, Z.; Liu, Q.; Cai, W.; and Wang, Y. 2022. SparseTT: Visual tracking with sparse Transformers. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*.

Gao, S.; Zhou, C.; Ma, C.; Wang, X.; and Yuan, J. 2022. AiATrack: Attention in attention for Transformer visual tracking. In *Proceedings of the European Conference on Computer Vision*.

Gao, Y.; Li, C.; Zhu, Y.; Tang, J.; He, T.; and Wang, F. 2019. Deep adaptive fusion network for high performance RGBT tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*.

Guo, M.-H.; Xu, T.-X.; Liu, J.-J.; Liu, Z.-N.; Jiang, P.-T.; Mu, T.-J.; Zhang, S.-H.; Martin, R. R.; Cheng, M.-M.; and Hu, S.-M. 2022. Attention mechanisms in computer vision: A survey. *Computational visual media*, 8(3): 331–368.

Hong, L.; Yan, S.; Zhang, R.; Li, W.; Zhou, X.; Guo, P.; Jiang, K.; Chen, Y.; Li, J.; Chen, Z.; et al. 2024. OneTracker: Unifying Visual Object Tracking with Foundation Models and Efficient Tuning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

Hou, R.; Ren, T.; and Wu, G. 2022. Mirnet: A robust RGBT tracking jointly with multi-modal interaction and refinement. In *IEEE International Conference on Multimedia and Expo*, 1–6.

Hou, X.; Xing, J.; Qian, Y.; Guo, Y.; Xin, S.; Chen, J.; Tang, K.; Wang, M.; Jiang, Z.; Liu, L.; et al. 2024. SDSTrack: Self-Distillation Symmetric Adapter Learning for Multi-Modal Visual Object Tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Hui, T.; Xun, Z.; Peng, F.; Huang, J.; Wei, X.; Wei, X.; Dai, J.; Han, J.; and Liu, S. 2023. Bridging search region interaction with template for RGB-T tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13630–13639.

Li, C.; Cheng, H.; Hu, S.; Liu, X.; Tang, J.; and Lin, L. 2016. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing*, 25(12): 5743–5756.

Li, C.; Liang, X.; Lu, Y.; Zhao, N.; and Tang, J. 2019a. RGB-T object tracking: Benchmark and baseline. *Pattern Recognition*, 96: 106977.

Li, C.; Liu, L.; Lu, A.; Ji, Q.; and Tang, J. 2020. Challenge-aware RGBT tracking. In *Proceedings of the European Conference on Computer Vision*, 222–237.

Li, C.; Lu, A.; Zheng, A.; Tu, Z.; and Tang, J. 2019b. Multi-adapter RGBT tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshop*, 2262–2270.

Li, C.; Xue, W.; Jia, Y.; Qu, Z.; Luo, B.; Tang, J.; and Sun, D. 2021. LasHeR: A large-scale high-diversity benchmark for RGBT tracking. *IEEE Transactions on Image Processing*, 31: 392–404.

Li, C.; Zhao, N.; Lu, Y.; Zhu, C.; and Tang, J. 2017. Weighted sparse representation regularized graph learning for RGB-T object tracking. *Proceedings of the 25th ACM international conference on Multimedia*.

Liu, L.; Li, C.; Xiao, Y.; Ruan, R.; and Fan, M. 2024. RGBT Tracking via Challenge-Based Appearance Disentanglement and Interaction. *IEEE Transactions on Image Processing*, 33: 1753–1767.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, A.; Li, C.; Yan, Y.; Tang, J.; and Luo, B. 2021. RGBT tracking via multi-Adapter network with hierarchical divergence Loss. *IEEE Transactions on Image Processing*, 30: 5613–5625.

Lu, A.; Qian, C.; Li, C.; Tang, J.; and Wang, L. 2022. Duality-gated mutual condition network for RGBT tracking.

IEEE Transactions on Neural Networks and Learning Systems, 1–14.

Luo, Y.; Guo, X.; Dong, M.; and Yu, J. 2023. Learning modality complementary features with mixed attention mechanism for RGB-T tracking. *Sensors*, 23(14): 6609.

Muller, M.; Bibi, A.; Giancola, S.; Alsubaihi, S.; and Ghanem, B. 2018. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European Conference on Computer Vision*, 300–317.

Pengyu, Z.; Zhao, J.; Wang, D.; Lu, H.; and Ruan, X. 2022. Visible-thermal UAV tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*.

Schlatt, F.; Fröbe, M.; and Hagen, M. 2024. Investigating the Effects of Sparse Attention on Cross-Encoders. In *European Conference on Information Retrieval*, 173–190.

Song, Z.; Luo, R.; Yu, J.; Chen, Y.-P. P.; and Yang, W. 2023. Compact Transformer tracker with correlative masked modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*, 5998–6008.

Wang, C.; Xu, C.; Cui, Z.; Zhou, L.; Zhang, T.; Zhang, X.; and Yang, J. 2020. Cross-modal pattern-propagation for RGB-T tracking. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 7064–7073.

Wang, H.; Liu, X.; Li, Y.; Sun, M.; Yuan, D.; and Liu, J. 2024. Temporal Adaptive RGBT Tracking with Modality Prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

Wu, Z.; Zheng, J.; Ren, X.; Vasluianu, F.-A.; Ma, C.; Paudel, D. P.; Van Gool, L.; and Timofte, R. 2024. Single-Model and Any-Modality for Video Object Tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

Xiao, Y.; Yang, M.; Li, C.; Liu, L.; and Tang, J. 2022. Attribute-based progressive fusion network for RGBT tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Xu, Q.; Zhao, W.; Lin, G.; and Long, C. 2023. Self-calibrated cross attention network for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 655–665.

Yan, B.; Peng, H.; Wu, K.; Wang, D.; Fu, J.; and Lu, H. 2021. LightTrack: Finding lightweight neural networks for object tracking via one-shot architecture search. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Ye, B.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2022. Joint feature learning and relation modeling for tracking: A one-stream framework. In *Proceedings of the European Conference on Computer Vision*, 341–357.

Zhang, H.; Zhang, L.; Zhuo, L.; and Zhang, J. 2020. Object tracking in RGB-T videos using modal-aware attention network and competitive learning. *Sensors*, 20(2): 393.

Zhang, L.; Danelljan, M.; Gonzalez-Garcia, A.; van de Weijer, J.; and Shahbaz Khan, F. 2019. Multi-modal fusion for end-to-end RGB-T tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*.

Zhang, T.; Guo, H.; Jiao, Q.; Zhang, Q.; and Han, J. 2023. Efficient RGB-T Tracking via Cross-Modality Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5404–5413.

Zhang, T.; Liu, X.; Zhang, Q.; and Han, J. 2021. SiamCDA: Complementarity-and distractor-aware RGB-T tracking based on Siamese network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3): 1403–1417.

Zhou, Q.; Shi, H.; Xiang, W.; Kang, B.; and Latecki, L. J. 2024. DPNet: Dual-path network for real-time object detection with lightweight attention. *IEEE Transactions on Neural Networks and Learning Systems*.

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

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020a. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.

Zhu, Y.; Li, C.; Luo, B.; Tang, J.; and Wang, X. 2019. Dense feature aggregation and pruning for RGBT tracking. In *Proceedings of the ACM International Conference on Multimedia*, 465–472.

Zhu, Y.; Li, C.; Tang, J.; and Luo, B. 2020b. Quality-aware feature aggregation network for robust RGBT tracking. *IEEE Transactions on Intelligent Vehicles*, 6(1): 121–130.

Zhu, Y.; Li, C.; Tang, J.; Luo, B.; and Wang, L. 2021. RGBT tracking by trident fusion network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2): 579–592.