SiamDL: Siamese Dual-level Fusion Attention Network for RGBT Tracking

Fengchen He[a], Mingyang Chen[b], Xiaoyu Chen[c], Jing Han[d,*] and Lianfa Bai[e]

[a]Jiangsu Key Laboratory of Spectral Imaging and Intelligent Sense, Nanjing University of Science and Technology, Nanjing 210094, China

ARTICLE INFO

Keywords:
RGBT tracking
Dual-level balance
Multi-domain aware
Mode-domain
Time-domain
Siamese network.

ABSTRACT

Most of the existing dual-mode tracking algorithms rely on feature fusion design. We propose a Siamese Dual-level Fusion Attention Network (SiamDL) for RGBT Tracking by combining dual-level balance module and multi-domain aware module. Dual-level balance module (DLBM) introduces a new dual-level fusion attention mechanism to utilize the two modal information at decision-level and feature-level, which is used to provide a more reasonable way to balance the two modal features' weight ratio. Multi-domain aware module (MDAM) introduces a new cross domain siamese attention mechanism to make mode-domain (referring to visible and infrared modal branches) and time-domain (referring to template and search branches) information interact with each other, which is used to enhance feature expression ability of network and adaptively update template features. The average tracking speed on rtx3060ti is about 45fps, which suggests that SiamDL has achieved state-of-the-art performance by carrying out experiments on three RGB-T tracking benchmark datasets.

1. Introduction

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 significant development. However, when the target is in dark light, high exposure or submerged in the background, the tracking effect of visible mode will be greatly reduced.

In most cases, the visible image is rich in the structure and color information of the target, and the infrared image is rich in the structure information of the target, which is highly complementary to the visible image. Thus, introduction of infrared mode will improve the performance of the tracker predictably [38]. Therefore, the two modal fusion tracking algorithm has emerged year by year, and we show the complementarity of infrared and visible images in Fig. 1.

Figure 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.

Based on the present problems, how to realize the fusion and utilization of visible and infrared modal images is worth discussing. The existing fusion tracking methods can be roughly divided into three categories: pixel-level fusion, feature-level fusion and decision-level fusion methods. Pixel-level [38] fusion methods fuse images with different modes to generate images with more information. Feature-level [37, 43, 32, 9, 23, 40] fusion methods extract the features of different modes and fuse them according to the fusion rules designed by different methods. Decision-level [36] fusion methods track each mode, and then fuse the results.

In the field of RGBT tracking, the tracker based on deep learning often adopts feature-level fusion strategy, which has improved significantly. For bimodal fusion features, equal treatment of its channel weight will hinder the expression ability [34]. 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. Therefore, there should be different weight ratios of fusion features for different scenes. Methods [26, 37, 9, 43, 32] directly use the feature-level fusion strategy to calculate the channel weight ratio of fused features, but because the background information accounts for a high proportion in the search image, it inevitably contains a large amount of background information, which greatly affects the calculation of the fusion features' weight ratio. We introduce a new dual-level fusion attention mechanism, which uses the information of decision-level and feature-level to balance the fusion features more reasonably.

Additionally, the feature expression ability of the network affects its decision ability of each mode, as well as the classification and regression results of the fused features. Inspired by [8, 34], we introduce a new cross domain siamese attention mechanism to realize the interaction of multi-domain information. For mode-domain, the spatial distribution of

ORCID(s): 0000-0003-1026-2824 (X. Chen); 0000-0002-1033-566X (J. Han)

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. Some researchers use gray images as fake infrared images for pre training to deal with the shortcomings of large-scale paired RGBT data sets, and then fine tune the RGBT data sets. However, due to the gray image is generated by visible image, the network has a strong dependence on visible image. We classify each pattern, which can also reduce this dependency.

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 misses the important information of two modal weight allocation. Besides, mode-domain and time-domain has rich context information, which is rarely used in the current research situation. In this study, we propose SiamDL to improve the tracking performance of siamese RGBT tracker in complex scenes. In SiamDL network, the dual-level fusion attention mechanism is used to utilize the decision-level and feature-level information to allocate the two modal weight ratio more reasonably, and the cross domain siamese attention mechanism is proposed to utilize rich context information to improve feature expression 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 expression ability of network.
- 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 modal weight ratio more reasonably.
- Based on SiamBAN [4], 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 [3, 11, 6] based on correlation filtering algorithm have high speed, high performance and strong expansibility, but the manual features restrict the discrimination ability of correlation filtering. To overcome this defect, SiamFC [1] introduced deep learning features into tracking, used siamese networks to replace manual features and implemented end-to-end training. Consequently, the structure is simple and efficient. Compare with the correlation filter trackers, SiamFC does not need to be updated for adopting high-level semantic features. SiamRPN [16] introduced regional proposal network [27] for classification and regression, which solved the problem of target scale transformation. SiamDW [41] had explored the adaptation of deep network in tracking and optimized the backbone network to avoid the impact of padding. SiamBAN, SiamCAR [9] and SiamFC++ [33] introduced the anchor-free mechanism to change the regression branch from anchor-base, which avoided hyper-parameters associated with the candidate boxes. UpdateNet [35] was designed to update templates online, avoiding interference in some complex scenarios. SiamAttn [34] adopted 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 [42] applied different scales for correlation operation, making the tracker more robust to target scale transformation. The above trackers only use the visible light mode as the information source, and can not deal with the complex lighting transformation scene.

2.2. RGBT 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 [39] introduced siamese network structure into RGBT tracking, combining two modal features for tracking while maintaining high speed. MANet [23] proposed a parallel network structure to extract single-mode unique features and two modal shared features, but its online tracking process can not achieve real-time. Guo et al. [9] balanced the weight of different modal features, and used the strategy 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 [36] constructed a fusion tracker by comprehensively considering the appearance information and motion information of the target. Li et al. [19] made full use of annotation attributes and proposed a challenge aware network framework to deal with the significant changes in the appearance of targets. Most of the above trackers focus on the fusion of feature-level information and rarely use important decision-level information.Wang et al.[31] propose an adaptive fusion algorithm based on response map evaluation for RGBT tracking. This demonstrates the availability and effectiveness of response graphs.But the difference is that we redesign the feature fusion using decision-level information.Shen et al.[28] propose a new algorithm, called cooperative low-rank graph model, to suppress background

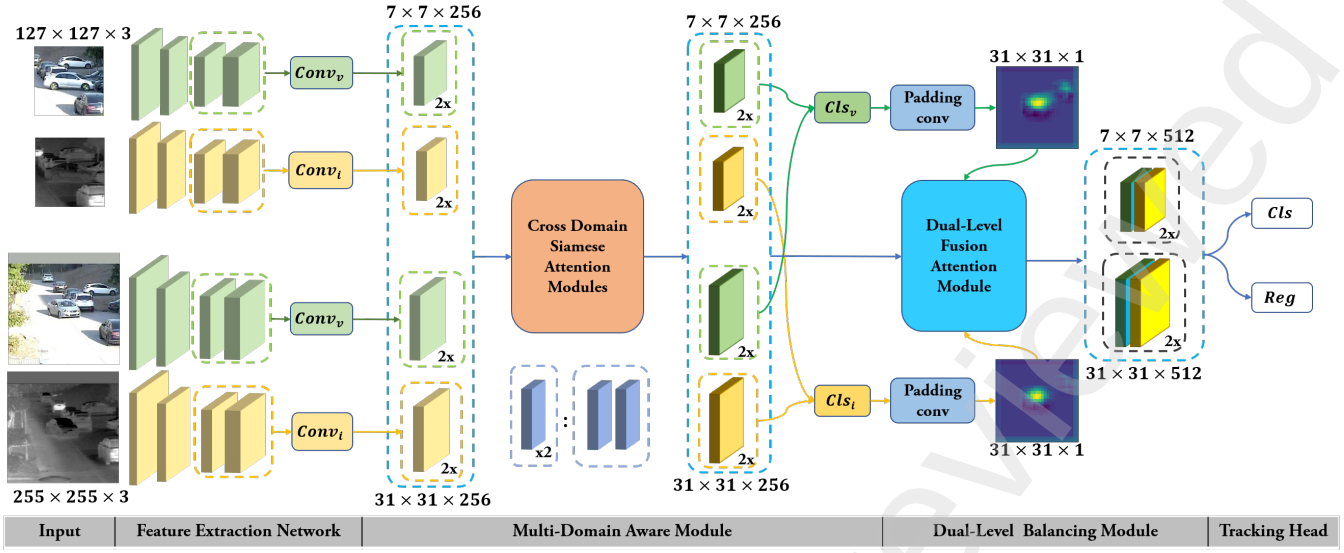

Figure 2: The framework of the proposed SiamDL, which consists of Feature Extraction Network, Multi-Domain Aware Module (MDAM) and Dual-Level Balance Module (DLBM). We feed the features of layers 3 and 4 in ResNet50 [30] into MDAM to enhance each modal feature, and then classify each modal feature to 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.

clutter.This also inspired us to add a classification branch to enhance the foreground and weaken the influence of background clutter.

2.3. Attention mechanism applied to tracking

RASNet [30] 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 [34] introduced a deformable siamese attention network to jointly calculate self-attention and cross channel attention, which could enhance the discrimination ability of the tracker. Xu et al. [32] constructed attention mechanisms for their modes at different backbone layers, but limited the speed of the tracker. Zhu et al. [43] used the channel attention mechanism to redistribute weights for different modal features, which could avoid the same contribution of different modes in complex scenes. SiamCDA [37] applied the attention mechanism to bridge the gap between the two modal features through complementary information. CMC2R[22] fuses local features and global representations under different resolutions through the transformer layer of the encoder block, and the two modalities are collaborated to get contextual information by the spatial and channel self-attention.For two modal scenarios, decision-level information can also be one of the important sources of attention mechanism applied to tracking.

3. Methods

This chapter describes the details of SiamDL network structure. As shown in Fig. 2, SiamDL takes SiamBAN as baseline, introduces dual-level fusion attention mechanism and cross domain siamese attention mechanism. Therefore, SiamDL includes feature extraction network, MDAM, DLBM and tracking head.

Overview: We use the first four layers of ResNet50 [30] as our backbone and feed the two modal template and search images to the feature extraction network to obtain the features. After that, the features are enhanced by the MDAM. The next step is 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 get the location of target through classification and regression head.

3.1. Efficient Feature Extraction Network Design

In tracking, it is proved to be very effective [10, 34, 4] that fuse the output results of the last three layers of ResNet50 [30]. 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, part of the receptive field is reduced with only a little loss of accuracy [4, 42]. 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 are 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

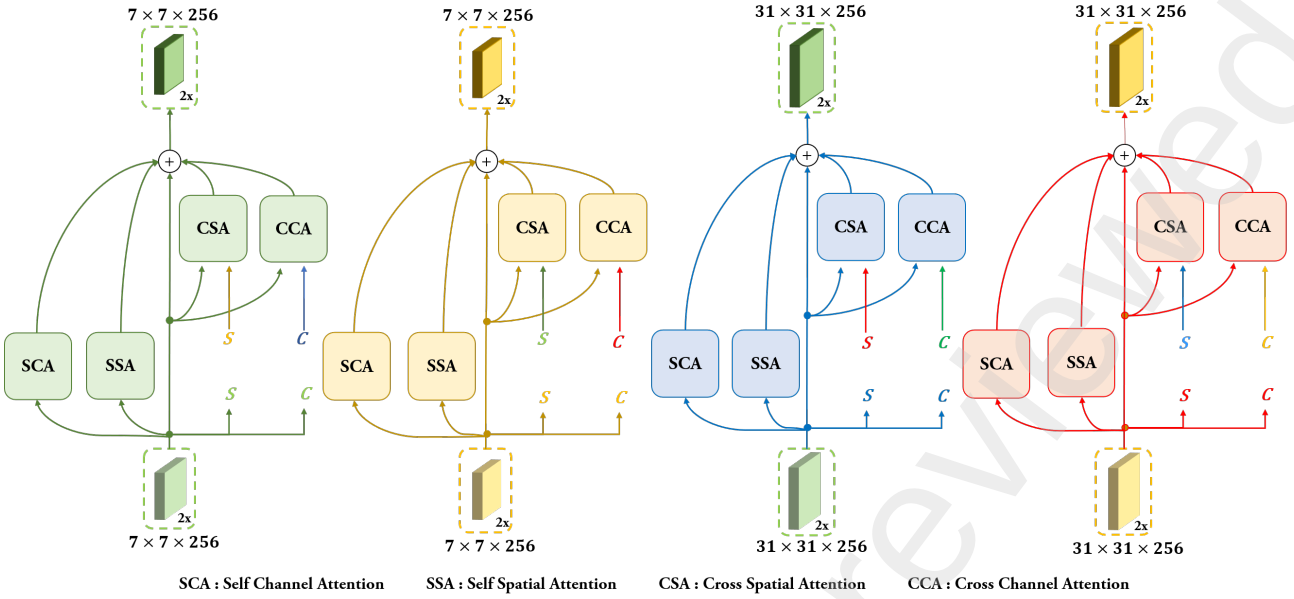

Figure 3: Illustration of cross domain siamese attention module. It consists of channel and spatial attention module, which is subdivided into self and cross attention modes for different modulation objects. This module is used to realize the interaction of time-domain and mode-domain information, furthermore, it can also update the template feature adaptively. To avoid complex wiring in Fig. 3, we simplified this figure by using jumper wires. And the corresponding wiring is lettered S or C in the same color.

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 [10, 34, 15]. For the search branches, we do not perform the crop operation. And 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, and the infrared search image as x_i. Then there are:

$$f_{zv} = conv_v(\phi_{v3,v4}(\phi_{1,2}(z_v))) \tag{1a}$$
$$f_{zi} = conv_i(\phi_{i3,i4}(\phi_{1,2}(z_i))) \tag{1b}$$
$$f_{xv} = conv_v(\phi_{v3,v4}(\phi_{1,2}(x_v))) \tag{1c}$$
$$f_{xi} = conv_i(\phi_{i3,i4}(\phi_{1,2}(x_i))) \tag{1d}$$

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.

3.2. Multi-Domain Aware Module

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

Yu, Y et al. [34] proposed that treating all channel features equally will hinder the representation ability. In addition, and the rational use of attention mechanism can alleviate this limitations. In particular, there is more information for us to interact in RGB-T target tracking. For example, we can interact more texture information in time-domain, while we can make each position on the feature map obtain two modal global context in mode-domain. Inspired by this, we design a cross domain siamese attention mechanism to interact, meanwhile, it can update template feature adaptively.

As shown in Fig. 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 Fig. 3, we introduce wire jumpers to simplify the figure, and corresponding wiring is lettered S or C of the same color.

Both channel and spatial attention mechanism include query matrix Q, key matrix K, and value matrix V [8, 34]. Take the modulation of feature X to feature Y as an example.

Spatial attention mechanism in MDAM. For feature X, Y, X, Y, $X, Y \in C \times H \times W$. Q is generated by X through a 1×1 convolution. The number of Q, 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 $N = H \times W$.Therefore, we can get the attention map A as:

$$A = softmax(Q'^{T} K'), A \in N \times N \tag{2}$$

$softmax$ means normalizing the data of the last dimension of the feature array. As for modulating feature Y , the value matrix V is generated by Y through a 1×1 convolution, and then reshape V, Y to $V', Y', V', Y' \in C \times N$.Therefore,

the spatial attention feature $S_X^{'Y}$ is modulated by the input feature X to the feature Y as:

$$S_X^{'Y} = \alpha \cdot V'A + Y', S_X^{'Y} \in C \times N \tag{3}$$

Where α is a scalar parameter. Finally, we reshape the modulated spatial attention feature to the size of feature Y to obtain $S_X^Y, S_X^Y \in C \times H \times W$.

Channel attention mechanism in MDAM. 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 is described as:

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

As for modulating feature Y, 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 $C_X^{'Y}$ modulated by the input feature X to the feature Y is described as:

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

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

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

$$F_{zv} = S_{f_{zv}}^{f_{zv}} + C_{f_{zv}}^{f_{zv}} + S_{f_{zi}}^{f_{zv}} + C_{f_{xv}}^{f_{zv}} \tag{6a}$$
$$F_{zi} = S_{f_{zi}}^{f_{zi}} + C_{f_{zi}}^{f_{zi}} + S_{f_{zv}}^{f_{zi}} + C_{f_{xi}}^{f_{zi}} \tag{6b}$$
$$F_{xv} = S_{f_{xv}}^{f_{xv}} + C_{f_{xv}}^{f_{xv}} + S_{f_{xi}}^{f_{xv}} + C_{f_{zv}}^{f_{xv}} \tag{6c}$$
$$F_{xi} = S_{f_{xi}}^{f_{xi}} + C_{f_{xi}}^{f_{xi}} + S_{f_{xv}}^{f_{xi}} + C_{f_{zi}}^{f_{xi}} \tag{6d}$$

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. And the classification head refers to SiamBAN. We feed F_{zv},F_{xv} to visible classification module Cls_v to obtain the visible light classification result V_{map}, feed F_{zv},F_{xv} to visible classification module Cls_i to get the visible light classification result I_{map}.

3.3. Dual-Level Balance Module

As shown in Fig. 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 features, which 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.

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 module 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.

The template and search images 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 network. 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 is 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 [26, 37, 9, 43, 32] 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.

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:

$$K_{xv} = V_{mask} \cdot F_{xv} \tag{7a}$$
$$K_{xi} = I_{mask} \cdot F_{iv} \tag{7b}$$

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 Fig. 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 B, 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 in DLBM. 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 obtain f_{avg}, f_{max}. We concat f_{avg}, f_{max} and then pass through the conv layer to generate

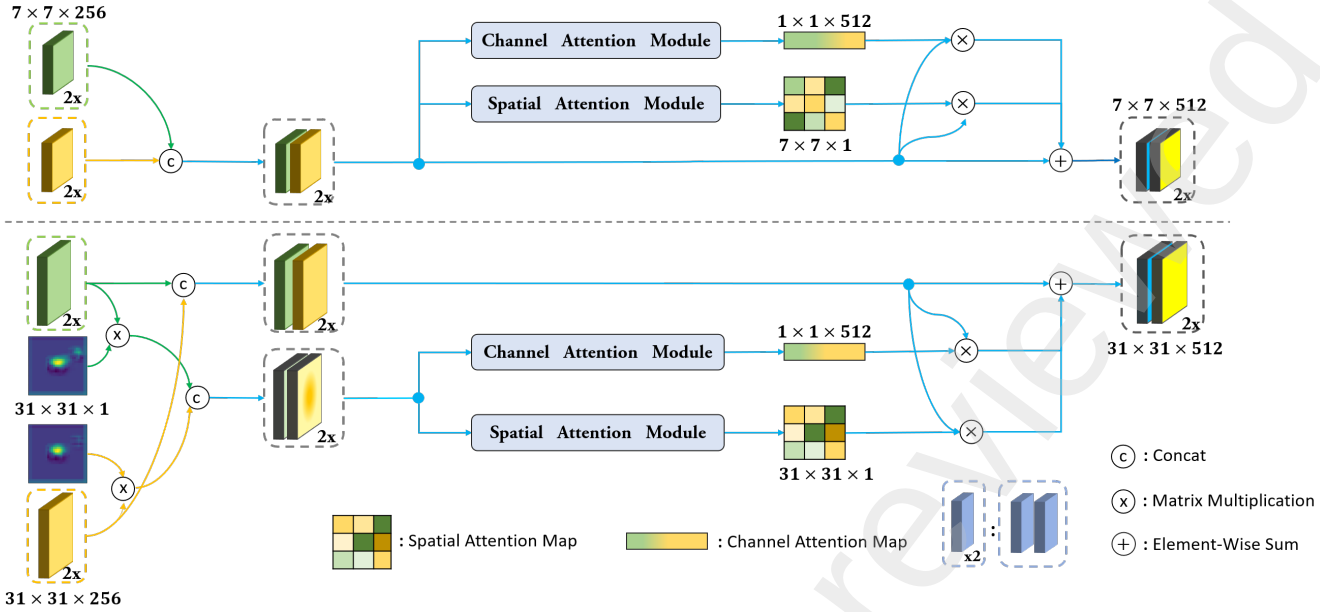

Figure 4: Illustration of dual-level fusion attention balance module, which uses the attention mechanism to allocate the weight ratio of two modal features. It can utilize the decision-level and feature-level information to balance the two modal features' weight ratio more reasonably.

a 2D spatial attention map $F, F \in 1 \times H \times W$. And F is calculated

$$\begin{aligned} F &= \sigma(Conv^{7\times7}([Avgpool(X); Maxpool(X)])) \\ &= \sigma(Conv^{7\times7}([f_{avg}; f_{max}])) \end{aligned} \tag{8}$$

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

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

$$S_X^Y = \alpha \cdot FY + Y, S_X^Y \in C \times H \times W \tag{9}$$

Where α is a scalar parameter.

Channel attention mechanism in DLBM. Input feature $X, X \in C \times H \times W$, 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:

$$\begin{aligned} F &= \sigma(FC(Avgpool(X)) + FC(Maxpool(X))) \\ &= \sigma(FC(f_{avg}) + FC(f_{max})) \end{aligned} \tag{10}$$

where FC represents the full connectivity layer. σ Represents sigmoid function. Then the channel attention feature is modulated by the input feature X to the feature Y as:

$$C_X^Y = \beta \cdot FY + Y, C_X^Y \in C \times H \times W \tag{11}$$

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 are described as:

$$F'_Z = C_{F_z}^{F_z} + S_{F_z}^{F_z} \tag{12a}$$

$$F'_X = C_{K_x}^{F_x} + S_{K_x}^{F'_x} \tag{12b}$$

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

3.4. GroundTruth And Loss

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

$$L = \lambda_1 L_{cls_v} + \lambda_2 L_{cls_i} + \lambda_3 L_{cls} + \lambda_4 L_{reg} \tag{13}$$

Among them, both classification head and regression head refer to SiamBAN. In details, the classification branch adopts cross entropy loss and elliptical classification label, while the regression branch adopts anchor-free method and IoU loss. Additionally, we set $\lambda_1 = 0.2, \lambda_2 = 0.2, \lambda_3 = 1, \lambda_4 = 1$ during trainings.

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, while the weight decay is set as 0.0001. We use 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 are initialized by the weights pre-trained on ImageNet [29]. 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.

Inference. We crop the template image from the first frame. For subsequent frames, we 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. Finally, 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], which generates two weight masks to update the classification results. Then we 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 [13] 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. Moreover, compared with similar tracking datasets, the classes are more abundant. Based on these advantages, GOT10K is very suitable for training tracking tasks.

LaSOT [7] 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. And each frame provides accurate manual annotation.

GTOT [17] 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 [18] 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 [24] contains 60 visible and infrared paired sequences, and these 60 sequences are a subset of RGBT234. For ease of use, we name the sequences without VOT-RGBT2020 as rgbt174.

LasHeR [20] 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. And it is the first large-scale data set in the two mode tracking challenge.

We adopt GOT10K and LaSOT to carry out 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. And 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 [14].

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 we set the PR threshold to 20. The detailed calculation of NPR refer to [25].

Table 1. Results on GTOT, including SiamRPN++, ATOM, DIMP, SiamFT, SGT, mfDIMP, MANet, SiamBAN, SiamBAN
(RGBT) and SiamDL . The red, blue, and green fonts represent the first three values.

Table 2. Results on VOT-RGBT2020, including SiamDL, SiamBAN(RGBT), and seven trackers from the VOT RGBT 2020 challenge [24]. The red, blue, and green fonts represent the first three values.

4.3. Comparison with State-of-the-art Trackers

GTOT: Table 1 and Figure 5 show the comparison results on GTOT with short sequences.Table 1 shows the percentage of PR/SR values of each tracker under each scene video sequence of the GTOT dataset.The comparison results include SiamRPN++[15], ATOM[5], DIMP[2], SiamFT, SGT[21], mfDIMP, MANet,
SiamBAN, SiamDL and SiamBAN (RGBT).Among them, SiamBAN (RGBT) is also the tracker implemented in this paper. After obtaining the visible light and infrared features through the ResNet50 network, the two features are directly combined by channel, and then the connected features are sent to the tracking classification regression head.Red, blue and green fonts represent the top three in that order.The PR of this algorithm tracker SiamDL is 0.888 and the SR is 0.731. Previously, the best performing tracker was MANet with PR of 0.894 and SR of 0.724.In contrast, SiamDL lags behind MANet in OCC (occlusion), FM (fast moving), and SO (small target) scenes, but performs better than MANet in LI (low illumination) and thermal crossover (TC) scenes.
Compared with the benchmark algorithm SiamBAN (RGBT) tracker, the proposed tracker achieves the PR of over 7.7% and the SR of over 4.4% in the GTOT full scene.Figure 5 shows the comparison results of PR/FPS and SR/FPS of each

Table 1
Results on GTOT

	SiamRPN++	ATOM	DIMP	SiamFT	SGT	mfDIMP	MANet	SiamBAN	SiamBAN (RGBT)	SiamDL
OCC	70.3/58.7	67.4/55.1	75.7/63.8	75.3/58.6	81.0/56.7	80.7/64.3	88.2/69.6	67.2/54.9	76.4/64.1	83.3/67.8
LSV	76.5/64.3	78.9/64.2	81.4/69.0	79.7/61.4	84.2/54.7	90.5/73.9	86.9/70.6	78.3/64.2	86.3/71.3	88.6/71.7
FM	75.9/65.9	74.8/63.0	78.9/68.0	72.1/60.1	79.9/55.9	81.3/68.7	87.9/69.4	74.3/62.0	80.2/68.5	84.8/70.6
LI	68.9/58.3	68.3/58.4	69.8/61.1	78.6/63.6	88.4/65.1	83.0/70.4	91.4/73.6	66.8/56.0	82.1/69.3	93.0/75.8
TC	76.6/64.0	79.0/63.3	84.2/68.7	76.0/59.3	84.8/61.5	80.4/65.2	88.9/70.2	76.3/61.0	72.7/62.0	86.1/70.9
DEF	71.0/59.3	69.1/58.8	69.9/59.9	72.5/61.9	91.9/73.3	80.7/67.1	92.3/75.2	66.1/55.5	80.9/67.3	91.2/73.8
SO	82.2/64.7	83.7/62.9	84.2/64.0	79.3/59.3	91.7/61.8	87.4/69.1	93.2/70.0	79.3/59.3	74.9/61.1	89.3/69.8
ALL	72.5/61.7	72.6/61.2	75.7/64.9	75.8/62.3	85.1/62.8	83.6/69.7	89.4/72.4	71.7/59.3	81.1/68.7	88.8/73.1

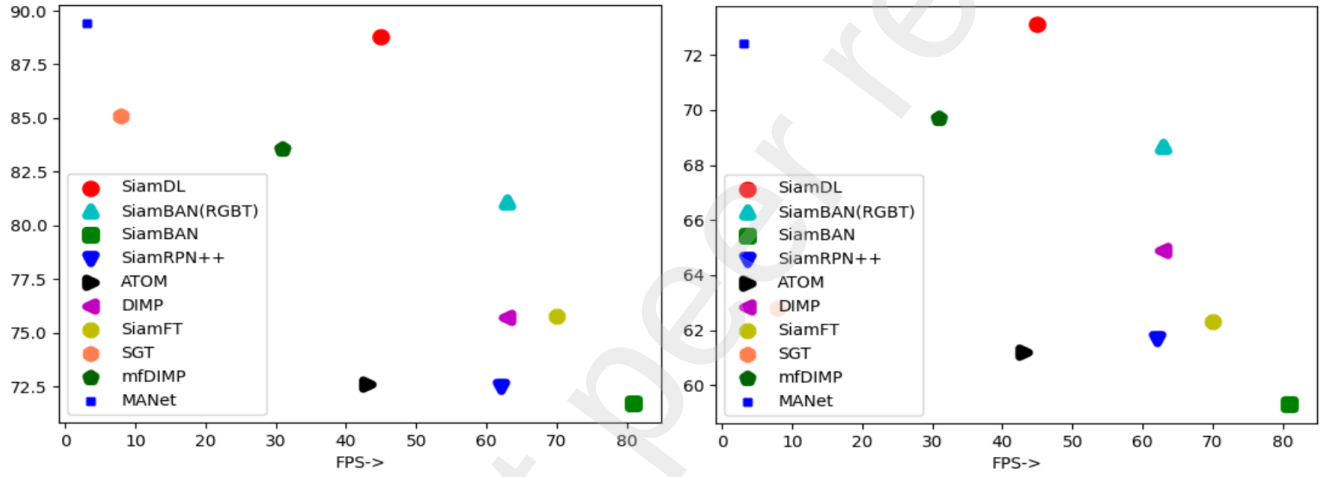

Figure 5: Speed comparison of various trackers on GTOT. The left figure compares PR and FPS, and the right figure compares SR and FPS.

tracker. It can be seen that SiamDL maintains high speed, and its performance is also close to SOTA.

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: Fig. 6 shows the comparison results on LasHeR with long sequences. While it is worth noting that only the test results of MANet and mfDIMP are published by other researchers [20]. And our tracker achieves 0.566 PR,0.437 SR and 0.521 NPR, which 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%.

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 SiamDL. CDSAM: cross domain siamese attention module. CLS: the classification heads for each modes. DLBM: dual-

Table 2
Results on VOT-RGBT2020

	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

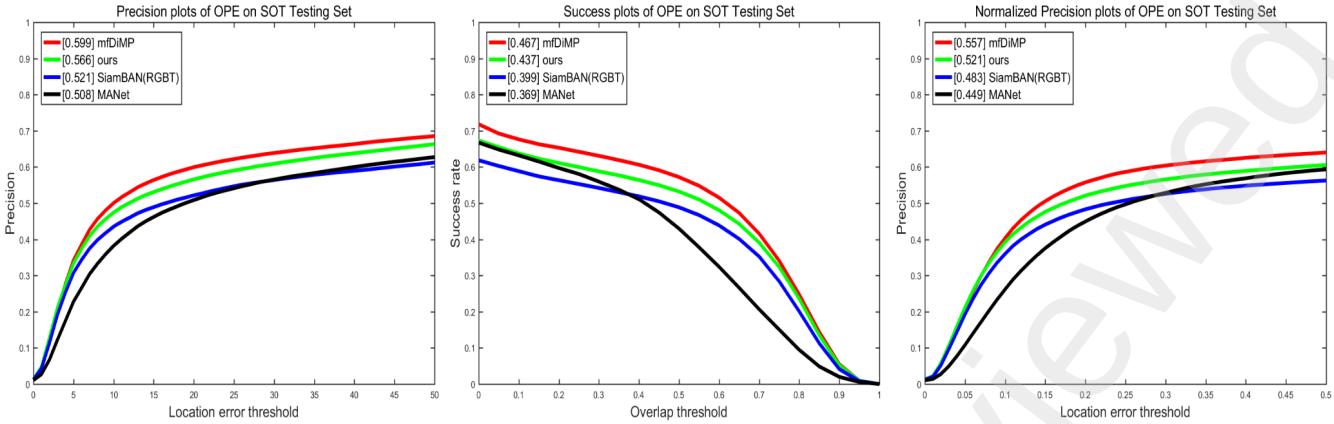

Figure 6: Results on LasHeR using precision (PR), success (SR) plots and normalizes precision (NPR) , including SiamDL, SiamBAN(RGBT), mfDIMP and MANet. Only the test results of MANet and mfDIMP are published in [20].

Table 3
Results of ablation experiments on model structures.

Method	PR	ΔPR	SR	ΔSR	NPR	ΔNPR
Baseline	0.521		0.399		0.483	
Baseline + CDSAM	0.540	+1.9%	0.416	+1.7%	0.496	+1.3%
Baseline + CDSAM + CLS	0.547	+2.6%	0.417	+1.8%	0.500	+1.7%
Baseline + CDSAM + CLS + DLBM	0.566	+4.5%	0.437	+3.8%	0.521	+3.8%

level balance module.

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.

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

4.4. Ablation Study

We study the impact of individual components in SiamDL, and conduct ablation study on LasHeR to test subset.

Model architecture. Table 3 shows the results of ablation experiments on model architecture. We use SiamBAN(RGBT) as baseline. By adding the cross domain siamese attention module, the indices PR, SR and NPR are improved from 0.521 to 0.540, 0.399 to 0.416 and 0.483 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 indices from 0.540 to 0.547, 0.416 to 0.417 and 0.496 to 0.500. 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 indices finally from 0.547 to 0.566, 0.417 to 0.437 and 0.500 to 0.521. And the final results of DLBM are 4.5 %, 3.8% and 3.8%, which are higher than these of 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.Among them, MDAM represents the multi-domain perception module, and DLBM represents the dual-level balance module. CDSAM stands for Cross-Domain Siamese Attention Mechanism, and CDSAM is an implementation of the attention mechanism in MDAM. DLFAM stands for Two-Level Fusion Attention Mechanism, which is an implementation method of the attention mechanism in DLBM. The implementation of reference [8] in the cross domain siamese attention mechanism and the implementation of reference [12] 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. As shown in Table 5, Feature-Level refers to the direct use of feature-level information to balance the weight ratio of modal features. Feature-Level + Decision-Level refers to the use of feature-level and decision-level information to balance the weight ratio of modal features. Different from [37, 9] and other studies, they directly use feature level information to balance the mode features'

Table 4
Results of ablation experiments on attention mode.

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

Table 5
Results of ablation experiments on source of balance module.

Method	PR	SR	NPR
Feature-Level	0.552	0.436	0.494
Feature-Level + Decision-Level	0.566	0.437	0.521

weight ratio. 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, as shown in Fig. 7, although using all layers of ResNet50 as the backbone has the highest performance, fps is only 20.

Qualitative analysis of ablation studies. Table 6 shows the percentage comparison results of PR/SR performance of each module in each scene sequence of GTOT dataset. Since the challenges faced in this chapter are illumination transformation, low-illumination scenes, and thermal crossovers, the qualitative analysis of ablation experiments in this subsection focuses on sequential low-lightness (LI) and thermal crossover (TC) scenarios accordingly.

The Multi-Domain Perception Module (MDAM) is added to the Baseline tracker to transfer the information in the temporal and modal domains to each other, which enhances the cross-domain representation of features, which makes the tracker accurate in low-light (LI) and thermal cross (TC) scenarios. The rate of PR has increased by 4% and 7.8%, and it can better cope with the small target (SO) scenario. The overall performance PR has increased by 3.5% and SR has increased by 2.2%.

The Baseline tracker adds a dual-level balance module (DLBM) to adaptively balance the weight distribution ratio after feature fusion using decision-level and feature-level information, which enhances the tracker's ability to adaptively select features, which makes the tracker in low illumination (LI) , the thermal crossover (TC) scene accuracy rate PR value has increased by 6% and 9.5%, the overall performance PR has increased by 4.4%, and the SR has increased by 2.8%. This shows that in the scene with complex illumination and temperature, the tracker is very effective for the

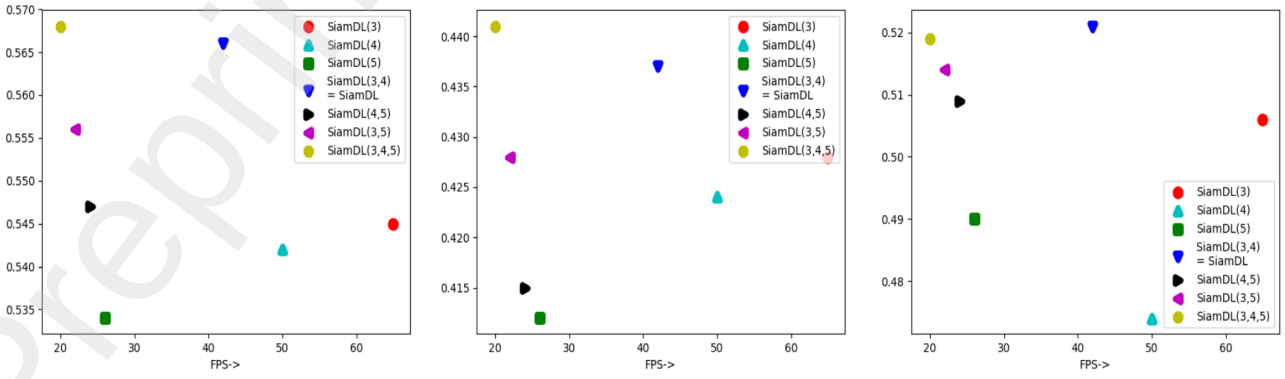

Figure 7: Results of ablation experiments on efficiency analysis. The left figure compares PR and FPS, the middle figure compares SR and FPS and the right figure compares NPR and FPS. The numbers 3, 4 and 5 represent that the output results of the ResNet50's corresponding layers are used to participate in the calculation of subsequent networks in SiamDL.

Table 6
Ablation experiment results of PR/SR qualitative analysis of each scene in the dataset GTOT

	OCC	LSV	FM	LI	TC	DEF	SO	ALL	FPS
Baseline	76.4/64.1	86.3/71.3	80.2/68.5	82.1/69.3	72.7/62.0	80.9/67.3	74.9/61.1	81.1/68.7	63
Baseline+ MDAM	78.0/64.6	83.8/68.9	78.1/68.3	86.1/71.4	80.5/68.0	87.1/71.6	82.1/65.3	84.6/70.9	54
Baseline+ DLBM	79.8/65.9	84.5/69.9	77.3/66.8	88.1/69.3	82.2/68.7	87.5/71.9	83.2/66.0	85.5/71.5	56
Baseline+ MDAM+DLBM =SiamDL	83.3/67.8	88.6/71.7	84.8/70.6	93.0/75.8	86.1/70.9	91.2/73.8	89.3/69.8	88.8/73.1	45

adaptive balance of fusion features.

Finally, the Baseline tracker adds a multi-domain perception module (MDAM) and a dual-level balance module (DLBM) at the same time to form the dual-modal tracking algorithm SiamDL based on feature-level and decision-level fusion attention proposed in this chapter, which makes the tracker in low illumination. (LI) and thermal crossover (TC) scenarios, the accuracy rate PR values are increased by 10.9% and 13.4%, the overall performance PR is increased by 7.7%, and the SR is increased by 4.4%. This shows that SiamDL maintains a high speed of 45FPS and has very robust performance when dealing with challenges such as illumination transformation, low-light scenes, and thermal crossover.

5. Conclusion

We design a siamese dual-level and multi-domain attention network for RGBT tracking. In details, cross domain siamese attention mechanism and dual-level fusion attention mechanism are introduced. And the former uses the rich context correlation of mode domain and time domain to improve the feature expression ability of network and adaptively update template features. While 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 trackers. We conduct several experiments on three data sets, and our tracker achieves state-of-the-art results and keeps high speed.In the future, we will optimize our network and design more concise fusion methods to achieve better performance. We will also try to use the network for related fields such as medical image fusion.

6. Acknowledgement

This work was supported by the China Postdoctoral Science Foundation (2021M691591).

References

[1] Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H., 2016. Fully-convolutional siamese networks for object tracking, in: European conference on computer vision, Springer. pp. 850–865.

[2] Bhat, G., Danelljan, M., Gool, L.V., Timofte, R., 2019. Learning discriminative model prediction for tracking, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 6182–6191.

[3] Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M., 2010. Visual object tracking using adaptive correlation filters, in: 2010 IEEE computer society conference on computer vision and pattern recognition, IEEE. pp. 2544–2550.

[4] Chen, Z., Zhong, B., Li, G., Zhang, S., Ji, R., 2020. Siamese box adaptive network for visual tracking, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6668–6677.

[5] Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M., 2019. Atom: Accurate tracking by overlap maximization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4660–4669.

[6] Danelljan, M., Bhat, G., Shahbaz Khan, F., Felsberg, M., 2017. Eco: Efficient convolution operators for tracking, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6638–6646.

[7] Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H., 2019. Lasot: A high-quality benchmark for large-scale single object tracking, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5374–5383.

[8] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3146–3154.

[9] 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.

[10] Guo, D., Wang, J., Cui, Y., Wang, Z., Chen, S., 2020. Siamcar: Siamese fully convolutional classification and regression for visual tracking, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6269–6277.

[11] Henriques, J.F., Caseiro, R., Martins, P., Batista, J., 2014. High-speed tracking with kernelized correlation filters. IEEE transactions on pattern analysis and machine intelligence 37, 583–596.

[12] Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141.

[13] Huang, L., Zhao, X., Huang, K., 2019. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence 43, 1562–1577.

[14] Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Kämäräinen, J.K., Danelljan, M., Zajc, L.Č., Lukežič, A., Drbohlav, O., et al., 2020. The eighth visual object tracking vot2020 challenge

results, in: European Conference on Computer Vision, Springer. pp. 547–601.

[15] Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J., 2019a. Siamrpn++: Evolution of siamese visual tracking with very deep networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4282–4291.

[16] Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X., 2018. High performance visual tracking with siamese region proposal network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8971–8980.

[17] Li, C., Cheng, H., Hu, S., Liu, X., Tang, J., Lin, L., 2016. Learning collaborative sparse representation for grayscale-thermal tracking. IEEE Transactions on Image Processing 25, 5743–5756.

[18] Li, C., Liang, X., Lu, Y., Zhao, N., Tang, J., 2019b. Rgb-t object tracking: Benchmark and baseline. Pattern Recognition 96, 106977.

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

[20] Li, C., Xue, W., Jia, Y., Qu, Z., Luo, B., Tang, J., Sun, D., 2021. Lasher: A large-scale high-diversity benchmark for rgbt tracking. IEEE Transactions on Image Processing 31, 392–404.

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

[22] Liu, X., Luo, Y., Yan, K., Chen, J., Lei, Z., 2022. Cmc2r: Cross-modal collaborative contextual representation for rgbt tracking. IET Image Processing 16, 1500–1510.

[23] 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 30, 5613–5625.

[24] M. Kristan, A. Leonardis, J.M.e.a., . Vot2020 challenge. www.votchallenge.net/vot20.

[25] Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B., 2018. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 300–317.

[26] Peng, J., Zhao, H., Hu, Z., Zhuang, Y., Wang, B., 2021. Siamese infrared and visible light fusion network for rgb-t tracking, in: Computer Vision and Pattern Recognition (IF).

[27] Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28.

[28] Shen, L., Wang, X., Liu, L., Hou, B., Jian, Y., Tang, J., Luo, B., 2022. Rgbt tracking based on cooperative low-rank graph model. Neurocomputing 492, 370–381.

[29] Song, Y., Ma, C., Wu, X., Gong, L., Bao, L., Zuo, W., Shen, C., Lau, R.W., Yang, M.H., 2018. Vital: Visual tracking via adversarial learning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8990–8999.

[30] Wang, Q., Teng, Z., Xing, J., Gao, J., Hu, W., Maybank, S., 2018. Learning attentions: residual attentional siamese network for high performance online visual tracking, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4854–4863.

[31] Wang, Y., Wei, X., Tang, X., Wu, J., Fang, J., 2022. Response map evaluation for rgbt tracking. Neural Computing and Applications 34, 5757–5769.

[32] Xu, Q., Mei, Y., Liu, J., Li, C., 2021. Multimodal cross-layer bilinear pooling for rgbt tracking. IEEE Transactions on Multimedia .

[33] Xu, Y., Wang, Z., Li, Z., Yuan, Y., Yu, G., 2020. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 12549–12556.

[34] Yu, Y., Xiong, Y., Huang, W., Scott, M.R., 2020. Deformable siamese attention networks for visual object tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6728–6737.

[35] Zhang, L., Gonzalez-Garcia, A., Weijer, J.v.d., Danelljan, M., Khan, F.S., 2019a. Learning the model update for siamese trackers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4010–4019.

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

[37] Zhang, T., Liu, X., Zhang, Q., Han, J., 2021b. Siamcda: Complementarity-and distractor-aware rgb-t tracking based on siamese network. IEEE Transactions on Circuits and Systems for Video Technology .

[38] Zhang, X., Ye, P., Leung, H., Gong, K., Xiao, G., 2020a. Object fusion tracking based on visible and infrared images: A comprehensive review. Information Fusion 63, 166–187.

[39] Zhang, X., Ye, P., Peng, S., Liu, J., Gong, K., Xiao, G., 2019b. Siamft: An rgb-infrared fusion tracking method via fully convolutional siamese networks. IEEE Access 7, 122122–122133.

[40] Zhang, X., Ye, P., Peng, S., Liu, J., Xiao, G., 2020b. Dsiammft: An rgb-t fusion tracking method via dynamic siamese networks using multi-layer feature fusion. Signal Processing: Image Communication 84, 115756.

[41] Zhang, Z., Peng, H., 2019. Deeper and wider siamese networks for real-time visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4591–4600.

[42] Zhang, Z., Peng, H., Fu, J., Li, B., Hu, W., 2020c. Ocean: Object-aware anchor-free tracking, in: European Conference on Computer Vision, Springer. pp. 771–787.

[43] Zhu, Y., Li, C., Tang, J., Luo, B., 2020. Quality-aware feature aggregation network for robust rgbt tracking. IEEE Transactions on Intelligent Vehicles 6, 121–130.