Template-based Uncertainty Multimodal Fusion Network for RGBT Tracking

Zhaodong Ding[1,2,4] , **Chenglong Li**[1,2,4*] , **Shengqing Miao**[6] and **Jin Tang**[1,3,5]
[1]National Key Laboratory of Opto-Electronic Information Acquisition and Protection Technology, Hefei, 230601, China
[2]Anhui Provincial Key Laboratory of Security Artificial Intelligence, Hefei, 230601, China
[3]Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Hefei, 230601, China
[4]School of Artificial Intelligence, Anhui University, Hefei, 230601, China
[5]School of Computer Science and Technology, Anhui University, Hefei, 230601, China
[6]School of Electronic and Information Engineering, Anhui University, Hefei, 230601, China
zhaodongding_ah@163.com, lcl1314@foxmail.com, 1325957414@qq.com, tangjin@ahu.edu.cn

Abstract

RGBT tracking is to localize the predefined targets in video sequences by effectively leveraging the information from both visible light (RGB) and thermal infrared (TIR) modalities. However, the quality of different modalities changes dynamically in complex scenes, and effectively perceiving modal quality for multimodal fusion remains a significant challenge. To address this challenge, we propose to employ the reliability of initial template to explore the uncertainty across different modalities, and design a novel template-based uncertainty computation framework for robust multimodal fusion in RGBT tracking. In particular, we introduce an Uncertainty-aware Multimodal Fusion Module (UMFM), which constructs the uncertainty of each modality by leveraging the correlation between the template and search region in the Subjective Logic framework, aiming to achieve robust multimodal fusion. In addition, existing methods focus on dynamic template update while overlooking the potential role of a reliable initial template in the template updating process. To this end, we design a simple yet effective Contrastive Template Update Module (CTUM) to assess the reliability of the new template by comparing its quality with that of the initial template. Extensive experiments suggest that our method outperforms existing approaches on four RGBT tracking benchmarks.

1 Introduction

In recent years, RGBT tracking has attracted increasing attention from researchers due to its broad and practical application prospects in fields such as assisted driving, human-computer interaction, and security surveillance [Zhang *et al.*, 2024]. The goal of RGBT tracking is to achieve robust all-weather and around-the-clock tracking by fully leveraging the complementary advantages of RGB and TIR modalities. However, the quality of the two modalities fluctuates dynamically in complex scenes. Thus, it is crucial for the designed model to be aware of this for adaption.

Current research on RGBT tracking primarily focuses on the fusion of modality information. These studies can be roughly divided into three categories. The first type of approach [Liu *et al.*, 2023; Chen *et al.*, 2024; Hui *et al.*, 2023a] typically focuses on designing suitable methods to achieve adaptive modal fusion. The second category is to realize the full utilization of different modal information by designing different types of prompts or adapters based on the frozen RGB tracker [Cao *et al.*, 2024; Hou *et al.*, 2024; Zhu *et al.*, 2023]. For instance, BAT [Cao *et al.*, 2024] designs a simple bidirectional adapter, effectively achieving comprehensive fusion of modality features and enhancing the performance of the tracker. The third type of approach [Li *et al.*, 2019b; Li *et al.*, 2024a; Liu *et al.*, 2024] focuses on mining the discriminative features of each modality through feature interaction or attribute decoupling, thereby improving the performance of RGBT trackers. However, existing methods typically generate the dynamic weights of the two modalities, with poor interpretability and difficulty in measuring the quality of the final multimodal outcome. In contrast, uncertainty-based multimodal fusion can effectively reflect the quality of each individual modality as well as the quality of the fused multimodal result to achieve robust multimodal fusion.

Based on the above discussion, we propose an Template-based Uncertainty Multimodal Fusion Network (TUMFNet) for RGBT tracking, which dynamically evaluates the quality of different modalities by modeling their uncertainty using the reliability of the initial template and Subjective Logic (SL) [Jsang, 2018]. Although the SL theory has made significant progress in uncertainty modeling, current works focus on multimodal classification tasks and is not suitable for directly applying SL to RGBT tracking. Therefore, we propose to employ the reliability of the initial template to explore the uncertainty of different modalities. To this end, we innovatively propose an Uncertainty-Aware Multimodal Fusion Module (UMFM) that leverages the correlation between the template and search region as well as SL to model modal uncertainty. Specifically, UMFM first fuses the tem-

[*]Corresponding author

plate features from both modalities and then uses the integrated template tokens and the search region tokens from each modality to perform correlation matrix computation separately for each modality, resulting in two distinct correlation matrices. Subsequently, UMFM utilizes these correlation matrices to construct the subjective opinions for each modality. To fully leverage the information from both modalities, we introduce an evidential multimodal fusion method to combine the subjective opinions of the two modalities for reliable token-level weight. Meanwhile, we construct modality-level weights based on the uncertainty of the two modalities and the uncertainty of the multimodal fusion. By leveraging both modality-level and token-level weights, we effectively suppress the impact of low-quality modality noise while enhancing the target features.

Additionally, the template plays a crucial role in tracking tasks. Existing tracking methods rely on the initial template and search regions to track the target, making it difficult to capture changes in the target's appearance. To solve this problem, some methods [He *et al.*, 2023; Wang *et al.*, 2024] update the template after a certain interval but fail to ensure the reliability of the template. Other methods [Zhang *et al.*, 2019; Chen *et al.*, 2022] design different modules to assess the reliability of the template but overlook the potential role of the initial template in the template update process. The initial template is inherently reliable, making it reasonable to rely on it to assess the quality of the new template. Therefore, to mine spatio-temporal contextual information from the target and fully leverage the reliable initial template, we design a simple yet effective contrastive template update module. Specifically, we first use the predicted maximum classification score to select candidate templates for template update. Then, we design a simple template quality perceptor that determines whether to update the template by comparing the quality of the initial and candidate templates.

In summary, our major contributions are as follows.

- We propose a novel template-based uncertainty computation framework, which models the uncertainty of different modalities by leveraging the correlation between the template and search region in the Subjective Logic framework.
- We design a novel uncertainty-aware multimodal fusion module, which evaluates the quality of different modalities by constructing modality-specific uncertainty using the correlation between the template and search region, thereby suppressing the interference from low-quality modalities.
- We develop a simple yet effective contrastive template update module that utilizes the initial template to assess the quality of templates and determine whether the dynamic template should be updated.
- Extensive experiments demonstrate that our method outperforms existing RGBT tracking methods on four popular RGBT tracking datasets, including GTOT [Li *et al.*, 2016], RGBT210 [Li *et al.*, 2017], RGBT234 [Li *et al.*, 2019a] and LasHeR [Li *et al.*, 2021].

2 Related Work

2.1 RGBT Tracking

RGBT tracking leverages the complementary strengths of both modalities to achieve robust visual tracking, effectively overcoming the limitations of the data from each modality. Existing RGBT tracking methods can be broadly classified into two categories. The first type of approach [Liu *et al.*, 2023; Chen *et al.*, 2024; Hui *et al.*, 2023a; Zhu *et al.*, 2020] achieves effective fusion of features from both modalities by designing specific fusion strategies. TGTrack [Chen *et al.*, 2024] designs a top-down cross-modal guidance attention mechanism to enable interaction between modality information. The second category of methods [Cao *et al.*, 2024; Hou *et al.*, 2024; Zhu *et al.*, 2023; Hong *et al.*, 2024] typically focuses on transferring the capabilities of an RGB tracker to a new RGBT tracker through prompt learning or adapters. BAT [Cao *et al.*, 2024] and SDSTrack [Hou *et al.*, 2024] introduce a simple adapter for multimodal information interaction. The third type of approach [Li *et al.*, 2024a; Liu *et al.*, 2024; Zhang *et al.*, 2021] focuses on enhancing representation learning across different modalities. However, these methods struggle to handle the dynamic changes in modality quality in complex scenarios. Therefore, we propose UMFM, which addresses the variation in modal quality by leveraging the correlation between the reliable template and search region to model the uncertainty of the modalities.

In addition, existing tracking methods typically rely on the initial template to track the target in the search region, making it difficult to adapt to changes in the target's appearance. To address this issue, some methods [He *et al.*, 2023; Wang *et al.*, 2024] update the template at specific intervals, but they struggle to ensure the reliability of the template. Other methods [Zhang *et al.*, 2019; Chen *et al.*, 2022] achieve template update by designing different modules to assess the reliability of the new template. For example, [Chen *et al.*, 2022] propose a regression model to control the template updating process. Although these methods improve tracking performance, they overlook the potential role of the reliable initial template in the template update process. To mine spatio-temporal contextual information of the target and fully leverage the reliable initial template, we propose a simple yet effective contrastive template update module, which determines whether to update the template by comparing the quality of the initial and new templates.

2.2 Uncertainty-aware Multimodal Learning

Uncertainty-aware multimodal learning has developed rapidly in recent years, researchers from different fields have explored the application of the Subjective Logic (SL) theory in their respective tasks fields [Li *et al.*, 2024b; Xu *et al.*, 2024; Kotelevskii *et al.*, 2024]. For instance, [Han *et al.*, 2022] pioneer the application of the SL framework to tackle the issue of low-quality data in multi-view classification tasks. [Xu *et al.*, 2024] further introduce reliable conflictive multiview learning to resolve potential conflicts in multiview data. In other fields, [Li *et al.*, 2024b] propose an adaptive uncertainty learning framework using SL for cross-modal person re-identification. These methods demonstrate the significant advantages of the SL theory in exploring uncertainty. However, directly transferring it to RGBT tracking is not suitable. To bridge this gap, we innovatively propose an uncertainty-aware multimodal fusion module that leverages the correlation between the template and search regions as well as SL to model modal uncertainty.

3 Method

3.1 Framework Overview

The overall structure of the proposed method is illustrated in Figure 1. This framework mainly consists of two core components: the Uncertainty-aware Multimodal Fusion Module (UMFM) and the Contrastive Template Update Module (CTUM). To dynamically perceive the quality of different modalities, UMFM models uncertainty to assess the quality of each modality by the correlation between the template and search region in the Subjective Logic framework. It is worth noting that UMFM is inserted into the 10th, 11th, and 12th encoders of the ViT. In addition, to mine spatio-temporal contextual information from the target during the tracking process, CTUM performs dynamic template update by comparing the quality of the candidate template with the initial template.

3.2 Uncertainty-aware Multimodal Fusion Module

Although most existing RGBT tracking methods [Hui *et al.*, 2023a; Cao *et al.*, 2024] have improved tracking performance by designing various modal fusion strategies, they struggle to achieve robust tracking when faced with dynamic changes in modality quality. To address this issue, we propose an Uncertainty-aware Multimodal Fusion Module that models the uncertainty of modalities to perceive changes in modality quality, thereby suppressing low-quality modality information and fully achieve multimodal fusion. It should be noted that we model the uncertainty of different modalities by employing the reliablility of initial template and Subjective Logic [Jsang, 2018].

Subjective Logic Framework. Subjective Logic (SL) provides a formal representation of the uncertainty allocation principle from Dempster-Shafer theory [Liu and Yager, 2008], which is modeled as a Dirichlet distribution. Thus, it offers a method to quantify uncertainty within a well-established theoretical framework using the principles of subjective logic. Specifically, we first need to obtain the evidence vector e_i for the *i*-th element. Then, we model the uncertainty u and the belief mass **p** = $\{p_k\}_{k=1}^{N}$ of each element, with the specific process outlined as follows:

$$p_k = \frac{e_k}{S}, \quad u = \frac{N}{S}, \tag{1}$$

where $S = \sum_{k=1}^{N}(e_k + 1)$ is the the intensity of Dirichlet distribution, p_k denotes the belief probability and the parameters of the corresponding Dirichlet distribution are $\boldsymbol{\alpha} = \{e_k + 1\}_{k=1}^{N}$. We can observe that uncertainty is inversely proportional to the total evidence, meaning that as the total evidence increases, uncertainty decreases. Additionally, the Dirichlet distribution represented by $\boldsymbol{\alpha}$ can be defined as follows:

$$D(\mathbf{p} \mid \boldsymbol{\alpha}) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha})} \prod_{j=1}^{N} p_j^{\alpha_j - 1} & \text{for } \mathbf{p} \in \mathcal{S}_N, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $B(\boldsymbol{\alpha})$ represents the N-dimensional beta function, and $\mathcal{S}_N$ is the N-dimensional unit simplex.

Template-based Uncertainty Calculation. To extend the SL to the RGBT tracking task, the initial step is to construct the evidence vector $\boldsymbol{e}_i$. In particular, we first concatenate the two template features T_r, T_t from both modalities and pass them through a 1x1 convolutional layer (*Conv*) for feature fusion to obtain the multimodal template features. Then, we calculate the correlation between the search region features S_r, S_t of each modality and the multimodal template features T_f, obtaining two correlation matrices, $C_r \in \mathbb{R}^{N_t \times N_s}$ and $C_t \in \mathbb{R}^{N_t \times N_s}$. N_t and N_s represent the number of tokens in the template and search region, respectively. Considering that the correlation matrix reflects the correlation between search region tokens and multimodal template tokens, we first compute the mean of the two correlation matrices along the column dimension and process them using the ReLU activation function. In this way, we obtain the overall correlation between the search region tokens and the reliable multimodal template features, which is the evidence $\mathbf{e}_i = \{e_i\}_{i=1}^{N_s}$ we need, as shown Figure 1. The specific processing procedure is as follows:

$$\begin{aligned} T_f &= \mathit{Conv}(\mathit{concat}(T_r, T_t)), \\ C_r &= \frac{QK^T}{\sqrt{d}} = \frac{S_r W_q (T_f W_k)^T}{\sqrt{d}}, \\ C_t &= \frac{QK^T}{\sqrt{d}} = \frac{S_t W_q (T_f W_k)^T}{\sqrt{d}}, \end{aligned} \tag{3}$$

$$\mathbf{e}_i^r = \mathrm{ReLu}\left(\mathrm{Mean}\left(C_r\right)\right), \mathbf{e}_i^t = \mathrm{ReLu}\left(\mathrm{Mean}\left(C_t\right)\right), \tag{4}$$

where $\mathbf{e}_i^r$ and $\mathbf{e}_i^t$ represent the evidence vectors of the RGB modality and the TIR modality, respectively.

Following the subjective logic framework mentioned above, we can obtain the parameters $\boldsymbol{\alpha}_i$ of the Dirichlet distribution as well as model the uncertainty **u** of each modality. Thus, we obtain the subjective opinions for each modality. The specific process is as follows:

$$\boldsymbol{\alpha}_i = \mathbf{e}_i + 1, \quad \mathbf{u} = \frac{N_s}{S}, \tag{5}$$

where $S = \sum_{i=1}^{N_s}(\boldsymbol{\alpha}_i)$ is the intensity of Dirichlet distribution.

Uncertainty-based Multimodal Fusion. To suppress the noise from low-quality modalities as well as to achieve effective multimodal fusion, we introduce a simple evidential multimodal fusion method. The process is as follows:

$$b_i^m = \frac{b_i^r u^t + b_i^t u^r}{u^r + u^t}, u^m = \frac{2u^r u^t}{u^r + u^t}, a_i^m = \frac{a_i^r + a_i^t}{2} \tag{6}$$

where b_i^r, a_i^r, u^r represent the belief mass, Dirichlet distribution parameters, and uncertainty corresponding to the RGB modality, respectively. Similarly, b_i^m, a_i^m, u^m represent the

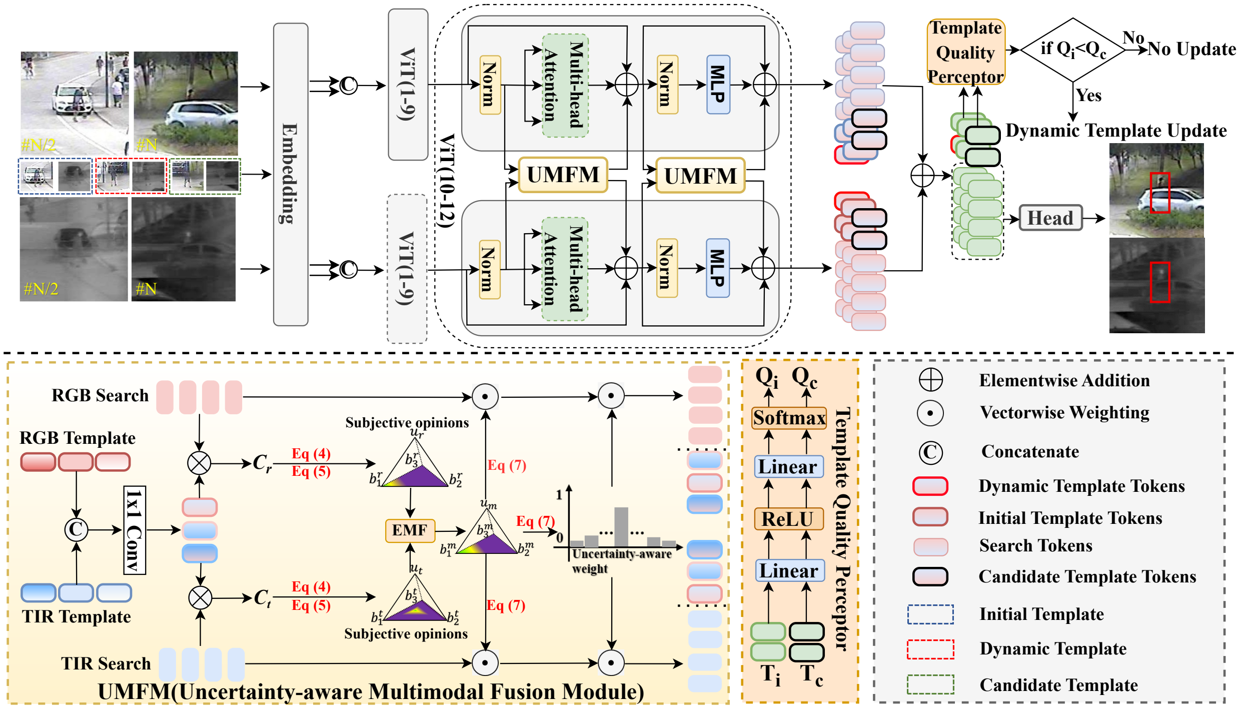

Figure 1: The overall architecture of our proposed method. EMF denotes evidential multimodal fusion. Every N frames, the model updates the dynamic template based on the quality score of templates. At this point, the input of the model includes the search regions of the N-th frame, the search regions of the $\frac{N}{2}$-th frame, the initial templates, the dynamic templates, and the candidate templates. If the quality score Q_c of the candidate template is higher than that Q_i of the initial template, the candidate template is updated as the dynamic template.

belief mass, Dirichlet distribution parameters, and uncertainty corresponding to the multimodal fusion. In addition, we obtain modality-level weights w_r, w_t based on uncertainty to suppress noise from low-quality modalities. Meanwhile, to further enhance the target features, we also introduce a token-level weight w_m. The detailed computation process is as follows:

$$w_r = \frac{u^m}{u^r}, w_t = \frac{u^m}{u^t}, w_m = \frac{a_i^m}{S^m} \tag{7}$$

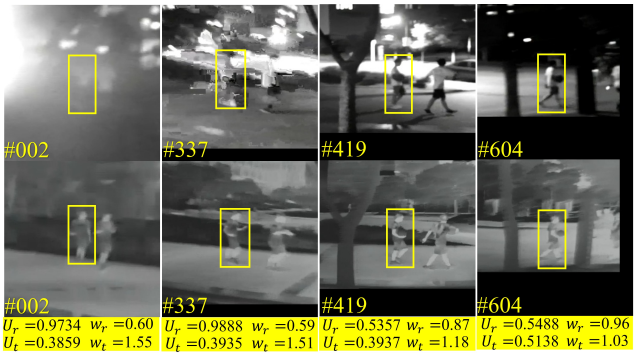
#002
#337
#419
#604
#002
#337
#419
#604
U_r =0.9734 w_r =0.60
U_t =0.3859 w_t =1.55
U_r =0.9888 w_r =0.59
U_t =0.3935 w_t =1.51
U_r =0.5357 w_r =0.87
U_t =0.3937 w_t =1.18
U_r =0.5488 w_r =0.96
U_t =0.5138 w_t =1.03

Figure 2: The uncertainty scores and dynamic weights of four frames in the sequence *boytakingbasketballfollowing* from the LasHeR dataset. U_r and U_t represent the uncertainty scores of the RGB and TIR modalities, respectively, while w_r and w_t represent the dynamic weights of the two modalities.

where $S^m = \sum_{i=1}^{N_s}(\boldsymbol{\alpha_i^m})$ is the intensity of Dirichlet distribution. Finally, to achieve effective uncertainty-aware learning, we propose an novel loss function $\mathcal{L}_u$, which supervises the learning process, as shown below:

$$\alpha_1 = (\boldsymbol{\alpha^m} - 1) \cdot y, \alpha_2 = (\boldsymbol{\alpha^m} - 1) \cdot (1 - y)$$

$$\mathcal{L}_U = -\log \frac{\sum_{i=1}^{N_s} e^{-\alpha_2}}{\sum_{i=1}^{N_s} e^{-\alpha_1} + \sum_{i=1}^{N_s} e^{-\alpha_2}} \tag{8}$$

where $\boldsymbol{\alpha^m}$ represents the Dirichlet distribution parameter, y denotes the mask of the target token, α_1 refers to the relevant parameters for the target, and α_2 refers to the relevant parameters for the background. We present the uncertainty scores for a sequence, as shown in Figure 2. It can be observed that under strong light or jitter, the quality of the RGB modality deteriorates, resulting in higher uncertainty scores, while the uncertainty scores of the TIR modality remain relatively stable.

3.3 Contrastive Template Update Module

In existing tracking methods [Zhang *et al.*, 2019; Wang *et al.*, 2024], template update is often used to capture the appearance changes of the target, thereby achieving robust tracking. However, current methods typically rely on classification scores and *IoU* scores to determine whether to update the template, overlooking the potential role of the initial template in the template update process. To explore the spatio-temporal context information and fully leverage the reliable

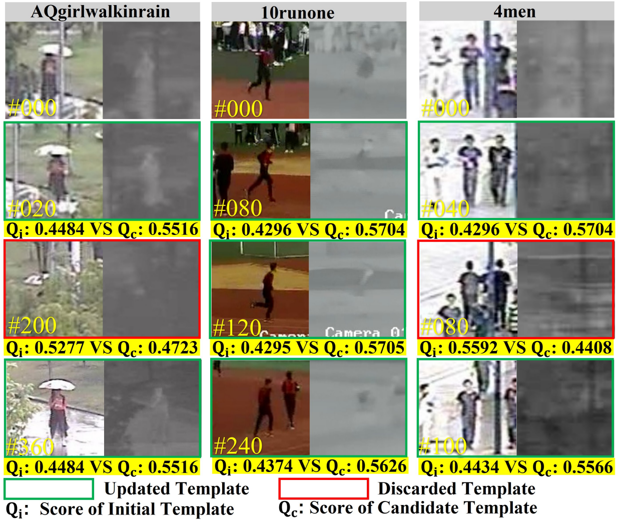
AQgirlwalkinrain
10runone
4men
#000
#000
#000
#020
#080
#040
Q_i: 0.4484 VS Q_c: 0.5516
Q_i: 0.4296 VS Q_c: 0.5704
Q_i: 0.4296 VS Q_c: 0.5704
#200
#120
#080
Q_i: 0.5277 VS Q_c: 0.4723
Q_i: 0.4295 VS Q_c: 0.5705
Q_i: 0.5592 VS Q_c: 0.4408
#360
#240
#100
Q_i: 0.4484 VS Q_c: 0.5516
Q_i: 0.4374 VS Q_c: 0.5626
Q_i: 0.4434 VS Q_c: 0.5566
Updated Template
Discarded Template
Q_i: Score of Initial Template
Q_c: Score of Candidate Template

Figure 3: The illustration of quality scores for the initial template and candidate templates in three sequences from the LasHeR dataset.

initial template, we design a simple yet effective CTUM including a template quality perceptor and a dual-template contrastive training strategy.

Specifically, the template quality perceptor consists of two linear layers, one ReLU layer and one Softmax layer, as shown in Figure 1. To effectively train the perceptor for perceiving the quality of different templates, we design the dual-template contrastive training strategy. During the training phase, for each sampling from the sequence, we select two distinct templates and a search region. Each of the different templates from the same sequence will interact with the same search region to extract features. The features of the two different templates are then input into the perceptor for quality perception, resulting in the quality scores Q_k, Q_j for both templates. In addition, to facilitate the learning of the perceptor, we generate pseudo labels based on the *IoU* derived from the prediction results. For two identical search regions S_k, S_j, we perform interactions with different templates. If the *IoU* obtained by a search region feature after being input into the predictor is higher, we infer that the template interacting with this search region exhibits superior quality. Therefore, the label generation process is described as follows:

$$y_k = \begin{cases} 0, & IoU_k < IoU_j \\ 1, & IoU_k > IoU_j \end{cases}, y_j = \begin{cases} 1, & IoU_k < IoU_j \\ 0, & IoU_k > IoU_j \end{cases}, \quad (9)$$

where IoU_k and IoU_j the *IoU* of the prediction results for the k-th and j-th search regions, respectively. In this way, we obtain the labels indicating the relative quality of the two templates.

In the tracking phase, our template update strategy is as follows:

- **Candidate Template Selection.** The classification score output by the tracking model reflects the discriminability between the foreground and background, and many methods [Wang *et al.*, 2024] use it as a criterion to assess the reliability of the prediction results. Therefore, we use the maximum classification score to obtain the candidate template. Specifically, if the maximum classification score of the current frame exceeds the average of the past classification scores, we consider the current prediction to be relatively reliable and update the candidate template accordingly.
- **Quality-contrast Template Update.** Although the maximum classification score enables continuous updating of the candidate template, its actual reliability needs further evaluation. The template quality perceptor assesses the quality of both the initial and candidate templates to determine whether the dynamic template should be genuinely updated. The specific template update process is shown in Figure 1. In the tracking process, we performs candidate template quality assessment every N frames. Specifically, for the N-th frame, both the initial template and the candidate template interact with the same two $\frac{N}{2}$-th frame. The features T_i, T_c of both templates are then input into the perceptor, which predicts the initial template quality score Q_i and candidate template quality score Q_c. If $Q_c > Q_i$, the dynamic template is updated. We present the template update process for three sequences, as shown in Figure 3. It can be observed that when the target in the template is occluded or tracking fails, the perceptor correctly determines whether the current candidate template should be discarded.

3.4 Loss Function

Our method adopts the traditional center point prediction method, where the regression and classification branches predict the target bounding box, and the model learns using both classification and regression losses. Additionally, we introduce a novel loss function $\mathcal{L}_U$ for the learning of UMFM. To effectively train the template quality perceptor, we also introduce a binary cross-entropy loss $\mathcal{L}_{BCE}$. The overall loss function is as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{\text{giou}} + \lambda_t \mathcal{L}_U + \lambda_3 \mathcal{L}_{BCE}, \quad (10)$$

where $\mathcal{L}_{\text{cls}}$ denotes the weighted focal loss, $\mathcal{L}_1$ represents the L1 loss, and the generalized *IoU* loss is represented by $\mathcal{L}_{\text{giou}}$. The parameters λ_1 and λ_2 are kept unchanged according to the settings in previous works [Ye *et al.*, 2022]. The parameter $\lambda_t = \min(0.1, t/T) \in (0, 0.1]$ is the annealing coefficient, t is the index of the current training epoch, and T is the annealing step. In our experiments, λ_1 is set to 5, λ_2 to 2 and λ_3 is set to 0.01.

4 Experiment

4.1 Implementation Details

We take OSTrack [Ye *et al.*, 2022] as the base tracker, which employs ViT as the backbone network for feature extraction. In addition, we use the parameters from DropMae[Wu *et al.*, 2023] as the pre-trained weights. Our model is implemented using PyTorch and experiments are conducted on one RTX 4090 GPU. We train the overall tracking network end-to-end using the LasHeR training set to evaluate GTOT, RGBT210, RGBT234, and LasHeR test set. The input search region and

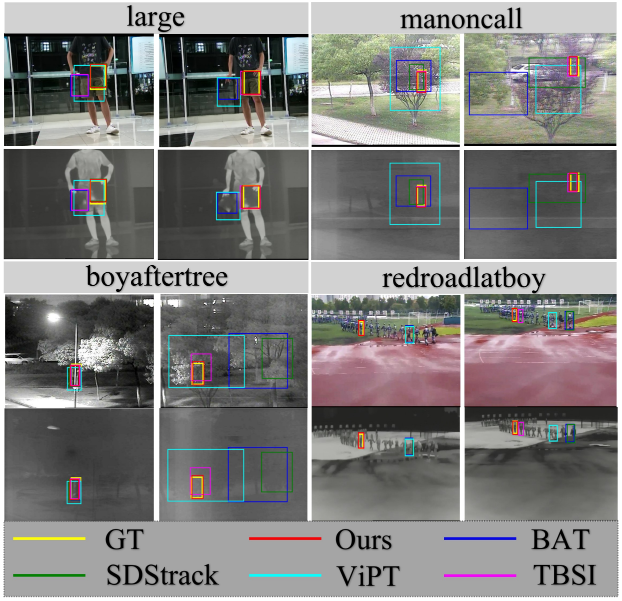

Figure 4: Visualization of our method with other RGBT trackers on four representative sequences that include multiple challenge attributes from LasHeR dataset.

template sizes of the model are 256×256 and 128×128, respectively. The learning rate for the backbone network is set to 1×10^{-5}, while the learning rate for the other parameters is set to 1×10^{-4}. The model is trained for a total of 15 epochs. Additionally, we use the AdamW optimizer with a weight decay of 1×10^{-4}. Note that our UMFM is inserted into the 10th, 11th, and 12th blocks of the ViT. λ_3 is set to 0.01 and N is set to 20.

Inference: During the inference phase, at every interval of N frames, the input of the model consists of the initial template, dynamic template, and the search region. When performing quality assessment of the candidate template every N frames, the input of model includes the initial template, dynamic template, candidate template, the search region of the N-th frame, and the search region of the $\frac{N}{2}$-th frame.

4.2 Evaluation Datasets and Metrics

GTOT [Li *et al.*, 2016] dataset is the earliest publicly available RGBT tracking dataset, containing a total of 50 sequences and approximately 15,000 frames. **RGBT210** [Li *et al.*, 2017] dataset consists of 210 pairs of RGBT video sequences, totaling 209.4K frames, and includes annotations across 12 attributes. **RGBT234** [Li *et al.*, 2019a] dataset extends RGBT210 dataset, offering more precise annotations. It contains a total of 234 pairs of RGBT video sequences, amounting to approximately 233.4K frames. **LasHeR** [Li *et al.*, 2021] is a large RGBT tracking dataset, consisting of 1,224 aligned video sequences with approximately 1,469.6K frames. It includes 245 test sequences and 979 training sequences, covering 19 real-world challenge attributes.

Following the previous works [Liu *et al.*, 2023; Hui *et al.*, 2023b], we use Precision (PR) and Success Rate (SR) as the main evaluation metrics in one-pass evaluation for quantitative performance analysis, which are commonly employed in current RGBT tracking tasks. To consider target size, we normalize PR to obtain Normalized Precision (NPR) for evaluating tracking performance on the LasHeR dataset.

4.3 Comparisons With State-of-the-Art Methods

We evaluate our algorithm on four popular RGBT tracking datasets and compare its performance with current state-of-the-art methods.

Evaluation on GTOT dataset. The comparison of the experimental results in GTOT dataset is shown in Table 1, where our method outperforms the current state-of-the-art methods, with gains over HMFT, QAT and US-Track in PR/SR by 4.3%/5.3%, 4.0%/4.7%, and 2.1%/1.9%, respectively. We further compare our method with the CKD tracker, which eliminates modality gaps. In terms of PR and SR, our method outperforms CKD by 2.3% and 3.0%, respectively.

Evaluation on RGBT210 dataset. We compare our method with 7 existing RGBT trackers on RGBT210 dataset, and the results are shown in Table 1. Specifically, compared to TBSI, QAT, and TATrack, our method achieves significant improvements of 5.4%/3.3%, 3.9%/3.9%, and 5.4%/4.0% in PR/SR, respectively. We further compare our method with the top-performing approach, CKD, and observe a performance improvement of 2.3%/0.6% in PR/SR.

Evaluation on RGBT234 dataset. To further validate the effectiveness of our method, we compare it with other 17 state-of-the-art RGBT trackers on RGBT234 dataset. When comparing our method with QAT, BAT, and US-Track on RGBT234 dataset, it is observed that our method achieves improvements of 2.4%/3.4%, 4.0%/3.7%, and 3.7%/4.1% in PR/SR, respectively. Finally, when comparing with the best-performing method on RGBT234 dataset, CKD, our method outperforms it by 0.8% in PR and 0.4% in SR.

Evaluation on LasHeR dataset. We compare our method with 13 state-of-the-art RGBT trackers on the most challenging LasHeR dataset, and the evaluation results are shown in Table 1. Our method significantly outperforms existing advanced methods due to its effective perception of modality quality and its ability to fully utilize dynamic templates to extract spatio-temporal information, achieving state-of-the-art performance with PR, NPR, and SR scores of 76.4%, 72.7%, and 61.4%, respectively. Compared to TATrack and BAT, our method exceeds them by 5.2% and 5.3% in both PR and SR metrics. We also compare our method with the top-performing tracker on this dataset, CKD. Our method outperforms CKD by 3.2%/3.4%/3.3% in terms of PR/NPR/SR.

Additionally, to further validate the effectiveness of the proposed method, we visualize the tracking results of four sequences with multiple challenges, as shown in Figure 4. We can observe that in the sequence *manoncall*, when the target is heavily occluded, BAT, ViPT, and SDSTrack fail to track the target, while only our method and TBSI manage to locate the target accurately. When faced with the sequence *boyaftertree*, includes challenges such as occlusion and low-light conditions, all other methods fail to track the target, while only our method achieves stable tracking.

Methods	Pub. Info.	Backbone	GTOT		RGBT210		RGBT234		LasHeR			FPS
			PR↑	SR↑	PR↑	SR↑	PR↑	SR↑	PR↑	NPR↑	SR↑	↑
APFNet [Xiao *et al.*, 2022]	AAAI 2022	VGG−M	90.5	73.7	–	–	82.7	57.9	50.0	43.9	36.2	1.3
DMCNet [Lu *et al.*, 2022]	TNNLS 2022	VGG−M	90.9	73.3	79.7	55.5	83.9	59.3	49.0	43.1	35.5	2.3
ProTrack [Yang *et al.*, 2022]	ACM MM 2022	ViT−B	–	–	–	–	78.6	58.7	50.9	–	42.1	30
HMFT [Pengyu *et al.*, 2022]	CVPR 2022	ResNet−50	91.2	74.9	78.6	53.5	78.8	56.8	–	–	–	–
MFG [Wang *et al.*, 2022]	TMM 2022	ResNet−18	88.9	70.7	74.9	46.7	75.8	51.5	–	–	–	–
CMD [Zhang *et al.*, 2023]	CVPR 2023	ResNet−50	89.2	73.4	–	–	82.4	58.4	59.0	54.6	46.4	30
ViPT [Zhu *et al.*, 2023]	CVPR 2023	ViT−B	–	–	–	–	83.5	61.7	65.1	–	52.5	–
TBSI [Hui *et al.*, 2023b]	CVPR 2023	ViT−B	–	–	85.3	62.5	87.1	63.7	69.2	65.7	55.6	36.2
QAT [Liu *et al.*, 2023]	ACM MM 2023	ResNet−50	91.5	75.5	86.8	61.9	88.4	64.4	64.2	59.6	50.1	22
TATrack [Wang *et al.*, 2024]	AAAI 2024	ViT−B	–	–	85.3	61.8	87.2	64.4	70.2	66.7	56.1	26.1
BAT [Cao *et al.*, 2024]	AAAI 2024	ViT−B	–	–	–	–	86.8	64.1	70.2	–	56.3	–
Un-Track [Wu *et al.*, 2024]	CVPR 2024	ViT−B	–	–	–	–	84.2	62.5	66.7	–	53.6	–
SDSTrack [Hou *et al.*, 2024]	CVPR 2024	ViT−B	–	–	–	–	84.8	62.5	66.5	–	53.1	20.9
OneTracker [Hong *et al.*, 2024]	CVPR 2024	ViT−B	–	–	–	–	85.7	64.2	67.2	–	53.8	–
US-Track [Xia *et al.*, 2024]	IJCAI 2024	ViT−B	93.4	78.3	–	–	87.1	63.7	–	–	–	84.2
CKD [Lu *et al.*, 2024]	ACM MM 2024	ViT−B	93.2	77.2	88.4	65.2	90.0	67.4	73.2	69.3	58.1	96.4
Ours	–	ViT−B	95.5	80.2	90.7	65.8	90.8	67.8	76.4	72.7	61.4	41

Table 1: PR/NPR and SR scores (%) for advanced trackers on four datasets. The best and second results are in *red* and *blue* colors, respectively.

Component		RGBT234		LasHeR		
UMFM	CTUM	MPR	MSR	PR	NPR	SR
		88.5	66.2	70.6	67.0	56.7
✓		90.0	67.1	73.1	69.4	58.9
✓	✓	**90.8**	**67.8**	**76.4**	**72.7**	**61.4**

Table 2: Component analysis on RGBT234 and LasHeR dataset.

Inserting Layers			RGBT234		LasHeR		
10	11	12	MPR	MSR	PR	NPR	SR
			88.5	66.2	70.6	67.0	56.7
✓			89.4	66.9	74.8	70.9	60.0
✓	✓		90.3	67.7	75.7	72.0	60.6
✓	✓	✓	**90.8**	**67.8**	**76.4**	**72.7**	**61.4**

Table 3: Inserting layers of the proposed UMFM.

4.4 Ablation Study

Analysis of Different Components. We first analyze the effectiveness of the designed UMFM and CTUM, with the results presented in Table 2. It can be seen that when UMFM is incorporated into the baseline, the model performance shows significant improvement. With the addition of CTUM, the performance is further enhanced due to the utilization of spatio-temporal information.

Analysis of Hyperparameters. We first evaluate different inserting layers of our proposed UMFM and summarize the experimental results in Table 3. As UMFM is added layer by layer, it can be observed that the overall performance of the model gradually improves. We then explore the impact of the interval N on the performance of the proposed method, and the experimental results are shown in Table 4. It can be observed that when the interval N is set to 20, the performance of our method achieves the best result. Compared to the other two template selection methods, although they also achieve significant performance gains, our method demonstrates a more prominent improvement in performance. **For more experimental results and visualizations, please refer to the supplementary materials**[*]**.**

N	MAX/RADNOM/CTUM		
	PR	NPR	SR
10	74.1/74.0/76.0	70.4/70.1/72.3	59.5/59.3/ 60.1
20	74.7/73.6/**76.4**	71.2/69.8/**72.7**	60.1/59.1/ **61.4**
30	74.6/73.8/75.6	71.4/70.0/71.9	59.9/59.1/60.5
40	74.1/73.9/74.6	70.7/70.0/70.9	59.7/59.3/60.0

Table 4: Analysis of hyperparameters N on LasHeR dataset. MAX denotes selecting the result with the highest classification score within the interval N as the updated template. RANDOM indicates that a result is randomly selected within the interval N for template updating.

5 Conclusion

In this work, we propose a novel template-based uncertainty framework consisting of an uncertainty-aware multimodal fusion module and a contrastive template update module, for RGBT tracking. UMFM utilizes the reliablility of initial template and Subjective Logic to model the uncertainty of different modalities, addressing the issue of dynamic modality quality changes in complex scenarios. CTUM explores the role of the initial template in the template update process, facilitating efficient dynamic template update. Extensive experiments on four publicly available RGBT tracking datasets demonstrate the effectiveness of our method compared to other state-of-the-art RGBT trackers. In the future, we will continue to explore novel and effective dynamic template update strategies and integrate the uncertainty-aware multimodal fusion module into other tracking frameworks.

[*] https://github.com/dongdong2061/IJCAI25-TUMFNet

Acknowledgments

This work was supported by the Natural Science Foundation of China (No.62376004) and the Natural Science Foundation of Anhui Province (No.2208085J18).

References

[Cao *et al.*, 2024] Bing Cao, Junliang Guo, Pengfei Zhu, and Qinghua Hu. Bi-directional adapter for multi-modal tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 927–935, 2024.

[Chen *et al.*, 2022] Xin Chen, Bin Yan, Jiawen Zhu, Huchuan Lu, Xiang Ruan, and Dong Wang. High-performance transformer tracking. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 45, pages 8507–8523, 2022.

[Chen *et al.*, 2024] Liang Chen, Bineng Zhong, Qihua Liang, Yaozong Zheng, Zhiyi Mo, and Shuxiang Song. Top-down cross-modal guidance for robust rgb-t tracking. In *IEEE Transactions on Circuits and Systems for Video Technology*. IEEE, 2024.

[Han *et al.*, 2022] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. In *IEEE transactions on pattern analysis and machine intelligence*, volume 45, pages 2551–2566. IEEE, 2022.

[He *et al.*, 2023] Kaijie He, Canlong Zhang, Sheng Xie, Zhixin Li, and Zhiwen Wang. Target-aware tracking with long-term context attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 1, pages 773–780, 2023.

[Hong *et al.*, 2024] Lingyi Hong, Shilin Yan, Renrui Zhang, Wanyun Li, Xinyu Zhou, Pinxue Guo, Kaixun Jiang, Yiting Chen, Jinglun Li, Zhaoyu 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*, pages 19079–19091, 2024.

[Hou *et al.*, 2024] Xiaojun Hou, Jiazheng Xing, Yijie Qian, Yaowei Guo, Shuo Xin, Junhao Chen, Kai Tang, Mengmeng Wang, Zhengkai Jiang, Liang 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*, pages 26551–26561, 2024.

[Hui *et al.*, 2023a] Tianrui Hui, Zizheng Xun, Fengguang Peng, Junshi Huang, Xiaoming Wei, Xiaolin Wei, Jiao Dai, Jizhong Han, and Si Liu. Bridging search region interaction with template for rgb-t tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13630–13639, 2023.

[Hui *et al.*, 2023b] Tianrui Hui, Zizheng Xun, Fengguang Peng, Junshi Huang, Xiaoming Wei, Xiaolin Wei, Jiao Dai, Jizhong Han, and Si Liu. Bridging search region interaction with template for rgb-t tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13630–13639, 2023.

[Jsang, 2018] Audun Jsang. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated, 2018.

[Kotelevskii *et al.*, 2024] Nikita Kotelevskii, Samuel Horváth, Karthik Nandakumar, Martin Takac, and Maxim Panov. Dirichlet-based uncertainty quantification for personalized federated learning with improved posterior networks. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 7127–7135. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track.

[Li *et al.*, 2016] Chenglong Li, Hui Cheng, Shiyi Hu, Xiaobai Liu, Jin Tang, and Liang Lin. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Trans. Image Process.*, 25(12):5743–5756, 2016.

[Li *et al.*, 2017] Chenglong Li, Nan Zhao, Yijuan Lu, Chengli Zhu, and Jin Tang. Weighted sparse representation regularized graph learning for rgb-t object tracking. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1856–1864, 2017.

[Li *et al.*, 2019a] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: Benchmark and baseline. In *Pattern Recognition*, volume 96, page 106977. Elsevier, 2019.

[Li *et al.*, 2019b] Chenglong Li, Andong Lu, Aihua Zheng, Zhengzheng Tu, and Jin Tang. Multi-adapter rgbt tracking. In *Proceedings of IEEE International Conference on Computer Vision Workshops*, 2019.

[Li *et al.*, 2021] Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, Jin Tang, and Dengdi Sun. Lasher: A large-scale high-diversity benchmark for rgbt tracking. In *IEEE Transactions on Image Processing*, pages 392–404, 2021.

[Li *et al.*, 2024a] Chenglong Li, Tao Wang, Zhaodong Ding, Yun Xiao, and Jin Tang. Dynamic disentangled fusion network for rgbt tracking. In *arXiv preprint arXiv:2412.08441*, 2024.

[Li *et al.*, 2024b] Shenshen Li, Chen He, Xing Xu, Fumin Shen, Yang Yang, and Heng Tao Shen. Adaptive uncertainty-based learning for text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3172–3180, 2024.

[Liu and Yager, 2008] Liping Liu and Ronald R Yager. Classic works of the dempster-shafer theory of belief functions: An introduction. In *Classic works of the Dempster-Shafer theory of belief functions*, pages 1–34. Springer, 2008.

[Liu *et al.*, 2023] Lei Liu, Chenglong Li, Yun Xiao, and Jin Tang. Quality-aware rgbt tracking via supervised reliability learning and weighted residual guidance. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3129–3137, 2023.

[Liu *et al.*, 2024] Lei Liu, Chenglong Li, Yun Xiao, Rui Ruan, and Minghao Fan. Rgbt tracking via challenge-based appearance disentanglement and interaction. In *IEEE Transactions on Image Processing*, 2024.

[Lu *et al.*, 2022] Andong Lu, Cun Qian, Chenglong Li, Jin Tang, and Liang Wang. Duality-gated mutual condition network for rgbt tracking. In *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2022.

[Lu *et al.*, 2024] Andong Lu, Jiacong Zhao, Chenglong Li, Yun Xiao, and Bin Luo. Breaking modality gap in rgbt tracking: Coupled knowledge distillation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9291–9300, 2024.

[Pengyu *et al.*, 2022] Zhang Pengyu, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8886–8895, 2022.

[Wang *et al.*, 2022] Xiao Wang, Xiujun Shu, Shilliang Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Mfgnet: Dynamic modality-aware filter generation for rgb-t tracking. In *IEEE Transactions on Multimedia*, pages 4335–4348, 2022.

[Wang *et al.*, 2024] Hongyu Wang, Xiaotao Liu, Yifan Li, Meng Sun, Dian Yuan, and Jing Liu. Temporal adaptive rgbt tracking with modality prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5436–5444, 2024.

[Wu *et al.*, 2023] Qiangqiang Wu, Tianyu Yang, Ziquan Liu, Baoyuan Wu, Ying Shan, and Antoni B Chan. Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14561–14571, 2023.

[Wu *et al.*, 2024] Zongwei Wu, Jilai Zheng, Xiangxuan Ren, Florin-Alexandru Vasluianu, Chao Ma, Danda Pani Paudel, Luc Van Gool, and Radu Timofte. Single-model and any-modality for video object tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 19156–19166, 2024.

[Xia *et al.*, 2024] Jianqiang Xia, Dianxi Shi, Ke Song, Linna Song, Xiaolei Wang, Songchang Jin, Chenran Zhao, Yu Cheng, Lei Jin, Zheng Zhu, Jianan Li, Gang Wang, Junliang Xing, and Jian Zhao. Unified single-stage transformer network for efficient rgb-t tracking. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 1471–1479. International Joint Conferences on Artificial Intelligence Organization, 2024.

[Xiao *et al.*, 2022] Yun Xiao, Mengmeng Yang, Chenglong Li, Lei Liu, and Jin Tang. Attribute-based progressive fusion network for rgbt tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2831–2838, 2022.

[Xu *et al.*, 2024] Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, and Xiyue Gao. Reliable conflictive multi-view learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16129–16137, 2024.

[Yang *et al.*, 2022] Jinyu Yang, Zhe Li, Feng Zheng, Ales Leonardis, and Jingkuan Song. Prompting for multi-modal tracking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3492–3500, 2022.

[Ye *et al.*, 2022] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European Conference on Computer Vision*, pages 341–357. Springer, 2022.

[Zhang *et al.*, 2019] Lichao Zhang, Abel Gonzalez-Garcia, Joost Van De Weijer, Martin Danelljan, and Fahad Shahbaz Khan. Learning the model update for siamese trackers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4010–4019, 2019.

[Zhang *et al.*, 2021] Pengyu Zhang, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Learning adaptive attribute-driven representation for real-time rgb-t tracking. In *International Journal of Computer Vision*, pages 2714–2729, 2021.

[Zhang *et al.*, 2023] Tianlu Zhang, Hongyuan Guo, Qiang Jiao, Qiang Zhang, and Jungong Han. Efficient rgb-t tracking via cross-modality distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5404–5413, 2023.

[Zhang *et al.*, 2024] Haiping Zhang, Di Yuan, Xiu Shu, Zhihui Li, Qiao Liu, Xiaojun Chang, Zhenyu He, and Guangming Shi. A comprehensive review of rgbt tracking. In *IEEE Transactions on Instrumentation and Measurement*. IEEE, 2024.

[Zhu *et al.*, 2020] Yabin Zhu, Chenglong Li, Jin Tang, and Bin Luo. Quality-aware feature aggregation network for robust rgbt tracking. In *IEEE Transactions on Intelligent Vehicles*, pages 121–130, 2020.

[Zhu *et al.*, 2023] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9516–9526, 2023.