SiamDMA: Siamese Dual-level And Multi-domain Fusion Attention Network

for RGBT Tracking

He Fengchen

Abstract: The trackers based on Siamese network shows good performance in dealing with RGB-T challenges, but in various scenarios, it can not reasonably allocate the visible and infrared modes' weight ratio.

We propose a dual-level balance module (DLBM) by introducing a new dual-level fusion attention mechanism to utilize the two modal information at decision-level and feature-level. The decision-level information includes the target location, the distinguishability between the target and the background. We combine the information of decision-level and feature-level, which provides a more reasonable way to balance the two modal features' weight ratio.

On the other hand, the decision-level information is limited by the feature extraction ability of the network for each mode. We propose a multi-domain aware module (MDAM) by introducing a new cross domain siamese attention mechanism. Multi-domain includes mode-domain (referring to visible and infrared modal branches) and time-domain (referring to template and search image branches), Which is full of rich context informations. Thus, the interactive multi-domain information can enhance feature representations ability of network and adaptively update template features.

Based on SiamBAN, combined with dual-level fusion attention mechanism and cross domain siamese attention mechanism, we propose a siamese dual-level dnd multi-domain fusion attention network (SiamDMA). Experiments on three RGB-T tracking benchmark datasets demonstrate that SiamDMA has achieved state-of-the-art performance. The average tracking speed on rtx3060ti is about 45fps.

Index Terms- RGBT tracking, dual-level balance, multi-domain aware, mode-domain, time-domain, siamese network

1. INTRODUCE

Target tracking is an important task in the field of computer vision, which gives an initial target template and estimates its position and size in subsequent frames. With the emergence of correlation filtering and deep learning, visible target tracking has achieved considerable

development. However, when the visible mode image is not enough to display the target, such as dark light, high exposure or submerged in the background, the visible tracking effect will be greatly reduced.

Most of the time, the infrared image is rich in the structure information of the target, and the visible image is rich in the structure and color information of the target. The features of the two mode are highly complementary, introduction of infrared mode will improve the performance of the tracker predictably [1]. Therefore, the two modal fusion tracking algorithm has emerged year by year. We show the complementarity of infrared and visible images in Figure 1.

Fig.1. Complementary image between two modes. (a) shows that visible information is dominant. (b) shows that both modes are important. (c) shows that infrared information is dominant.

How to make rational use of visible and infrared modal images is a problem worthy of study. The existing fusion tracking methods can be roughly divided into three categories: pixel-level fusion, feature-level fusion and decision-level fusion. Pixel-level[1] fusion is to fuse images with different modes to generate images with more information. Feature-level[2,3,4,6,9,10] fusion is to extract the features of different modes and fuse them according to the fusion rules designed by different methods. Decision level[11] fusion is to track each mode, and then fuse the results.

In the field of RGBT tracking, the trackers based on deep learning often adopts the feature-level fusion strategy, which have good performance. Although the unique information of different modes can complement each other, in some scenarios, the information that different modes can interact with is very limited, and even provides negative information. For example, those [2,6] who work in Siamese series account for 7/8 of the background information in the search image. Many works directly use feature-level fusion strategy to calculate the channel weight ratio of fused features, which inevitably contains a lot of background information and greatly affects the calculation of this ratio. We introduce a new dual-level fusion attention mechanism, which uses the decision-level and feature-level information to balance the infrared and visible modal features more reasonably.

On the other hand, the feature representation ability of the network affects its decision ability

for each mode, as well as the classification and regression results of the fused features. Inspired by [7,8], We introduce a new cross domain Siamese attention mechanism to realize the interaction of multi-domain information. For mode-domain, The spatial distribution of infrared and visible features should be related, and the crossed spatial attention can transmit spatial information to different modes. For time-domain, cross channel attention can use rich context information and provide an implicit way to update the template feature adaptively. Then, we classify the enhanced features to provide decision information for the dual-level balance module. Due to the lack of large-scale paired RGBT data sets, Some studies use gray images to replace infrared images for pre training, and then finetune on RGB-T dataset. The gray image is generated by the visible image, so the network has a strong dependent on the visible image. We classify each modes, which can also alleviate this dependence.

In recent years, the rise of RGBT tracking tasks and the wide application of attention mechanism have inspired the current work. However, at present, decision-level information is rarely introduced to participate in tracking, which can not reasonably allocate their modal weight ratio. Secondly, mode and time domain has rich context information, which is rarely used in the current research situation. In this study, we propose SiamMDA to improve the tracking performance of Siamese RGBT tracker in complex scenes. In SiamDMA network, the dual-level fusion attention mechanism can utilize the decision-level and feature-level information to allocate the two modal weight ratio more reasonably, the cross domain siamese attention mechanism can utilize rich context information to improve feature representation ability of network.

The main contributions of this work can be summarized as follows:

- By introducing the cross domain siamese attention mechanism, we propose a multi-domain aware module (MDAM). It can update the template feature adaptively, utilize rich context information of mode-domain and time-domain to improve feature representation ability of network. This module can be easily embedded in other work.
- By introducing the dual-level fusion attention mechanism, we propose a dual-level balance module (DLBM).It can utilize the decision-level and feature-level information to balance the two mode weight ratio more reasonably. This module can be easily embedded in other work.
- Based on SiamBAN [19] network, our method introduces multi-domain aware module and dual-level balance module to meet the challenge of RGBT tracking. Several tests were conducted on GTOT, VOT-RGBT2020 and LasHeR, our tracker achieves state-of-the-art results and keeps high speed （45FPS,RTX3060ti）。

2. Related Work

RGBT fusion tracking is one of the effective methods to improve the performance of tracker in recent years. This chapter will introduce the related work from the following three aspects.

1. Siamese tracker. 2. RGBT tracker. 3. Attention mechanism applied to tracking.

2.1 Siamese tracker

The trackers [12,13,14] based on correlation filtering algorithm has high speed, high performance and strong expansibility, but the manual characteristics restrict the discrimination ability of correlation filtering. SiamFC [15] introduced deep learning features into tracking, used siamese networks to replace manual features and implemented end-to-end training. The structure is simple and efficient. Compare with the correlation filter trackers, because siamFC adopted high-level semantic features, the template does not need to be updated, but the similarity is that siamfc utilized multi-scale prediction method to deal with the scale transformation of the target. SiamRPN [16] introduced regional proposal network [17] for classification and regression, which solves the problem of target scale transformation. SiamDW [18] has explored the adaptation of deep network in trackingand optimized the backbone network to avoid the impact of padding. SiamBAN, SiamCAR [6] and siamFC++ [20] introduced the anchor-free mechanism to change the regression branch from anchor-base, which avoided hyper-parameters associated with the candidate boxes. UpdateNet [21] was designed to update templates online, avoid interference in some complex scenarios. SiamAttn [8] used the deformable siamese attention mechanism to contact the context information between the template and the search branch which could implicitly update the target template. Ocean [22] used different scales for correlation operation, making the tracker more robust to target scale transformation.

2.2RGBT tracker

When the target is in a high exposure, low illumination environment or submerged in the background, it is difficult for the tracker to maintain good discrimination ability. SiamFT [23] introduced Siamese network structure into RGBT tracking, combined two modes' features for tracking while maintaining high speed. MANet [9] proposed a parallel network structure to extract single-mode unique features and two mode shared features, but its online tracking process can not achieve real-time. In [6], Guo et al. balanced the weight of different modal features, and used the idea of decision level fusion to construct the classification branches of two modes, so as to avoid the same contribution of different modes in complex scenes. JMMAC [11] considered both the appearance information and motion information of the target, and constructed a fusion tracker based on this. In [24], Li et al. made full use of annotation attributes and proposed a challenge aware network framework to deal with the significant changes in the appearance of targets

2.3 Attention mechanism applied to tracking

RASNet[25] introduced the attention mechanism for Siamese series trackers, added spatial

and channel attention to the target template, but only modified the template and ignored the search image branch. SiamAttn [8] introduced a deformable siamese attention network to jointly calculate self attention and cross channel attention, which could enhance the discrimination ability of the tracker. In [4], Xu et al. constructed attention mechanisms for their modes at different backbone layers, but limited the speed of the tracker.In [6], Guo et al. used the channel attention mechanism to redistribute weights for different modal features, could avoid the same contribution of different modes in complex scenes. SiamCDA [2] used the attention mechanism to reduce the gap between the two modal features through complementary information.

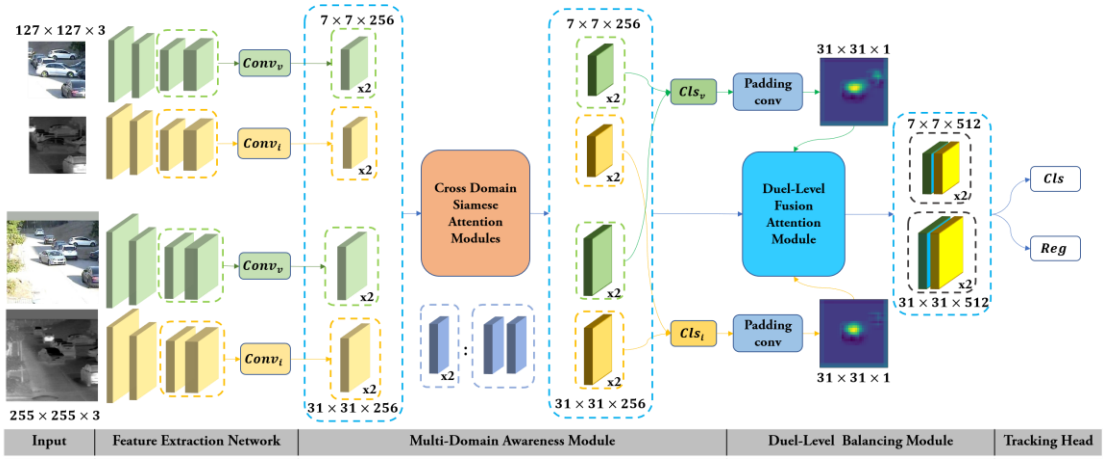

Figure 2. The framework of the proposed SiamDMA. It consists of Feature Extraction Network, Multi-Domain Aware Module and Duel-Level Balance Module. We feed the features of layers 3 and 4 in resnet50 [25] into MDAM to enhance each modal feature, and then classify each modal feature to obtain decision-level information. DLBM modulates decision-level and feature-level information to obtain fusion features. Finally, the fusion features are fed into the classification and regression head.

3. Methods

This chapter describes the details of SiamDMA network structure. As shown in Figure 2, SiamDMA network framework is improved on SiamBAN network, introducing dual-level fusion attention mechanism and cross domain siamese attention mechanism. Therefore, SiamDMA includes feature extraction network, MDAM, DLBM and tracking head.

Overview: We use the first four layers of resnet50 [25] as our backbone and feed the two modal template and search images to the feature extraction network to obtain the features. The features are enhanced by the MDAM. We classify each modal feature to obtain decision information, and then the decision-level and feature-level information are fed to the DLBM to balance the fused features, Finally, we got the location of target through classification and regression head.

3.1 Feature Extraction Network

In tracking, it is proved to be very effective [5,8,19] that fuse the output results of the last three layers of resnet50 [25]. However, in RGB-T tracking, if all resnet50 layers are used to extract features, the tracking speed will be greatly slowed down. If the layer 5 is removed, although part of the receptive field is reduced, the accuracy is only a little lost [19,22].
We use the first four layers of Resnet50 as our backbone to extract features, and the outputs of 3 and 4 layers are involved in the calculation of the following networks.
In the fourth layer network, the downsampling operation is replaced by atrous convolution. In order to extract each modal unique features and balance the speed and parameter quantity ,we set the parameters of the first two layers of our backbone as shared in all domains, and all parameters as shared in time-domain.
The first two layers of our backbone are marked as $\phi_{1,2}$,the 3, 4 layers of each mode-domain are marked as $\phi_{v3,v4}$, $\phi_{i3,i4}$。

In our backbone, the number of output channels of the 3 and 4 layers is different , so we reduce all features to 256 channels through 1x1 convolution layer. For the visible and infrared template branches, we crop the convoluted features and keep the center of 7×7 area. It can not only keep the whole target information, but also weaken the impact of the background [5,8,26]. For the search branches, we don't perform the crop operation. These convolution and crop operations of each mode are marked as $conv_v$, $conv_i$.

We mark the input visible template image as z_v, the infrared template image as z_i, the visible search image as x_v, the infrared template image as x_i. Then there are:

$$\begin{aligned} f_{zv} &= conv_v(\phi_{v3,v4}(\phi_{1,2}(z_v))) \\ f_{zi} &= conv_i(\phi_{i3,i4}(\phi_{1,2}(z_i))) \\ f_{xv} &= conv_v(\phi_{v3,v4}(\phi_{1,2}(x_v))) \\ f_{xi} &= conv_i(\phi_{i3,i4}(\phi_{1,2}(x_i))) \end{aligned} \quad (1)$$

Where f_{zv},f_{zi},f_{xv},f_{xi} represents the visible template, infrared template, visible search and infrared search features, output by the feature extraction network

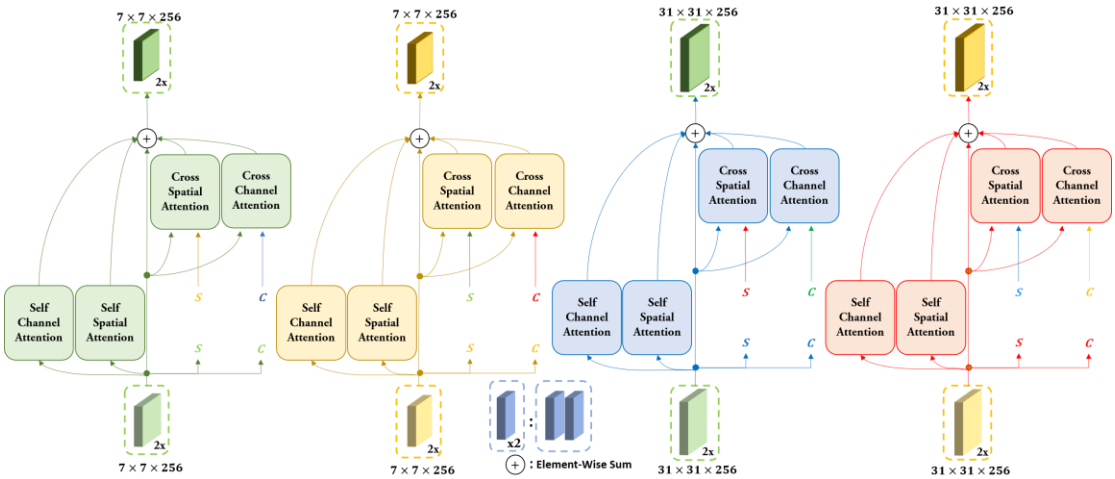

Figure 3. Illustration of cross domain siamese attention module. which consists of channel and spatial attention module, is subdivided into self and cross attention modes for different modulation objects. This module can interact the information of time-domain and mode-domain, and can also update the template feature adaptively. To avoid complex wiring in Figure 3, we simplified this figure by using jumper wires. The Corresponding wiring is letter S or C in the same color.

3.2 Multi-Domain Aware Module

As shown in Fig. 2, our proposed MDAM consists of a cross domain siamese attention module and two classification heads. The features obtained from the feature extraction network are input into the cross domain siamese attention module, modulated and interacted with multi domain context information. Then the modulated features are input into the classification heads to obtain the classification results. These classification results can be fed to the follow-up network as decision information.

[8] indicates that treating all channel features equally will hinder the representation ability. In addition, due to the limitation of convolution receptive field, only local information can be obtained at each position on the feature map, and the global context information cannot. Reasonable use of attention mechanism can alleviate the above limitations. In particular, for RGB-T target tracking, the information we can use to interact is more. In time-domain, we can interact more texture information. In mode-domain, we can make each position on the feature map obtain two modal global context. Inspired by this, we design a cross domain siamese attention mechanism to interact, meanwhile, it can update template feature adaptively.

As shown in Figure 3, the cross domain siamese attention module consists of channel and spatial attention module, which is subdivided into self and cross attention way for different modulation objects. In order to avoid complex wiring in Figure 3, we use wire jumpers to simplify the figure., and corresponding wiring is letter S or C of the same color.

For channel and spatial attention mechanism, both include query matrix Q, key matrix K and value matrix V [7,8]. Take the modulation of feature X to feature Y as an example.
Spatial attention mechanism. For feature X, Y, $X, Y \in C \times H \times W$. Q is generated by X through a 1x1 convolution. K is generated by Y through a 1x1 convolution. The number of Q and K matrix channels is modulated to 1/8 of the original number, $Q, K \in C' \times H \times W$. Where $C' = C/8$. Then reshape Q, K to $Q', K' \in C' \times N$, Where $N = H \times W$。 We can get that the attention attention map A as,

$$A = softmax_{-1}(Q'^{T} K'), A \in N \times N, (2)$$

$softmax_{-1}$ means normalizing the data of the last dimension of the feature array. For the feature Y to be modulated, the value matrix V is generated by Y through a 1x1 convolution, and then reshape V, Y to V', Y', $V', Y' \in C \times N$。Therefore, the spatial attention

feature S'^{Y}_{X} modulated by the input feature X to the feature Y as,

$$S'^{Y}_{X} = \alpha \cdot V'A + Y', S'^{Y}_{X} \in C \times N, (3)$$

Where α is a scalar parameter. Finally, reshape the modulated spatial attention feature to the size of feature Y to obtain S^{Y}_{X}, $S^{Y}_{X} \in C \times H \times W$。

Channel attention mechanism. The implementation method of matrix Q, K, V is different from that in spatial attention mechanism, For feature X, Y, $X \in C \times H_1 \times W_1, Y \in C \times H_2 \times W_2$, Where H_1 and W_1 need not be equal to H_2 and W_2. Feature X reshapes to generate Q, K, $Q, K \in C \times N_1, N_1 = H_1 \times W_1$. Then the attention map A as,

$$A = softmax\_{1}(QK^{T}), A \in C \times C, (4)$$

For feature Y to be modulated, Y reshapes the generated value matrix V, $V \in C \times N_2, N_2 = H_2 \times W_2$. Then the spatial attention feature C'^{Y}_{X} modulated by the input feature X to the feature Y as,

$$C'^{Y}_{X} = \beta \cdot reshape(AV) + Y, C^{Y}_{X} \in C \times N_2, (5)$$

Where β is a scalar parameter. Finally, the modulated channel attention feature is reshaped back to the size of feature Y to obtain C^{Y}_{X}, $C^{Y}_{X} \in C \times H_2 \times W_2$.

The features obtained by feature extraction network are $f_{zv}, f_{zi}, f_{xv}, f_{xi}$. After the cross domain siamese attention module, the features we obtained as,

$$\begin{aligned} F_{zv} &= S^{f_{zv}}_{f_{zv}} + C^{f_{zv}}_{f_{zv}} + S^{f_{zv}}_{f_{zi}} + C^{f_{zv}}_{f_{xv}} \\ F_{zi} &= S^{f_{zi}}_{f_{zi}} + C^{f_{zi}}_{f_{zi}} + S^{f_{zi}}_{f_{zv}} + C^{f_{zi}}_{f_{xi}} \\ F_{xv} &= S^{f_{xv}}_{f_{xv}} + C^{f_{xv}}_{f_{xv}} + S^{f_{xv}}_{f_{xi}} + C^{f_{xv}}_{f_{zv}} \\ F_{xi} &= S^{f_{xi}}_{f_{xi}} + C^{f_{xi}}_{f_{xi}} + S^{f_{xi}}_{f_{xv}} + C^{f_{xi}}_{f_{zi}} \end{aligned} \quad (6)$$

Where $F_{zv}, F_{zi}, F_{xv}, F_{xi}$ represent visible template, infrared template, visible search and infrared image features after cross domain aware module.

Finally, the modulated features are classified by two classification heads. The classification head refers to SiamBAN. We feed F_{zv},F_{xv} to visible classification module Cls_v. The visible light classification result V_{map} is obtained。Feed F_{zv},F_{xv} to visible classification module Cls_i The visible light classification result I_{map} is obtained。

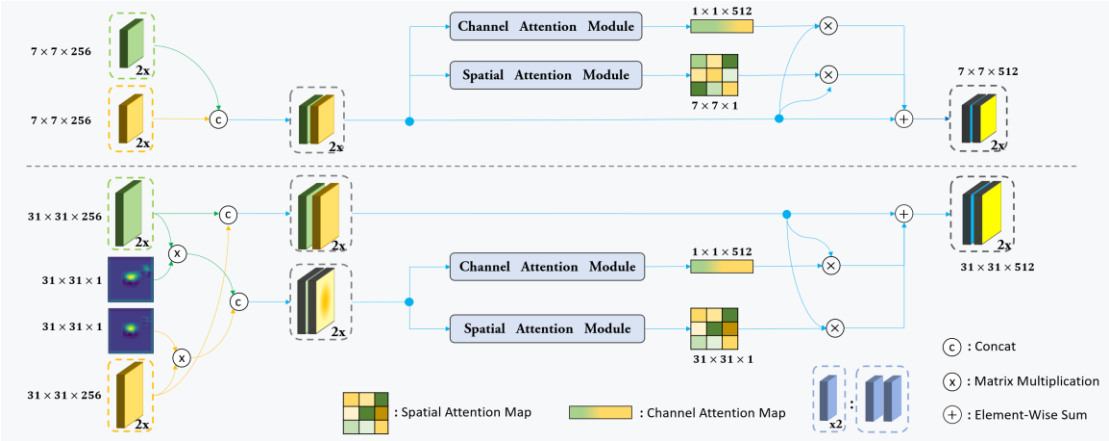

Figure 4. Illustration of dual-level fusion attention balance module, which uses the attention mechanism to

allocate the weight ratio of two modal features. It can utilize the decision-level and feature-level information to balance the two modal features' weight ratio more reasonably.

3.3 Dual-Level Balance Module

As shown in Figure 2, the DLBM is composed of two paddingconv modules and a dual-level fusion attention balance module. Since the classification results obtained in the multi-domain sensing module are two 25×25 maps, through the paddingconv module, the size of the maps is increased to 31×31 and aligned with the search featrues. It dilates the classification results. Then the features and dilated classification results are fed to the dual-level fusion attention module to allocate the weight ratio of fusion features.

Our work is based on SiamBAN, and the template and search area branches are fed to the network in the same way as SiamBAN. For template branches, we cut an area about twice the size of the target as our template, which is centered on the target.Obviously, the background area accounts for about 3/4. After multiple convolution layers, we only use the central region feature to feed the subsequent networ. The influence of the background is not significant. Therefore, in template branching, we directly use feature F_{zv},F_{zi} as the allocation source of fusion features.

However, for the search branch,an area about four times the size of the target are cropped out, and the background area accounts for about 7/8. After multiple convolution layers, the size becomes 31 × 31, and no crop operation is performed. The influence of the background is significant. Some studies [2,6] directly use the fusion features of the search area for weight allocation, which can not avoid the influence of the background. We use the mask generated by paddingconv module as an auxiliary to allocate the weight through the information of decision-level and feature-level.

Paddingconv module consists of two conv layers with padding and one relu layer, which adaptively dilates the classification results. We believe that the weight allocation of fusion features can not rely on the information of the whole graph, but on the distinguishability between the target and the background. Therefore, we use paddingconv to adaptively dilate the classification result to generate a mask that extracts only the features of the target and part of the background around the target.

We feed the classification results V_{map} and I_{map} to the paddingconv module to generate masks V_{mask} and I_{mask}. Through the mask, the key information such as the target's own information and the distinguishability between the target and the background are extracted.

$$K_{xv} = V_{mask} \cdot F_{xv}$$
$$K_{xi} = I_{mask} \cdot F_{iv}, \quad (11)$$

Where K_{xv} is visible feature's key information, K_{xi} is infrared feature's key information.

After obtaining the decision-level information, we use the dual-level information to balance the existing features. As shown in Figure 4, the dual-level fusion attention module uses the attention mechanism to allocate the weight ratio of two modal features. Different from the attention mechanism in Section 3.2, the attention mechanism in this section aims to realize the weight allocation of fused feature. Take the modulation of feature X to feature Y as an example.

Spatial attention mechanism. Input feature $X, X \in C \times H \times W$ Calculate the average pool and maximum pool in channel dimension, aggregate the channel information of feature X, and obtainf_{avg}, f_{max}。We concat f_{avg}, f_{max} and then pass through the conv layer to generate a 2D spatial attention map $F, F \in 1 \times H \times W$. F is calculated as

$$\begin{aligned} F &= \sigma\left(Conv^{7\times7}([Avgpool(X); Maxpool(X)])\right) \\ &= \sigma\left(Conv^{7\times7}([f_{avg}; f_{max}])\right) \end{aligned}, (7)$$

Where $Conv^{7\times7}$ represents a conv operation with 7×7 kernel size. σ Represents sigmoid function.

The spatial attention feature modulated by the input feature X to the feature Y as,

$$S_X^Y = \alpha \cdot FY + Y, S_X^Y \in C \times H \times W, (8)$$

Where α is a scalar parameter.

Channel attention mechanism. Input featrue $X, X \in C \times H \times W$ Calculate the average pool and maximum pool in spatial dimension, aggregate the spatial information of feature X, and obtain f_{avg}, f_{max}。We input f_{avg}, f_{max} into the full connection layer to generate the channel attention map $F, F \in C \times 1$. F is calculated as

$$\begin{aligned} F &= \sigma\left(FC(Avgpool(X)) + FC\left(Maxpool(X)\right)\right) \\ &= \sigma\left(FC\left(f_{avg}\right) + FC(f_{max})\right) \end{aligned}, (9)$$

Where FC represents the full connectivity layer. σ Represents sigmoid function.

Then the channel attention feature modulated by the input feature X to the feature Y as,

$$C_X^Y = \beta \cdot FY + Y, C_X^Y \in C \times H \times W, (10)$$

Where β is a scalar parameter.

Finally, we use concat function to combine F_{zv},F_{zi} into F_z , F_{xv},F_{xi} into F_x and K_{xv}, K_{xi} into F_x. Then the weight allocation method of two mode features as,

$$\begin{aligned} {F'}_Z &= C_{F_z}^{F_z} + S_{F_z}^{F_z} \\ {F'}_X &= C_{K_x}^{F_x} + S_{K_x}^{F_x}, (12) \end{aligned}$$

Where ${F'}_Z$,${F'}_X$ represent the template and search features after passing through the DLBM.

3.4GroundTruth And Loss

We use end-to-end training. Training loss is a weighted combination of visible mode classification loss, infrared mode classification loss, fusion feature classification and regression loss.

$$L = \lambda_1 L_{cls_v} + \lambda_2 L_{cls_v} + \lambda_3 L_{cls} + \lambda_4 L_{reg} \ , (20)$$

Among them, both classification head and regression head based on SiamBAN. The classification branch adopts cross entropy loss and elliptical classification label. The regression branch adopts anchor-free method and IoU loss. When training, we set $\lambda_1 = 0.2, \lambda_2 = 0.2, \lambda_3 = 1, \lambda_4 = 1$.

4.Experiments

4.1 Implementation Details

Training. The template image size is 127×127, the search image size is 255×255. Our model is trained for 20 epochs with Adaptive Moment Estimation(Adam) with a minibatch of 16 pairs. Weight decay is set as 0.0001. We using a warmup learning rate of 0.001 to 0.005 in the first 5 epochs and a learning rate exponentially decayed from 0.005 to 0.00005 in the last 15 epochs. Our backbone networks is initialized by the weights pre-trained on ImageNet [27]. At the beginning 10 epochs of training, freeze all the parameters of backbone networks, and then finetune the parameters of backbone networks' last two layers. In addition, we add high exposure, low illumination and blur strategies to make the image quality worse for data augmentation. We alternately degrade the image quality of the two modes, which is helpful to enhance the performance of our tracker.
(The specific way of data augmentation can be found in our project).

Inference. Refer to SiamBAN, we crop the template image from the first frame. For subsequent frames, crop the search image from each frame, and then we feed it to the network together with the template image to get the results of classification and regression. We use the regression results to punish the scale change and the cosine window to punish the distance from the search image's center [16]. This generates two weight masks, which are used to update the classification results. Then find the spatial position with the highest score in the updated classification result, and select the regression prediction box corresponding to the spatial position to update the current tracking box.

Our method is implemented in Python using PyTorch ,and we use Nvidia RTX 3060ti.

4.2 Dataset and Evaluation Metrics

GOT10K [28] contains more than 10000 visible sequences and 560 classes of objects, covering most moving objects comprehensively and fairly. On average, each sequence contains 150 frames, and each frame provides accurate manual annotation. Compared with similar tracking datasets, the classes are more abundant, which is very suitable for training tracking tasks.
LaSOT [29] contains 1400 visible sequences and 70 classes of objects. On average, each sequence contains 2500 frames, but the time interval between frames is smaller than GOT10K. Each frame provides accurate manual annotation.
GTOT [30] contains 50 visible and infrared paired sequences. On average, each single-mode sequence contains 150 frames. GTOT has 7 challenging attributes. However, the dataset have few classes, low resolution and poor quality.
RGBT234 [31] contains 234 visible and infrared paired sequences. On average, each single-mode sequence contains 150 frames. RGBT234 has 12 challenging attributes, and there are fewer classes of this dataset.
VOT-RGBT2020 [32] contains 60 visible and infrared paired sequences, and these 60 sequences are a subset of RGBT234. For ease of use, the sequences without VOT-RGBT2020 in RGBT234 we call it as rgbt174.
LasHeR [33] contains 1224 visible and infrared paired sequences. On average, each single-mode sequence contains 600 frames. It has 19 challenging attributes and 32 classes of objects. It is the first large-scale data set in the two mode tracking challenge

We use GOT10K, LaSOT pre training network, and use the gray image to replace the infrared image for end-to-end training. In order to be consistent with the training datasets used by the methods we compared. We finetune on RGBT234 to test GTOT, finetune on GTOT and rgbt174 to test VOT-RGBT2020, finetune on LasHeR training subset to test LasHeR testing subset.

When testing GTOT,We use precision rate (PR) and success rate (SR) as evaluation Metrics. PR is the percentage of frames whose distance between the output position and the ground truth position is within a threshold. We set this threshold to 5 pixels. SR is the percentage of frames whose overlap ratio between the output bounding box, and the ground truth bounding box is larger than the overlap threshold. We count the area under the curves (AUC) as SR score.

When testing VOT-RGBT2020, accuracy (A), robustness (R) and expected average overlap (EAO) are used to evaluate our trackers. Refer to the new EAO agreement in [34].

When testing LasHeR, precision rate (PR), success rate (SR) and normalized precision rate (NPR) are used to evaluate our trackers. PR and SR are the same as above, and sets the PR threshold to 20. The detailed calculation of NPR refer to [35].

	Ours	SiamBAN +RGBT	MANet	DAFNet	DAPNet	MACNet	SGT	M5L	FANet
PR	0.921	0.86	0.894	0.891	0.882	0.885	0.853	0.894	0.891
SR	0.756	0.706	0.724	0.712	0.707	0.698	0.636	0.709	0.728

Table 1. Results on GTOT, including SiamDMA, SiamBAN+RGBT,MANet [9], DAFNet [36], DAPNet [37], MACNet [38], SGT [39], M5L [40] and FANet [3]. SiamBAN+RGBT is our implemented tracker. After the visible and infrared features pass through the neck of SiamBAN, they are directly concatenated and then fed concatenated features to the SiamBAN's head. The red, blue, and green fonts represent the first three values.

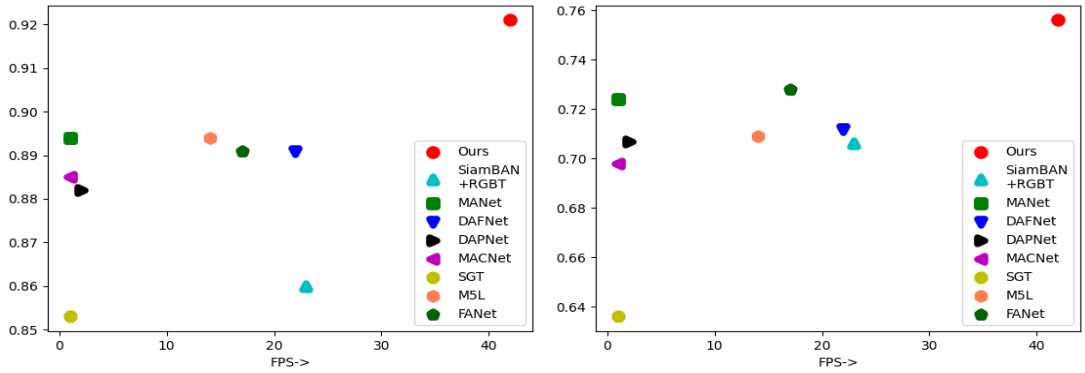

Figure 5. Speed comparison of various trackers on GTOT. The left figure compares PR and FPS, and the right figure compares SR and FPS

	Ours	SiamBAN +RGBT	JMMAC	AMF	DFAT	SiamDW -T	mfDiMP	SNDCFT	M2C2Frgbt
A	0.637	0.654	0.662	0.63	0.672	0.654	0.638	0.630	0.636
R	0.816	0.751	0.818	0.822	0.779	0.791	0.793	0.789	0.722
EAO	0.39	0.355	0.42	0.412	0.39	0.389	0.38	0.378	0.332

Table 2. Results on VOT-RGBT2020, including SiamDMA, SiamBAN+RGBT, and seven trackers from the VOT RGBT 2020 challenge [32]. The red, blue, and green fonts represent the first three values.

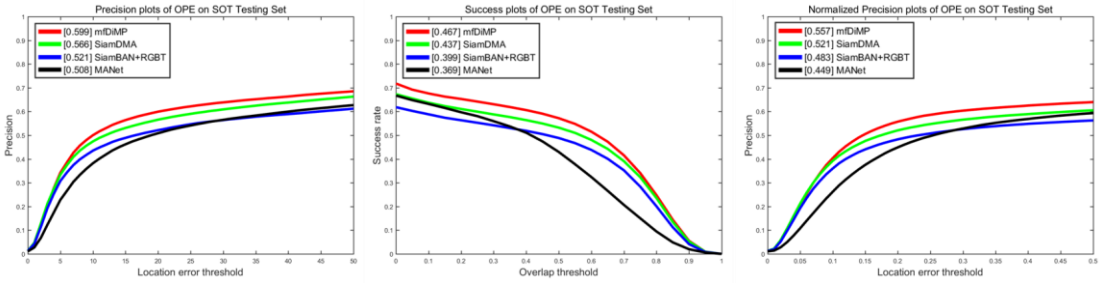

Figure 6. Results on LasHeR using precision (PR), normalizes precision (NPR)and success (SR) plots, including SiamDMA, SiamBAN+RGBT, mfDIMP and MANet. Only the test results of MANet and mfDIMP are published in [33].

4.3 Comparison with State-of-the-art Trackers

GTOT： Table 1 and Figure 5 show the comparison results on GTOT with short sequences. Our tracker obtained a PR of 0.921 and an SR of 0.756. Previously, the tracker with best performance was MANet, which achieved PR of 0.894 and SR of 0.724. Compared with it, our tracker surpasses its PR of 2.7% and SR of 3.2%, and our FPS exceeds it too. Compared with our benchmark SiamBAN+RGBT, our tracker surpasses its PR of 6.1% and SR of 5%.

VOT-RGBT2020: Table 2 shows the comparison results on VOT-RGBT2020 with long sequences. Our tracker achieves 0.637 accuracy, 0.816 robustness and 0.39 EAO. The EAO value is consistent with the DFAT, which is the champion of the VOT RGBT 2020 challenge. Compared with our benchmark SiamBAN+RGBT, our tracker surpasses its robustness of 6.5% and EAO of 3.5%.

LasHeR: Figure 6 shows the comparison results on LasHeR with long sequences. In [33], It is worth noting that only the test results of MANet and mfDIMP are published. Our tracker achieves 0.566 PR,0.437 SR and 0.521 NPR. The three indexes are lower than mfDIMP and higher than MANet. Compared with our benchmark siamban(rgbt), our tracker surpasses its PR of 4.5% and SR of 3.8%, beyond its NPR of 3.8%.

Method	PR	ΔPR	SR	ΔSR	NPR	ΔNPR
Baseline	**0.519**	**-**	**0.402**	**-**	**0.486**	**-**
Baseline + CDSAM	**0.540**	**+2.1%**	**0.417**	**+1.5%**	**0.496**	**+1.0%**
Baseline + CDSAM + CLS	**0.547**	**+2.8%**	**0417**	**+1.5%**	**0.501**	**+1.5%**
Baseline + CDSAM + CLS + DLBM	**0.566**	**4.7%**	**0.437**	**+3.5%**	**0.521**	**+3.5%**

Table 3. Results of ablation experiments on model structures. SiamBAN(no layer 5)+RGBT is our baseline which is a tracker obtained by removing all added modules from SiamDMA. CDSAM: cross domain siamese attention module. CLS: the classification heads for each modes. DLBM: dual-level balance module.

MDAM		DLBM		PR	SR	NPR
CDSAM	DLFAM	CDSAM	DLFAM			
✓		✓		**0.559**	**0.432**	**0.514**
✓			✓	**0.566**	**0.437**	**0.521**
	✓	✓		**0.542**	**0.424**	**0.515**
	✓		✓	**0.545**	**0.428**	**0.506**

Table 4. Results of ablation experiments on attention mode. MDAM: multi-domain aware module. DLBM:

dual-level balance module. CDSAM: cross domain siamese attention mechanism which is the implementation method of attention mechanism in MDAM. DLFAM: dual-level fusion attention mechanism which is the implementation method of attention mechanism in DLBM.

Method	PR	SR	NPR
Feature-Level	**0.552**	**0.436**	**0.406**
Feature-Level + Decision-Level	**0.566**	**0.437**	**0.521**

Table 5. Results of ablation experiments on source of balance module. Feature-Level means that directly use feature level information to balance the mode features' weight ratio. Feature-Level + Decision-Level means that use feature-level and Decision-Level information to balance the mode features' weight ratio.

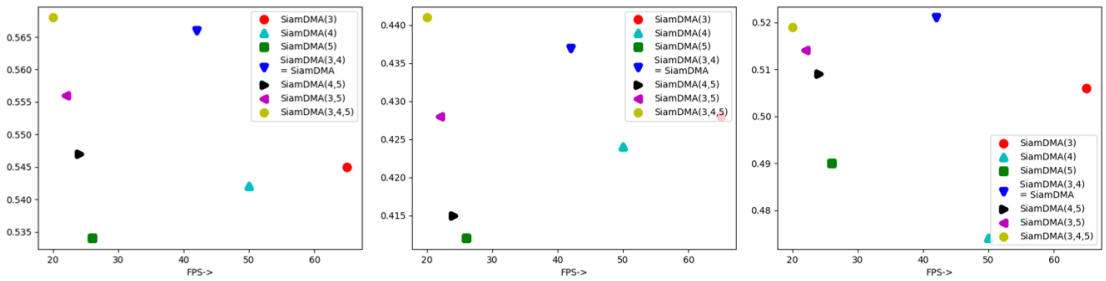

Figure 7. Results of ablation experiments on efficiency analysis. The left figure compares PR and FPS, the mid figure compares SR and FPS and the right figure compares NPR and FPS. The numbers 3, 4 and 5 represent that the output results of the resnet50's corresponding layers are used to participate in the calculation of subsequent networks in SiamDMA.

4.4 Ablation Study

We study the impact of individual components in SiamDMA, and conduct ablation study on LasHeR testting subset.

Model architecture. Table 3 shows the results of ablation experiments on model architecture. We use SiamBAN(no layer 5)+RGBT as baseline. SiamBAN(no layer 5)+RGBT is a tracker obtained by removing all added modules from SiamDMA. Unlike SiamBAN+RGBT, it does not use the layer 5 of resnet50. By adding the cross domain siamese attention module, the indexs PR, SR and NPR are improved from 0.519 to 0.540, 0.402 to 0.417 and 0.486 to 0.496. It shows that the interaction of rich context information is very important for two modal tracking, which makes the tracker more robust. Then we classify the enhanced features, which improves the indexs from 0.540 to 0.547, 0.417 to 0.417 and 0.496 to 0.501. It can alleviate the dependence of the network on the visible mode and ensure that each mode can be extracted with rich features. Finally, we introduce the DLBM, which makes our indexs finally from 0.547 to 0.566,

0.417 to 0.437 and 0.501 to 0.521. The final results are 4.7%, 3.5% and 3.5% higher than the baseline respectively.

Attention mode. In this paper, the attention mechanism of cross domain siamese attention mechanism and dual-level fusion attention mechanism are different. We have made replacement tests on their implementation methods, as shown in Table 4. The implementation of reference [7] in the cross domain siamese attention mechanism and the implementation of reference [42] in the dual-level fusion attention mechanism have the highest index performance in LasHeR.

Source of balance module. Our method uses decision-level and feature-level information as the input of dual-level fusion attention mechanism. Different from [2,6] and other studies, they directly use feature level information to balance the mode features' weight ratio. We also conducted the experiment of changing the input source on the DLBM, as shown in Table 5. The results show that, the introduction of decision-level information is a more reasonable way to balance two mode features' weight ratio.

Efficiency analysis. In order to give consideration to accuracy and speed, we use the first four layers of Resnet50 as our backbone to extract features, and the outputs of 3 and 4 layers are involved in the calculation of the following networks. Our tracker has reached 45 fps on 3060ti. We replace our backbone, and the results are shown in figure 7. Although using all layers of resnet50 as the backbone has the highest performance, fps is only 20.

5.Conclusion

We design a siamese dual-level and multi-domain attention network for RGBT tracking. Cross domain siamese attention mechanism and dual-level fusion attention mechanism are introduced. The former uses the rich context correlation of mode domain and time domain to improve the feature extraction ability of network and adaptively update template features. The latter combines the information of decision-level and feature-level, which provides a more reasonable way to balance the two mode features' weight ratio. They can be easily embedded in other tracking work. We conducted several experiments on three data sets, our tracker achieves state-of-the-art results and keeps high speed.

REFERENCES

[1] Zhang, X. , Ye, P. , Leung, H. , Gong, K. , & Xiao, G. . (2020). Object fusion tracking based on visible and infrared images: a comprehensive review. Information Fusion, 63(November 2020), 166-187.

[2] Zhang, T. , X Liu, Qiang, Z. , & Han, J. . (2021). Siamcda: complementarity-and distractor-aware rgb-t tracking based on siamese network. IEEE Transactions on Circuits and Systems for Video Technology, PP(99), 1-1.

[3] Zhu, Y. , Li, C. , Tang, J. , & Luo, B. . (2020). Quality-aware feature aggregation network for robust rgbt tracking. IEEE Transactions on Intelligent Vehicles, PP(99), 1-1.

[4] Xu, Q. , Mei, Y. , Liu, J. , & Li, C. . (2021). Multimodal cross-layer bilinear pooling for rgbt tracking. IEEE Transactions on Multimedia, PP(99), 1-1.

[5] Guo, D. , Wang, J. , Cui, Y. , Wang, Z. , & Chen, S. . (2020). SiamCAR: Siamese Fully Convolutional Classification and Regression for Visual Tracking. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.

[6] Guo, C. , Yang, D. , Li, C. , & Song, P. . (2021). Dual siamese network for rgbt tracking via fusing predicted position maps. The Visual Computer, 1-13.

[7] Fu, J. , Liu, J. , Tian, H. , Li, Y. , Bao, Y. , & Fang, Z. , et al. (2020). Dual Attention Network for Scene Segmentation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.

[8] Yu, Y. , Xiong, Y. , Huang, W. , & Scott, M. R. . (2020). Deformable siamese attention networks for visual object tracking.

[9] Lu, A. , Li, C. , Yan, Y. , Tang, J. , & Luo, B. . (2021). Rgbt tracking via multi-adapter network with hierarchical divergence loss. IEEE Transactions on Image Processing, PP(99), 1-1.

[10] Zhang, X. , Ye, P. , Peng, S. , Liu, J. , & Xiao, G. . (2020). Dsiammft: an rgb-t fusion tracking method via dynamic siamese networks using multi-layer feature fusion. Signal Processing Image Communication.

[11] Zhang, P. , J Zhao, Wang, D. , Lu, H. , & Yang, X. . (2020). Jointly modeling motion and appearance cues for robust rgb-t tracking.

[12] Bolme, D. S. , Beveridge, J. R. , Draper, B. A. , & Lui, Y. M. . (2010). Visual object tracking using adaptive correlation filters. The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010. IEEE.

[13] Henriques, J. F. , Caseiro, R. , Martins, P. , & Batista, J. . (2015). High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis & Machine Intelligence, 37(3), 583-596.

[14] Danelljan, M. , Bhat, G. , Khan, F. S. , & Felsberg, M. . (2016). ECO: Efficient Convolution Operators for Tracking. IEEE Computer Society. IEEE Computer Society.

[15] Bertinetto, L. , Valmadre, J. , Henriques, J. F. , Vedaldi, A. , & Torr, P. . (2016). Fully-Convolutional Siamese Networks for Object Tracking. European Conference on Computer Vision. Springer, Cham.

[16] Bo, L. , Yan, J. , Wei, W. , Zheng, Z. , & Hu, X. . (2018). High Performance Visual Tracking with Siamese Region Proposal Network. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.
[17] Ren, S. , He, K. , Girshick, R. , & Jian, S. . (2017). Faster r-cnn: towards real-time object detection with region proposal networks.
[18] Zhang, Z. , & Peng, H. . (2019). Deeper and wider siamese networks for real-time visual tracking.
[19] Chen, Z. , Zhong, B. , Li, G. , Zhang, S. , & Ji, R. . (2020). Siamese Box Adaptive Network for Visual Tracking.
[20] Xu, Y. , Wang, Z. , Li, Z. , Yuan, Y. , & Yu, G. . (2020). Siamfc++: towards robust and accurate visual tracking with target estimation guidelines. Proceedings of the AAAI Conference on Artificial Intelligence, 34(7), 12549-12556.
[21] Zhang, L. , Gonzalez-Garcia, A. , Joost, V. , Danelljan, M. , & Khan, F. S. . (2019). Learning the model update for siamese trackers.
[22] Zhang, Z. , & Peng, H. . (2020). Ocean: Object-aware Anchor-free Tracking.
[23] Zhang, X. , Ye, P. , Peng, S. , Liu, J. , Gong, K. , & Xiao, G. . (2019). Siamft: an rgb-infrared fusion tracking method via fully convolutional siamese networks. IEEE Access.
[24] C. Li, L. Liu, A. Lu, Q. Ji, and J. Tang. . (2020). Challenge-Aware RGBT Tracking. European Conference on Computer Vision (ECCV 2020),
[25] Wang, Q. , Teng, Z. , Xing, J. , Gao, J. , & Maybank, S. . (2018). Learning Attentions: Residual Attentional Siamese Network for High Performance Online Visual Tracking. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE.
[26] Li, B. , Wu, W. , Wang, Q. , F Zhang, J Xing, & J Yan. (2020). SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.
[27] Song, Y. , Chao, M. , Wu, X. , Gong, L. , & Yang, M. H. . (2018). VITAL: VIsual Tracking via Adversarial Learning. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE.
[28] Lianghua, Huang, Xin, Zhao, & Kaiqi. (2019). Got-10k: a large high-diversity benchmark for generic object tracking in the wild. IEEE transactions on pattern analysis and machine intelligence.
[29] H Fan, H Ling, Lin, L. , Yang, F. , & Liao, C. . (2019). LaSOT: A High-Quality Benchmark for Large-Scale Single Object Tracking. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.
[30] C. L. Li, H. Cheng, S. Y. Hu, X. B. Liu, J. Tang, and L. Lin, Learning Collaborative Sparse Representation for Grayscale-Thermal Tracking, IEEE Transactions on Image Processing, vol. 25, no. 12, pp. 5743-5756, Dec, 2016.
[31] C. L. Li, X. Y. Liang, Y. J. Lu, N. Zhao, and J. Tang, RGB-T object tracking: Benchmark and baseline, Pattern Recognition, vol. 96, Dec, 2019.
[32] www.votchallenge.net/vot2020
[33] arxiv.org/pdf/2104.13202.pdf (Chenglong Li, LasHeR: A Large-scale High-diversity Benchmark for RGBT Tracking)
[34] Memarmoghadam, A. . (2021). The Eighth Visual Object Tracking VOT2020 Challenge Results. The Eighth Visual Object Tracking VOT2020 Challenge Results.

[35] Mueller, M. , Bibi, A. , Giancola, S. , Alsubaihi, S. , & Ghanem, B. . (2018). Trackingnet: a large-scale dataset and benchmark for object tracking in the wild. Springer, Cham.
[36] Gao, Y. , Li, C. , Zhu, Y. , Tang, J. , & Wang, F. . (2019). Deep Adaptive Fusion Network for High Performance RGBT Tracking. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). IEEE.
[37] Y. Zhu, C. Li, B. Luo, J. Tang, and X. Wang, Dense feature aggregation and pruning for rgbt tracking, Proceedings of ACM International Conference on Multimedia, 2019.
[38] H Zhang, Zhang, L. , Zhuo, L. , & Zhang, J. . (2020). Object tracking in rgb-t videos using modal-aware attention network and competitive learning. Sensors, 20(2), 393.
[39] C. Li, N. Zhao, Y. Lu, C. Zhu, and J. Tang, Weighted sparse representation regularized graph learning for rgb-t object tracking, in Proceedings of the 25th ACM international conference on Multimedia, 2017.
[40] Tu, Z. , Lin, C. , Li, C. , Tang, J. , & Luo, B. . (2020). M^{5}l: multi-modal multi-margin metric learning for rgbt tracking.
[41] Zhang, L. , Danelljan, M. , Gonzalez-Garcia, A. , Weijer, J. , & Khan, F. S. . (2019). Multi-modal fusion for end-to-end rgb-t tracking. IEEE.
[42] Jie, H. , Li, S. , Gang, S. , & Albanie, S. . (2017). Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, PP(99).