M^{5}L: Multi-Modal Multi-Margin Metric Learning for RGBT Tracking

Zhengzheng Tu, Chun Lin, Wei Zhao, Chenglong Li, and Jin Tang

***Abstract*— Classifying hard samples in the course of RGBT tracking is a quite challenging problem. Existing methods only focus on enlarging the boundary between positive and negative samples, but ignore the relations of multilevel hard samples, which are crucial for the robustness of hard sample classification. To handle this problem, we propose a novel Multi-Modal Multi-Margin Metric Learning framework named M^{5}L for RGBT tracking. In particular, we divided all samples into four parts including normal positive, normal negative, hard positive and hard negative ones, and aim to leverage their relations to improve the robustness of feature embeddings, e.g., normal positive samples are closer to the ground truth than hard positive ones. To this end, we design a multi-modal multi-margin structural loss to preserve the relations of multilevel hard samples in the training stage. In addition, we introduce an attention-based fusion module to achieve quality-aware integration of different source data. Extensive experiments on large-scale datasets testify that our framework clearly improves the tracking performance and performs favorably the state-of-the-art RGBT trackers.**

***Index Terms*— Deep metric learning, multiple modalities, feature fusion, hard samples, RGBT tracking.**

I. Introduction

VISUAL tracking, a fundamental task in computer vision, aims at locating the specific object with a changeable bounding box in the consecutive video frames. However, there are still many challenges to be solved, specifically when the tracked object receives tremendous influences from the environment. RGBT tracking takes advantages of different spectrum data that are visible images and thermal infrared images to allow the object being tracked in day and night, and thus receives more and more attentions in recent years [1]–[3].

Manuscript received March 31, 2021; revised September 19, 2021; accepted October 16, 2021. Date of publication November 16, 2021; date of current version November 30, 2021. This work was supported in part by the Natural Science Foundation of Anhui Higher Education Institution of China under Grant KJ2020A0033, in part by the Anhui Provincial Natural Science Foundation under Grant 2108085MF211, in part by the Joint Funds of the National Natural Science Foundation of China under Grant U20B2068, in part by the Anhui Provincial Key Research and Development Program under Grant 202104d07020008, and in part by the National Natural Science Foundation of China Key Project of International (Regional) Cooperation and Exchanges under Grant 61860206004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zhao Zhang. *(Corresponding author: Chenglong Li.)*

Zhengzheng Tu, Chun Lin, Wei Zhao, and Jin Tang are with the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: zhengzhengahu@163.com; lc20191001@163.com; 944258047@163.com; tangjin@ahu.edu.cn).

Chenglong Li is with the Information Materials and Intelligent Sensing Laboratory of Anhui Province, School of Artificial Intelligence, Anhui University, Hefei 230601, China (e-mail: lcl1314@foxmail.com).

Digital Object Identifier 10.1109/TIP.2021.3125504

As a multi-modal visual task, a pair of RGB and thermal infrared images are input into the RGBT tracking model, at last the output of RGBT tracking is the bounding box around the target in each frame of sequences. Li *et al.* [1] are the first to propose a comprehensive benchmark datasets to facilitate the research of RGBT tracking. Based on the benchmark [1], they propose a method based on the collaborative sparse representation in the Bayesian filtering framework [3]. In the similar tracking framework, Lan *et al.* [4] propose to optimize the modality weights for adaptive fusion of different modalities by a max-margin principle on the basis of classification scores. To improve the robustness of feature representations to background clutter in bounding boxes, Li *et al.* [5] propose a collaborative graph learning algorithm to construct a spatially-ordered weighted patch descriptor and perform object tracking via the structured support vector machine algorithm. Some works introduce deep learning techniques to boost RGBT tracking performance significantly. For example, Li *et al.* [6] present a multi-adapter convolutional neural network to learn the modality-shared, modality-specific and instance-aware target representations. Gao *et al.* [7] adopt an adaptive attention mechanism to fuse useful information from multiple modalities. However, the existing RGBT tracking methods only focus on the fusion of different modalities, but ignore the relations of multilevel hard samples, which are crucial for the robustness of hard sample classification.

There are several losses designed to consider the relations of samples. Triplet loss [8] is made up of three parts, an anchor, a positive sample and a negative sample. It aims to pull one positive sample closer to the anchor than the negative ones, only considering the relation between samples. N-pair-mc [9] investigates structural characteristics of multiple samples to construct the embedding function. It pulls one positive sample in positive class from $N-1$ negative samples in $N-1$ negative classes (one negative sample per class). However, it also ignores the information of rest positive and negative samples. Based on N-pair-mc, Lifted Struct [10] unites all negative samples to learn or optimize the embedding function. It aims to pull two positive samples chosen randomly as closer as possible and push all negative samples away from any of positive pairs with the distance larger than a margin. Actually, there are more diverse relations of samples, which needs to be explored, but not all of positive samples are utilized for exploring that. Ranked list loss [11] proposes to use all non-trivial samples to explore more relations of samples, which aims to rank all positive samples before negative ones. However, it neglects to mine the relations of multilevel hard samples. Therefore,

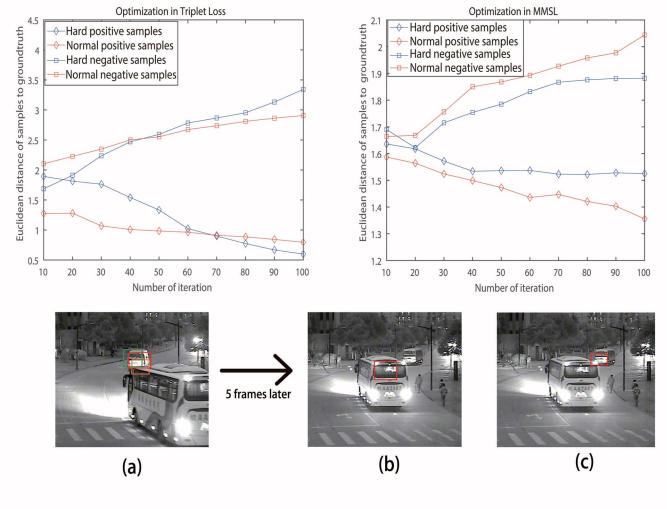


Fig. 1. Illustration of effectiveness of the proposed loss against conventional triplet loss. (a) represents the current frame, and the green, orange and red bounding boxes represent the normal sample, the hard sample and the tracking result respectively. (b) represents the tracking result using traditional triple loss, which has deviated from the true position of the object. (c) represents the tracking result using MMSL loss, which is closer to the true location of the object.

existing works usually design different loss functions to mine relations between positive and negative samples. However, the relations between hard samples and simple samples also greatly affect the performance. For example, Lin *et al.* [12] point out that simple samples have little effect on improving network performance, but due to their excessive number, eventually dominate the overall loss of the network. Similarly, hard samples greatly improve the performance of network, but the discriminative information of hard samples is hard to explored and exploited. However, existing networks almost use the relations between positive and negative samples but ignore the relations between multilevel hard samples. Therefore, the performance might not be so satisfied. In this work, we aim to make full use of the relations between samples, not only positive and negative samples but also simple and hard ones.

In this paper, in order to make full use of the relationship between samples, we divide all samples into four categories: normal positive, normal negative, hard positive and hard negative ones. As shown in Fig. 2, according to the Euclidean distance between each sample and the ground truth, we define four boundaries for separating all samples: the boundary $\alpha-\beta$, the boundary α, the boundary $\alpha+m$, the boundary $\alpha+m+\beta$. Herein, α, β and m are predefined parameters. In specific, the samples within the boundary $\alpha-\beta$ are defined as normal positive, the samples between boundary $\alpha-\beta$ and boundary α are hard positive, the samples beyond the boundary $\alpha+m+\beta$ are normal negative, and the samples between the boundary $\alpha+m$ and $\alpha+m+\beta$ are hard negative. In experiments, when $\alpha=1.6$, $\beta=0.1$, $m=0.2$, the network achieves best performance. We present a visualization to explain these types of samples clearer, as shown in Fig. 3. And we define their structural relations as follows. 1) Normal positive samples are closer to the ground truth than hard positive ones. 2) Hard negative samples are closer to the ground truth than normal negative ones. 3) Positive samples are almost closer to the ground truth than negative ones. These relations are crucial for the robustness of RGBT tracking, and we show their effectiveness in Fig. 1. Specifically, we predefine four margins to represent the structure of all samples. The first one is the normal-positive margin that constrains the distance between normal positive samples and the ground truth. The second one is the hard positive margin that makes normal positive samples closer to the ground truth than hard positive ones. The third one is the positive-negative margin that enlarges the boundary between all positive samples and all negative ones, and the last one is the hard-negative margin that makes hard negative samples closer than normal negative ones. The details are illustrated in Fig. 2.

In this paper, we propose a Multi-Modal Multi-Margin Metric Learning approach (called M^5L) for RGBT tracking. According to the definition of these four types of samples, in order to preserve the structural relations in learning process, we design a new multi-modal multi-margin structural loss(MMSL). We first mine hard samples in different modalities by computing the Euclidean distance between samples and the ground truth. Then we optimize the hard samples into the predefined regions determined by hard-positive margin and hard-negative margin respectively. The details are shown in Fig. 2. Moreover, we also use an improved triplet loss [13] to decrease the difference of two modalities. Fig. 1 shows the effectiveness of MMSL loss.

In addition, in the stage of feature extraction, we add a feature aggregation step to the first two layers of the feature extraction module, which can adaptively transfer complementary information between modalities and provide more discriminative information for the backbone network. Finally, we design an attention-based feature fusion to achieve adaptive fusion of different modalities. In specific, after extracting features of two modalities, we integrate them adaptively for more discriminative representations [14], which could make the fusion of different modalities adaptive for more discriminative representations. Compared with the methods proposed by Zhu *et al.* [14] and Gao *et al.* [7], our attention-based fusion module is simpler and more efficient. We just use two convolution layers to compute the weights of modalities, for reducing feature dimensions and computing weights of modalities respectively.

To our best knowledge, it is the first time to investigate the deep metric learning for RGBT tracking. At the same time, our multi-modal multi-margin structural loss focusing on the hard sample classification is proved to be effective for the improvement of RGBT tracking performance. We summarize the main contributions of this work as follows.

- We propose a novel deep metric learning framework, which exploits the structural information of hard samples to improve the robustness of hard sample classification in RGBT tracking.
- We propose a novel multi-modal multi-margin structural loss to preserve the structural relations of hard samples from both RGB and thermal modalities.

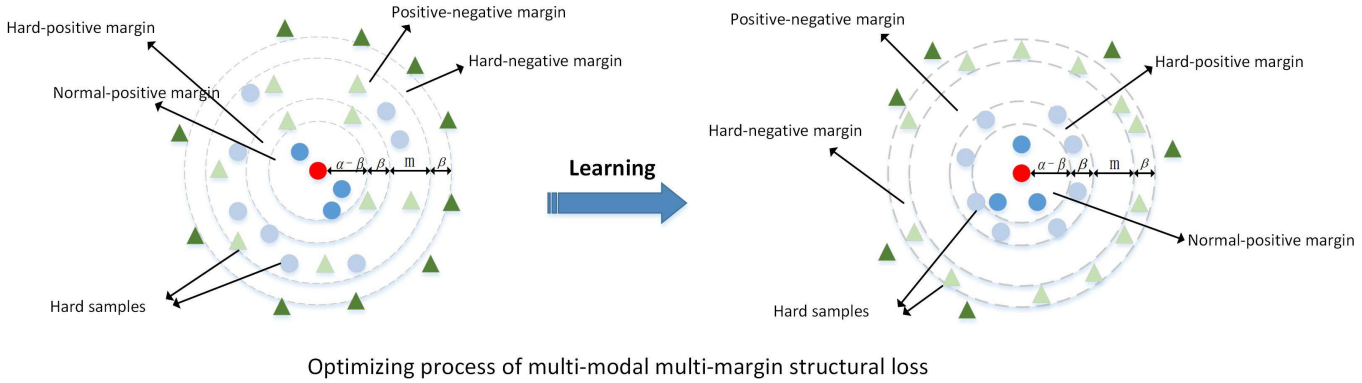


Fig. 2. Illustration of optimization process using the multi-modal multi-margin structural loss. The red circle is the ground truth, and the green triangles represent negative samples while the blue circles represent positive samples. It is worth mentioning that the lighter blue circles and lighter green triangles are hard samples which are hard to discriminate. The arrow denotes the gradient direction of optimization. The multi-modal multi-margin structural loss aims to make the hard positive samples into the region by hard-positive margin and hard negative samples into the region constrained by hard-negative margin. Meanwhile, all positive samples and negative samples are separated by positive-negative margin m, which utilizes their relations to construct a more robust feature embedding.

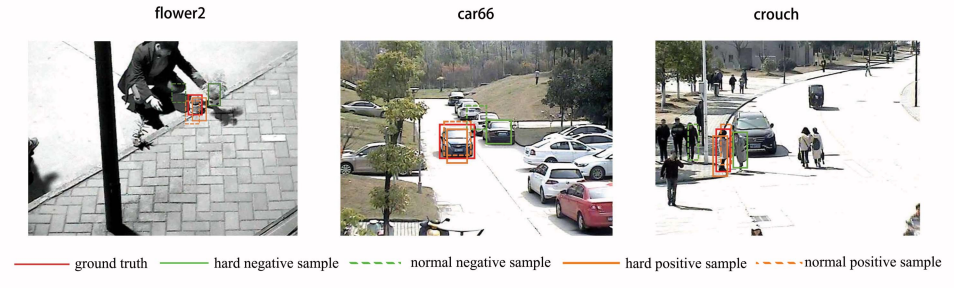


Fig. 3. Visualization of four types of samples in some sequences. The red box represents the ground truth, the green solid line represents the hard negative sample, the orange solid line represents the hard positive sample, the green dotted line represents the normal negative sample, and the orange dotted line represents the normal positive sample.

- We integrate an attention-based fusion module to achieve quality-aware fusion of different source data in an end-to-end trained deep learning framework.
- Extensive experiments on two large-scale datasets have demonstrated that our RGBT Tracking framework outperforms the state-of-the-art methods.

II. Related Work

RGBT tracking is a challenging topic in the field of computer vision. There has been a growing interest in the research of RGBT tracking in recent years. In this section, we mainly discuss the following relevant works including RGBT tracking and deep metric learning based visual tracking.

A. RGBT Tracking

With the help of thermal infrared modality, visual tracking can track the object better under difficult challenges like occlusion, background clutter and illumination variations. Some researches focus on the fusion of RGB and thermal modalities, a kind of fusion strategies are to learn the modality weights to obtain more robust feature fusion [15], [16]. Li *et al.* [1], [3] put forward a method based on the collaborative sparse representation to fuse multi-modal data in the Bayesian filtering framework. Lan *et al.* [4] propose to optimize the modality weights by the max-margin principle on the basis of classification scores. However, when classification scores of candidates or reconstructed residues become unreliable, the tracking performance will decrease.

There are some researches focusing on learning robust representation from a large number of multimodal data [5], [17]. Li *et al.* [5] propose a collaborative graph learning algorithm to construct a spatially-ordered weighted patch descriptor and perform object tracking via the structured support vector machine algorithm. To fuse the representations from different modalities and avoid introducing noises at the same time, Li *et al.* [17] also create a FusionNet to choose the most discriminative representations from the outputs of two stream ConvNet. However, these methods rely on handcrafted features instead of fusing multimodal data with predicting weights from different modalities. As the latest researches, Zhu *et al.* [18] present a recursive strategy (DAPNet) to fuse features of different layers, and Li *et al.* [6] propose a multi-adapter convolutional neural network (MANet) to learn the modality-shared, modality-specific and instance-aware target representations. Zhu *et al.* [19], [20] propose a novel method called CMRT which depends on a cross-modal manifold ranking algorithm to suppress background effect. Zhang *et al.* [21] propose three

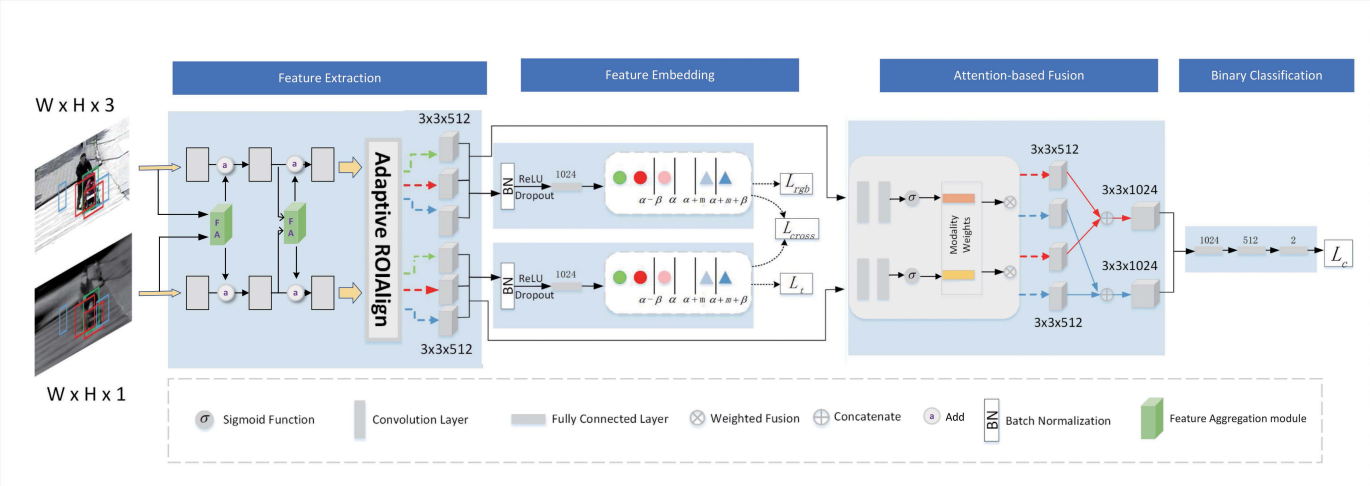


Fig. 4. Pipeline of our M^{5}L architecture which contains the feature extraction module, the feature embedding module, the attention-based fusion module and the binary classification module. The green, red and blue dotted lines denote the network streams of the anchor, positive and negative samples respectively. The green circle, red circles and blue triangles are corresponding to the ground truth, positive samples and negative samples respectively. The lighter red circle and lighter blue triangles denote the hard samples.

novel end-to-end fusion architectures, which consist of pixel-level fusion, response-level fusion, and feature level fusion.

There are also some traditional methods of RGBT tracking. For example, Wang *et al.* [22] diffuse instance patterns across RGBT data on spatial domain and temporal domain by presenting a novel cross-modal pattern-propagation (CMPP) tracking framework. Zhai *et al.* [23] propose to learn filters with the low-rank constraint and introduce an effective ADMM algorithm for optimization. Zhang *et al.* [24] propose a RGBT tracking method to fuse multilayers based on two dynamic Siamese networks. Kuai *et al.* [25] design a twofold Siamese network composed of a thermal branch and a RGB branch for RGBT tracking. Two response maps are derived from two branches respectively, and then fused according to their confidence degrees. A novel spatial-temporal regularized correlation filter is proposed by Feng *et al.* [26] to prevent the model degradation and the tracker drift caused by occlusion and background clutter. Zhang *et al.* [27] present a RGBT tracking framework combining modeling appearance and motion cues. Zhang *et al.* [27] also propose a tracker switcher to switch between the appearance and the motion trackers flexibly. Unlike these methods, we adopt an adaptive attention-based fusion module to achieve quality-aware fusion of different source data, meanwhile, multi-modal multi-margin structural loss is proposed to increase classification accuracy by exploiting relations of multilevel hard samples.

B. Deep Metric Learning-Based Visual Tracking

Based on a principled metric learning framework, Wang *et al.* [28] propose a discriminative model, which takes the visual matching and the appearance modeling into a single objective, achieving more persistent results. Based on the particle filter, Hu *et al.* [29] propose a deep metric learning tracker for robust tracking, which learns the nonlinear distance to classify the target and the background. In recent work of object tracking, a triplet loss [30] is added into Siamese network by training extracted deep feature maps. Compared with traditional pairwise loss, the triplet loss achieves more powerful and discriminative features. Li *et al.* [31] utilize the online distance metric learning to acquire codependent relationship of various feature dimensions. Their experiments show that online distance metric learning can improve the robustness of tracking. Unlike the above methods, we present a novel loss in deep metric learning by utilizing the relations of normal positive, normal negative, hard positive and hard negative samples to remain structural relations of hard samples for all modalities, and decrease the difference of two modalities by an improved triplet loss in RGBT tracking.

III. M^{5}L Framework

In this section, we will describe the details of our M^{5}L (Multi-Modal Multi-Margin Metric Learning) tracking framework, including overview of our M^{5}L, network architecture, training procedure and tracking details.

A. Overview of M^{5}L

We build our network based on RT-MDNet [32] as shown in Fig. 4. We first extract RGB and thermal features by the two-stream CNN architecture and the feature aggregation module. To exploit the relations of multilevel hard samples, we design a feature embedding module and propose a multi-modal multi-margin structural loss. Then, we propose an attention-based fusion module for achieving quality-aware integration of RGB and thermal data. At last, we use a binary classification module to discriminate the target from background.

B. Network Architecture

The overall network structure of our M^{5}L is illustrated in Fig. 4. M^{5}L mainly consists of four parts, including a feature extraction module, a feature embedding module, an attention-based fusion module and a binary classification module.

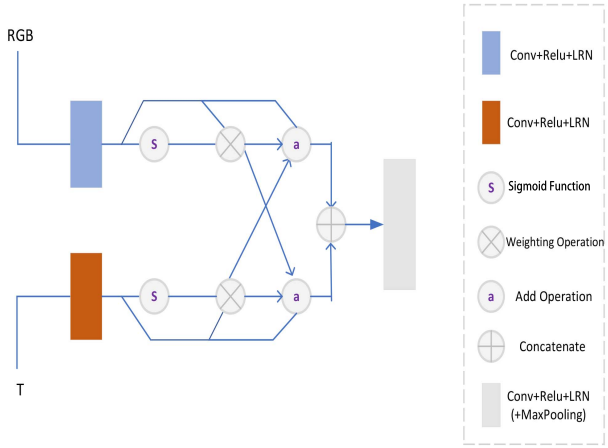


Fig. 5. Overall architecture of the feature aggregation(FA) module. In the first layer of feature extraction subnetwork, the maxpooling layer is added into FA module to reduce the dimension of features.

1) Feature Extraction Module: M^{5}L is for RGBT tracking task and adopts RT-MDNet [32] as the baseline. RT-MDNet uses VGG-M [33] as its backbone and several fully connected layers to capture global relations for robust classification. These fully connected layers are updated online to adapt appearance variations of object. If a more complex network is adopted as the backbone, the efficiency will drop significantly. As shown in Fig. 4, we use the first three layers of VGG-M network [33] (denoted as $Conv1-Conv3$) which is pre-trained on the ImageNet dataset [34]. Similar to RT-MDNet [32], we discard the max pooling layer in VGG-M to obtain the rich semantic information. We take the original RGBT image pairs as the input of network, and adopt the adaptive ROIAlign [35] layer in RT-MDNet [32] to improve the efficiency of feature extraction.

In addition, we add a feature aggregation module(FA) into the first two layers of the feature extraction subnetwork. The structure of FA is shown in Fig. 5. In different scenarios, RGB data and thermal data have different contributions to the network. For example, at night, thermal image can provide more discriminative information than RGB image, so we design the FA module to adaptively fuse the two modalities. Specifically, the FA module first uses the convolution layer to extract the features of different modalities, and then adopts the activation function to obtain the modal weight matrix, which will be multiplied by the modal features to obtain the model guidance feature map. Detailed operations are as follows:

$$g_{rgb} = Conv(f_{rgb}) * S(Conv(f_{rgb}))$$
$$g_t = Conv(f_t) * S(Conv(f_t)) \quad (1)$$
$$z = Conv((g_{rgb} + Conv(f_t)) \oplus (g_t + Conv(f_{rgb}))) \quad (2)$$

where S represents the sigmoid activation function. f_{rgb} and f_t are feature maps from RGB and thermal modalities. g_{rgb} and g_t represents the modal guidance feature maps, which will be used to adaptively strengthen the connection between the modalities, for conveying complementary information. $Conv$ is the convolution operation. Finally, we concatenate RGB features and thermal features, and use convolution and pooling operations to extract information to obtain the robust object representation.In Eq. (2), $+$ and $\oplus$ are addition operation and concatenation operation respectively. It is worth noting that we only add FA modules to the first two layers of the feature extraction subnetwork. As the network deepens, the number of feature channels will increase. Feature interaction will bring more parameters, so we only add the FA modules into the two layers.

2) Feature Embedding Module: To project all features of samples into the same space, we design a feature embedding module, which consists of fully connected (FC) layers and batch normalization (BN).

To handle the hard samples, we design a multi-modal multi-margin structural loss for accurate RGBT tracking. First, we get the ground truth bounding box, and take positive and negative samples in Gaussian space. After inputting all features of samples and the ground truth into the feature embedding module, we obtain their feature vectors. Then, by computing the Euclidean distance between each sample and the ground truth, we get the similarity between each sample and the ground truth. Next, we mine more hard samples as the training samples for our multi-modal multi-margin structural loss. Similar to ranked list loss [11], we mine hard samples referred to Eq. (3) and Eq. (4):

$$P^*_{r,i} = \left\{V_j | j \neq i, d^r_{ij} > \alpha\right\}$$
$$N^*_{r,i} = \left\{V_k | k \neq i, d^r_{ik} < \alpha + m\right\} \quad (3)$$
$$P^*_{t,i} = \left\{I_j | j \neq i, d^t_{ij} > \alpha\right\}$$
$$N^*_{t,i} = \left\{I_k | k \neq i, d^t_{ik} < \alpha + m\right\} \quad (4)$$

where r and t indicate the indexes of RGB and thermal modalities respectively. V and I are samples from RGB and thermal modalities. i, j, k are the indexes of ground truth, positive samples and negative samples respectively. In a mini-batch, we take V_i as the ground truth in RGB modality, so the positive set in RGB modality is referred to $P_{r,i} = \left\{V_j | j \neq i\right\}$, and similarly, $N_{r,i} = \{V_k | k \neq i\}$ is the negative set. And $P^*_{r,i}$ and $N^*_{r,i}$ are hard positive sample set and hard negative sample set in RGB modality respectively, herein $*$ means the hard sample set. d^r_{ij} is the Euclidean metric, and $d^r_{ij} = (f_r(V_j) - f_r(V_i))^2$ or $d^r_{ik} = (f_r(V_k) - f_r(V_i))^2$. Here, f_r is the distance metric function for feature embedding in RGB modality. α and m are the boundary of positive samples and the positive-negative margin between all positive and negative samples.

We expect positive samples are close to the ground truth as much as possible, while negative samples move against the trail of positive samples. Simultaneously, in order to guarantee the structural relations of hard samples, we divide all samples into four parts including normal positive, normal negative, hard positive and hard negative ones. Mathematically, we formulate the above objective as follows:

$$L(V_i, V_h; f_r) = y^r_{ih}(\left|\alpha - \beta - d^r_{ih}\right| + \left|\alpha - d^r_{ih}\right|)$$
$$+(1 - y^r_{ih})(\left|\alpha + m - d^r_{ih}\right| + \left|\alpha + m + \beta - d^r_{ih}\right|) \quad (5)$$
$$L(I_i, I_h; f_t) = y^t_{ih}(\left|\alpha - \beta - d^t_{ih}\right| + \left|\alpha - d^t_{ih}\right|)$$
$$+(1 - y^t_{ih})(\left|\alpha + m - d^t_{ih}\right| + \left|\alpha + m + \beta - d^t_{ih}\right|) \quad (6)$$

where f_r and f_t are the distance metric functions for feature embedding in different modalities. Here, h is the index of hard samples. V_i is the ground truth in RGB modality. $y^r_{ih} = 1$ if V_h is a hard positive sample in RGB modality. $y^r_{ih} = 0$ if V_h is a hard negative sample in RGB modality. From our point of view, positive samples and the ground truth belong to the same category, nevertheless negative samples are not. α, β, m are predefined parameters. If V_h is a hard positive sample, it will be optimized into the region $[\alpha - \beta, \alpha]$ by the hard-positive margin. If V_h is a hard negative sample, it will be optimized into the region $[\alpha + m, \alpha + m + \beta]$ by the hard-negative margin. Simultaneously, hard positive samples and hard negative samples are separated by the positive-negative margin. Therefore, the relations of multilevel hard samples are preserved by constructing a more informative multilevel feature embedding. The optimization process is shown in Fig. 2 and the equations are listed as follows:

$$L^r_P(V_i; f_r) = \frac{1}{\left|P^*_{r,i}\right|} \sum_{V_p \in P^*_{r,i}} L(V_i, V_p; f_r) \tag{7}$$

$$L^r_N(V_i; f_r) = \frac{1}{\left|N^*_{r,i}\right|} \sum_{V_n \in N^*_{r,i}} L(V_i, V_n; f_r) \tag{8}$$

$$L^t_P(I_i; f_t) = \frac{1}{\left|P^*_{t,i}\right|} \sum_{I_p \in P^*_{t,i}} L(I_i, I_p; f_t) \tag{9}$$

$$L^t_N(I_i; f_t) = \frac{1}{\left|N^*_{t,i}\right|} \sum_{I_n \in N^*_{t,i}} L(I_i, I_n; f_t) \tag{10}$$

where $P^*_{r,i}$ and $N^*_{r,i}$ are the hard sample sets in RGB modality which are decided by Eq. (3). V_p, V_n and I_p, I_n belong to the sets of hard positive and negative sample respectively in RGB and Thermal modalities.

Enlarging the distance between positive and negative samples in each modality is not enough, since the difference of RGB modality and thermal modality will influence the fusion of different modalities. Therefore, we further propose the cross-modality constraint between two modalities, which is based on the triplet loss [8];

$$\begin{aligned} L_{cross} = &\frac{1}{n} \sum \max\left\{d^r(P^+_r, A_t) - d^r(N^-_r, A_t) + \delta, 0\right\} \\ &+ \frac{1}{n} \sum \max\left\{d^t(P^+_t, A_r) - d^t(N^-_t, A_r) + \delta, 0\right\} \end{aligned} \tag{11}$$

where A_t and A_r are the ground truths from RGB modality and thermal modality respectively. $d^r(P^+_r, A_t)$ and $d^r(N^-_r, A_t)$ denote the Euclidean distance between one positive sample from RGB modality and the ground truth from thermal modality, and the Euclidean distance between one negative sample from RGB modality and ground truth from thermal modality, respectively. Likewise, $d^t(P^+_t, A_r)$ and $d^t(N^-_t, A_r)$ are the Euclidean distance between one positive sample from thermal modality and the ground truth from RGB modality, and the Euclidean distance between one negative sample from thermal modality and ground truth from RGB modality, respectively. δ is an enforced margin between a pair of positive and negative samples. n is the number of triplet pairs.

In multi-modal multi-margin structural loss (MMSL), we minimize objectives($L_{rgb}(V_i; f_r)$, $L_t(I_i; f_t)$, L_{cross}) equally and jointly optimize them:

$$L_{rgb}(V_i; f_r) = L^r_P(V_i; f_r) + L^r_N(V_i; f_r) \tag{12}$$

$$L_t(I_i; f_t) = L^t_P(I_i; f_t) + L^t_N(I_i; f_t) \tag{13}$$

$$L_{MMSL} = L_{rgb}(V_i; f_r) + L_t(I_i; f_t) + L_{cross} \tag{14}$$

where MMSL belongs to a part of the total loss, which plays a crucial role in training.

3) Attention-Based Fusion Module: Since properties of RGB and thermal modalities are different and their qualities are also changing, several methods compute and then fuse the weights of modalities for enhancing the performance of tracking. In the framework of Bayesian filtering, Li *et al.* [1] propose an adaptive fusion scheme based on collaborative sparse representation. Li *et al.* [3] also combine the modal reliability with the Laplacian sparse representation to realize the adaptive fusion of RGB and thermal infrared modalities. Lan *et al.* [36] present a feature representation and fusion model to fuse the object features from RGB and infrared modalities for object tracking. However, these methods may misclassify the object because of unreliable weights of modalities. Compared with the above traditional fusion methods, attention based fusion schemes are always very useful for deep learning based RGBT tracking. For example, Xu *et al.* [37] use the global average pooling and the global maximum pooling to realize the channel attention, and perform adaptive feature channel calibration on all convolutional layers. Gao *et al.* [7] use the convolution layer and the global average pooling to achieve channel attention, and use the adaptive channel weight to fuse RGB and thermal modalities. In addition, multiple convolutional layers and Softmax activation functions are introduced to achieve the spatial attention and enhance the localization ability of the network. In order to adaptively aggregate RGB and thermal modalities, Zhu *et al.* [38] use the global average pooling and three fully connected layers to achieve the channel attention, and generate a weight matrix to rebalance the contributions of the two modalities. However, in order to obtain the modal weight matrix, excessive parameters are introduced in above methods, greatly increasing calculation cost of the network. In this work, we introduce a channel attention-based fusion module in an end-to-end CNN framework to calculate the weight of each modality, which can improve the accuracy and has almost no effect on the tracking speed. Compared with the channel attention in [38], the structure of our attention-based fusion module is simpler. Specifically, our module only contains two convolutional layers, two ReLU activation functions and sigmoid activation functions. We take the convolution layer to extract the features of RGB and thermal modalities respectively, then obtain the weight matrix by the sigmoid activation function, and finally obtain more accurate feature representation of the two modalities by multiplying the weight matrix with the corresponding feature map.

As we know, attention mechanism makes the model focus on finding the significant information in input data related to the current output, so as to improve the quality of the output. In computer vision tasks, the attention mechanism always aims

at learning the feature mask by introducing new convolution layers with weights. The feature mask is corresponding to the regions of interest in the feature maps. Hence, the network can automatically focus on the regions of interest in the feature maps and highlight the target in the feature. In the course of RGBT tracking, there is one modality being relatively more beneficial to tracking at one moment. Inspired by the idea of attention mechanism, we design the new layers with parameters to learn the weights of modalities and make the network focus on a better modality. Therefore, attention mechanism in our method aims to amplify the superior modality and suppress the other modality over time. And the weights of modalities calculated by the attention mechanism are relatively more reliable.

To be specific, to explicitly measure the importance of features in different modalities, we introduce modality weights to fuse them adaptively, and all weights are collaboratively learned. After obtaining the modality weights, we aggregate feature maps of different modalities adaptively according to their weights. Different from adaptive feature recalibration which is used to compute weights in SENet [39], we utilize a simpler and more efficient module to compute modality weights. Specifically, this module includes two convolution layers, two ReLU activation functions and a sigmoid function. More details are presented in Fig. 4. We briefly use the following formulas to deliver the process of attention-based fusion module:

$$x = (\sigma(Conv(D_R)) \otimes D_R) \oplus (\sigma(Conv(D_T)) \otimes D_T) \quad (15)$$

where D_R and D_T are respectively the features of RGB and thermal modalities from feature extraction module, $Conv$ is the convolution layer, $\sigma(\cdot)$ is sigmoid function, $\otimes$ and $\oplus$ are weighting operation(that is element-wise multiplication) and concatenating operation respectively, and x is the concatenated RGBT feature. Fig. 6 demonstrates the efficiency of our attention-based fusion module. We compare our method with FANet [14]. In Fig. 6, the color in the heatmap represents the weight of the pixel in each modality, a larger response value in the heatmap(the color is close to orange) represents a larger weight in the modality. From Fig. 6, we can find the weights of modalities in our method are more accurate.

4) Binary Classification Module: In this module, we use three fully connected layers with ReLUs and dropouts for binary classification. Softmax cross-entropy loss is adopted as binary classification and we select the bounding boxes with higher classification scores as candidate bounding boxes for tracking. More training and tracking details can be found in next section.

C. Training Procedure

The whole network is end-to-end trained. We first take the images from RGB modality and thermal modality as the inputs of the network. To be specific, we extract 64 positive and 196 negative proposals from the Gaussian space, which are drawn according to the overlap ratios($IoUs$) with the ground-truth bounding box, where bounding boxes with $IoUs$ being larger than 0.7 are treated as positive samples and $IoUs$ of negative samples are less than 0.5. Then, we use two pre-trained VGG-M models to extract the features of first three layers of RGB and thermal images respectively, and adopt adaptive ROIAlign [35] mapping the ground truth and proposals to the feature maps of RGB and thermal modalities. With the feature embedding module, we can obtain the multi-modal multi-margin structural loss. At the same time, output from the attention-based fusion module, the fused features are then input into three fully connected layers to obtain the binary classification loss. Ultimately, the multi-modal multi-margin structural loss and softmax loss are combined together to jointly optimize the whole network.

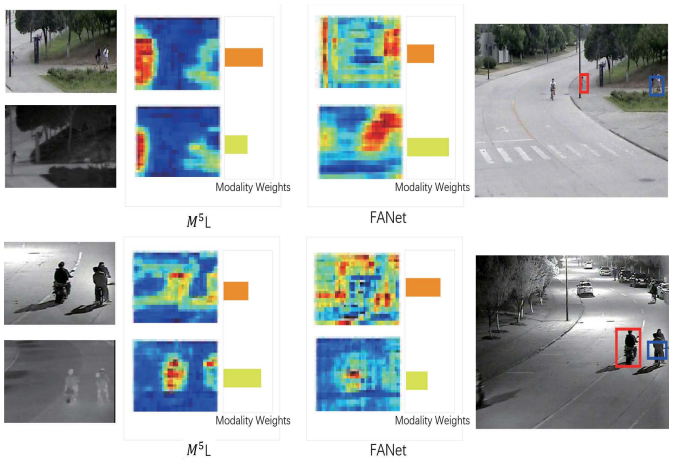


Fig. 6. Illustration of heatmaps of our method and FANet. The red and blue bounding boxes represent the tracking results of our method and FANet respectively. The orange rectangle bar and the green rectangle bar represent modality weights of RGB and thermal respectively. The larger the rectangle area is, the greater the weight of this modality is.

In the feature embedding module, we set α and β to 1.6 and 0.1 respectively. And the positive-negative margin(m) between positive and negative samples is set to 0.2. These parameters are validated with experiments. With respect to the network training, the learning rate is 0.0001 for convolutional layers and 0.001 for fully connected layers. And we use stochastic gradient descent(SGD) as the optimizer, and train our network for 300 epochs after loading the pre-trained model VGG-M. Finally, we alternatively train and test our M^{5}L on dataset GTOT [1] and RGBT234 [2].

D. Tracking Details

For the task of RGBT tracking, the object bounding box which contains the position and size of the object in the first frame is given in each video sequences. Given the tracking model trained offline, a typical process of RGBT tracking is to leverage the object state in the first frame to fine-tune the model and then locate the object in the next frame. To adapt appearance variations caused by some challenging factors such as illumination change, deformation and scale change, the tracking model is usually updated using reliable tracking results. Specifically, in the first frame, we draw a set of samples according to the initial position of object, and use these samples to fine-tune our model. In subsequent

frames, we use the model to locate the object. To adapt appearance variations of object over time, we update the model by the proposed structural loss using reliable tracking results in some frames. Similar to RT-MDNet [32], the network extracts 500 positive samples and 5000 negative samples around the object in first frame to train the network and update the weight parameters. Attention-based fusion module is used to integrate RGBT feature maps after extracting features of different modalities. Then these feature maps are sent to three fully connected layers, and it is worth mentioning the fully connected layers are responsible for test sequences, as the parameters of the fully connected layers will be updated according to different test sequence when the object is tracked online. Specifically, we update fully connected (FC) layers using the ground truth given in the first frame, or tracking results in subsequent frames to capture appearance dynamics of target object. FC layers consist of two domain-shared layers and one domain-specific layer (the last FC layer). In the tracking phase, given a new testing sequence, the domain-specific layer is re-initialized and all FC layers are updated online while the parameters of convolution layers are fixed. As same as MDNet does, we also adopt the principles of short-term and long-term [40] for update the parameters of three fully connection layers. Given the j-th frame, we update 256 candidates from a Gaussian distribution of last tracking result. For each candidate, we compute its positive and negative scores, and the target location in current frame is decided by the candidate with the maximum positive score. Eventually, we also adopt the technique of bounding box regression [40] to improve the location accuracy.

E. Discussion

Here, we will discuss the difference between ranked list loss [11] and our multi-modal multi-margin structural loss. In ranked list loss, instead of a certain positive data point, the whole positive sample set is expected to be close to the anchor as soon as possible, while the whole negative sample set moves in the opposite direction. However, it ignores structural relations of multilevel hard samples and also is not suitable for multi-modal samples. After we select hard samples, ranked list loss [11] optimizes all hard samples, but ignores the relations of multilevel hard positive and negative samples, which are crucial for robustness of hard sample classification. We divide all samples into four parts including normal positive, normal negative, hard positive and hard negative ones, which aims to use their relations to improve the robustness of feature embedding. Fig. 9 demonstrates the feature embedding visualization of samples for ranked list loss and our MMSL. Our MMSL also explores a cross-modality loss to minimize the difference of distribution of two modalities.

IV. Experiments

To verify the effectiveness of our Multi-Modal Multi-Margin Metric Learning for RGBT Tracking(M^5L), we conduct the experiments on two RGBT benchmark: GTOT [1] dataset and RGBT234 [2] dataset, and compare our $\mathbf{M^5L}$ method with many state-of-the-art methods.

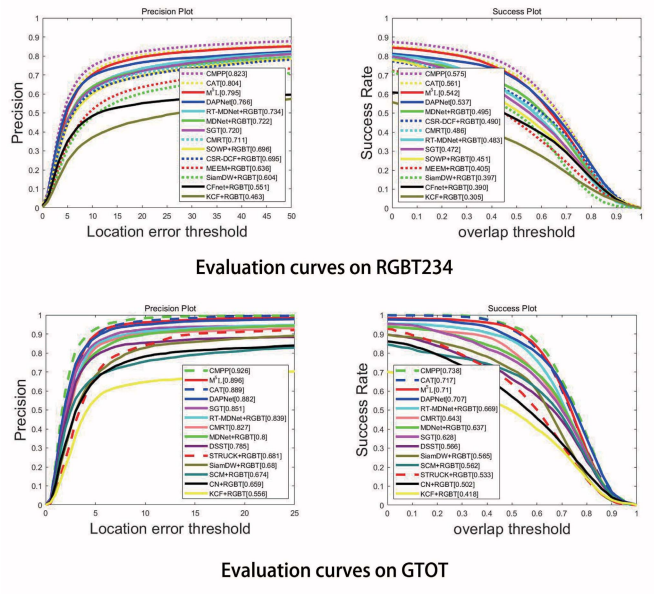


Fig. 7. PR/SR curves on RGBT234 and GTOT datasets respectively. Results on RGBT234 are on the top and GTOT on the bottom.

TABLE I

Results of Trackers w/o Thermal Modality Evaluated on GTOT. '+T' Means Adding Thermal Channel and '−T' Denotes Removing Thermal Channel

Trackers	+T			-T		
	PR	SR	Speed	PR	SR	Speed
(1) DAT	0.835	0.675	0.73fps	0.771	0.618	1.21fps
(2) SiamDW	0.680	0.565	**95.34fps**	0.688	0.550	**154.8fps**
(3) CN	0.659	0.502	52.72fps	0.469	0.36	100.58fps
(4) MDNet	0.800	0.637	0.89fps	0.812	0.633	1.61fps
(5) RT-MDNet	0.829	0.667	15.6fps	0.745	0.613	28.75fps
(6) M^5L	**0.896**	**0.710**	9.75fps	**0.870**	**0.679**	20.19fps

A. Evaluation Setting

1) Datasets: GTOT dataset consists of 50 sequences, and is annotated with 7 attributes. RGBT234 has a total of almost 234,000 frames and 234 video sequences with different lengths. RGBT234 is annotated with 12 attributes. We train all comparative methods on GTOT and test results on RGBT234, or train the methods on RGBT234 and test results on GTOT.

2) Evaluation Metrics: We use two following criteria for evaluation. Precision rate(PR) is the ratio of frames in which the location of tracked object is within the threshold distance of the groundtruth, and success rate(SR) is the percentage of frames in which the overlap between the output bounding box and the groundtruth bounding box exceeds the given threshold.

B. Evaluation on GTOT Dataset

1) Overall Performance: On GTOT dataset, we compare our method with existing 13 trackers, including CAT [41], CMPP [22], RT-MDNet [32]+RGBT, MDNet [40]+RGBT, SGT [5], KCF [42]+RGBT, SCM [43]+RGBT, DSST [44], CN [45]+RGBT, STRUCK [46]+RGBT, DAPNet [18], CMRT [20] and SiamDW+RGBT [47]. DSST [44], SGT [5], DAPNet [18], CMRT [20] are RGBT-based trackers and

TABLE II

RESULTS OF TRACKERS W/O THERMAL MODALITY EVALUATED ON RGBT234. '+T' MEANS ADDING THERMAL CHANNEL AND '−T' DENOTES REMOVING THERMAL CHANNEL

Trackers	+T			-T		
	PR	SR	Speed	PR	SR	Speed
(1) CSR-DCF	0.695	0.490	1.6fps	0.620	0.432	3.5fps
(2) SOWP	0.696	0.451	8.93fps	0.642	0.411	18.1fps
(3) SiamDW	0.604	0.397	**92.75fps**	0.590	0.392	**151.20fps**
(4) CFNet	0.551	0.390	32.42fps	0.521	0.386	65.72fps
(5) MDNet	0.722	0.495	0.82fps	0.707	0.490	1.27fps
(6) RT-MDNet	0.734	0.483	14.68fps	0.714	0.476	28.62fps
(7) M^{5}L	**0.795**	**0.542**	9.75fps	**0.744**	**0.523**	20.08fps

TABLE III

COMPARISON RESULTS OF M^{5}L WITH MDNET, DAPNET, RT-MDNET, CMRT AND SIAMDW. MDNET AND RT-MDNET ARE BOTH TRANSFORMED TO RGBT TRACKERS. THE BEST PR/SR SCORE ON EACH DATASET IS MARKED IN BOLD

Method	Publication	Date	RGBT234		GTOT		Speed
			PR	SR	PR	SR	
(a) RT-MDNet	ECCV	2018	0.734	0.483	0.829	0.667	15.6fps
(b) DAPNet	ACMMM	2019	0.766	0.537	0.882	0.707	2.11fps
(c) MDNet	CVPR	2016	0.722	0.483	0.8	0.637	0.89fps
(d) CMRT	ECCV	2018	0.711	0.486	0.827	0.643	8.30fps
(e) SiamDW	CVPR	2019	0.604	0.397	0.68	0.565	**95.34fps**
(f) M^{5}L	—-	—-	**0.795**	**0.542**	**0.896**	**0.710**	9.75fps

other methods are all RGB-based ones which are transformed to RGBT trackers. We concatenate features from RGB and thermal modalities into a single vector or view the thermal modality as an extra channel for the fair comparison. Fig. 7 shows that our method has a satisfied performance on GTOT, which is better than all trackers except CMPP and CAT. In the section IV-D.1, we will separately compare M^{5}L with CMPP and CAT in detail. In this section, we first compare the M^{5}L with other 11 trackers.

To be more specific, our method is 5.7%/4.1% higher than RT-MDNet+RGBT and 1.4%/0.3% higher than DAPNet in PR/SR. At the same time, M^{5}L has the advantage of tracking speed, which is much faster than DAPNet, as shown in Table III. More details can be seen in Table III.

We also compare our tracker with other RGB trackers for demonstrating the importance of thermal modality, including DAT [48], SiamDW [47], CN [45], MDNet [40] and RT-MDNet [32]. It can be seen from Table I that RGBT trackers have more advantages than RGB trackers and our method is superior to other popular trackers. RGBT tracking takes advantages of different spectrum data that are visible images and thermal infrared images to allow the object being tracked in day and night. In general, the models added with features from thermal channel have higher accuracy than models with one modality. However, the training time will be longer, the speed of tracking will also be slower after adding thermal channel, which is acceptable. Hence, considering the balance of performance and speed, it is worth conducting tracking on RGBT instead of RGB data.

C. Evaluation on RGBT234 Dataset

1) Overall Performance: To further verify the effectiveness of our approach, we evaluate our algorithm on RGBT234 dataset. The results of our approach compared with other 13 trackers (CAT [41], CMPP [22], MDNet [40]+RGBT, RT-MDNet [32]+RGBT, CSR-DCF [1]+RGBT, SGT [5], SOWP [49]+RGBT, MEEM [50]+RGBT, CFNet [51]+RGBT, KCF [42]+RGBT, DAPNet [18], CMRT [20] and SiamDW+RGBT [47]) are shown in Fig. 7, where SGT [5], DAPNet [18], CMRT [20] are RGBT-based trackers and other methods are RGB-based trackers which are transformed to RGBT trackers. From Fig. 7, we can see that our method is 6.1%/5.9% higher than RT-MDNet+RGBT in PR/SR, which, to a certain extent, demonstrating the effectiveness of the proposed modules. And, among all 13 trackers, our network's performance on PR/SR was second only to CMPP and CAT, again, in IV-D.1 we give a detailed comparison of these three networks. More comparisons with some trackers with most outstanding performances are presented in Table III.

We also compare our tracker with other RGB trackers including CSR-DCF [1], SOWP [49], SiamDW [47], CFNet [51], MDNet [40] and RT-MDNet [32]. More details can be seen form Table II. In general, the models added with features from thermal channel have higher accuracy than models with the single modality.

2) Challenges-Based Performance on RGBT234: There are 12 challenges in RGBT234, including background cluster (BC), camera motion (CM), deformation (DEF), fast motion (FM), heavy occlusion (HO), low illumination (LI), low resolution (LR), motion blur (MB), no occlusion (NO), partial occlusion (PO), scale variation (SV) and thermal crossover (TC). The results of comparison with other state-of-the-art trackers including RT-MDNet+RGBT, MDNet+RGBT, SOWP+RGBT, CSC-DCF+RGBT, MEEM+RGBT, CFNet+RGBT, DSST and KCF+RGBT, are listed in Table IV. The overall performance of our M^{5}L is better than our baseline method RT-MDNet.

In addition, our M^{5}L has achieved good performance for HO and PO, which proves that our attention-based fused features can boost the robustness of tracking in case of occlusion. Furthermore, BC and HO demonstrate that our feature embedding module can discriminate positive and negative samples and remain the relations of multilevel hard and normal samples, which greatly helps to classify target accurately, and improves the tracking performance. In brief, our network performs well for background cluster, camera motion, deformation, motion blur, and many other attributes, meaning that our framework has strong robustness.

3) Qualitative Performance: The comparison between our M^{5}L and other four trackers in term of qualitative performance has presented in Fig. 8. For example, in (b)and (c), our tracker performs best for partial occlusions and fast motion. The tracked object is occluded partially by other hard samples in (b). With the multi-modal multi-margin structural loss, our M^{5}L locates the object accurately. Bad illumination as shown in (a) and (d), our M^{5}L performs outstanding, demonstrating fused feature based on modality weight is more robust. In (b) and (c), a person is disturbed by surrounding pedestrians so that the general trackers could not discriminate which one is the target, similarly, when a football moves fast, football is similar to white crossings, which also disturbs the tracker. However, our M^{5}L aims at distinguishing positive and negative

TABLE IV

COMPARISON RESULTS OF PR/SR SCORES(%) OF DIFFERENT TRACKERS ON DIFFERENT CHALLENGES IN RGBT234

	RT-MDNet+RGBT	MDNet+RGBT	SOWP+RGBT	CSR-DCF+RGBT	SiamDW+RGBT	CFNet+RGBT	CMRT	DAPNet	M^{5}L
BC	0.725/0.455	0.644/0.432	0.647/0.419	0.618/0.410	0.519/0.323	0.463/0.308	0.631/0.398	0.717/**0.484**	**0.750**/0.477
CM	0.644/0.455	0.640/0.454	0.652/0.430	0.611/0.445	0.562/0.382	0.417/0.318	0.629/0.447	0.668/0.474	**0.752/0.529**
DEF	0.670/0.466	0.668/0.473	0.650/0.460	0.630/0.462	0.558/0.390	0.523/0.367	0.667/0.473	0.717/**0.578**	**0.736**/0.511
FM	0.637/0.387	0.586/0.363	0.703/0.435	0.529/0.358	0.597/0.365	0.454/0.299	0.613/0.384	0.670/0.443	**0.728/0.495**
HO	0.618/0.404	0.619/0.421	0.570/0.379	0.593/0.409	0.520/0.337	0.417/0.290	0.563/0.377	0.660/0.444	**0.665/0.450**
LI	0.737/0.474	0.670/0.455	0.723/0.468	0.691/0.474	0.600/0.399	0.523/0.369	0.742/0.498	0.775/0.530	**0.821/0.547**
LR	0.760/0.483	0.759/0.493	0.725/0.462	0.720/0.476	0.605/0.370	0.551/0.365	0.687/0.420	0.750/0.510	**0.823/0.535**
MB	0.612/0.429	0.654/0.463	0.639/0.421	0.580/0.425	0.494/0.340	0.357/0.271	0.600/0.427	0.653/0.467	**0.738/0.528**
NO	0.894/0.586	0.862/0.611	0.868/0.537	0.826/0.600	0.783/0.534	0.764/0.563	0.895/0.616	0.900/0.644	**0.931/0.646**
PO	0.780/0.517	0.761/0.518	0.747/0.484	0.737/0.522	0.608/0.396	0.597/0.417	0.777/0.536	0.817/0.544	**0.863/0.589**
SV	0.735/0.482	0.735/0.505	0.664/0.404	0.707/0.499	0.609/0.405	0.596/0.433	0.710/0.493	0.772/0.513	**0.796/0.542**
TC	0.786/0.513	0.756/0.517	0.701/0.442	0.668/0.462	0.569/0.368	0.457/0.327	0.675/0.443	0.768/0.538	**0.821/0.564**
ALL	0.734/0.483	0.722/0.495	0.696/0.451	0.695/0.490	0.604/0.397	0.551/0.390	0.711/0.486	0.766/0.537	**0.795 /0.542**

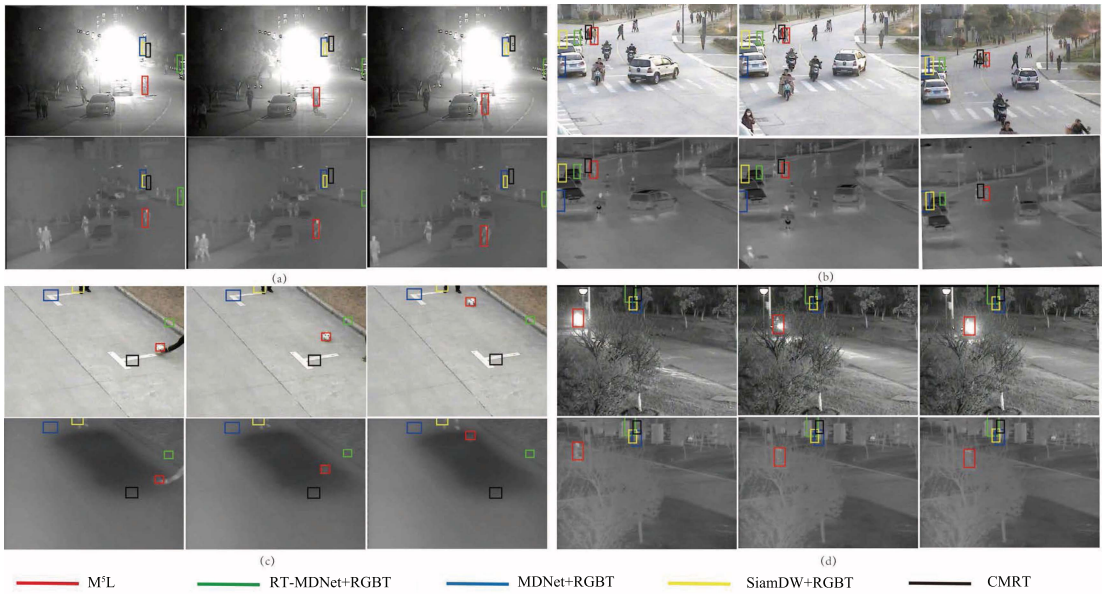


Fig. 8. Comparison of tracking results of our M^{5}L with other four trackers on four RGBT sequences. (a), (b), (c) and (d) are corresponding four challenges: high illumination, background cluster, fast motion and heavy occlusion, respectively.

samples and remaining the structural relations of samples including normal positive and negative samples, hard positive and negative ones, which enables the tracker locate the object.

D. Analysis of Our Approach

1) Comparison With the Relevant and Latest RGBT Trackers: We also make a comparison with the relevant and latest RGBT trackers. Compared with the CAT [41], our M^{5}L achieves comparable performance on GTOT dataset, and slightly lower performance on RGBT234 dataset, as shown as Fig. 7. M^{5}L has some advantages over CAT. CAT requires manual annotations for challenges to learn challenge-aware branches and needs three-stage training scheme. M^{5}L does not need challenge annotations and is end-to-end trained. In addition, compared with M^{5}L, CAT introduces many extra network branches and thus has more network parameters. From Fig. 7, we can see CMPP [22] outperforms M^{5}L on both GTOT and RGBT234 datasets. Although CMPP outperforms M^{5}L, our M^{5}L still has the following advantages. First, CMPP uses the inter-modal pattern propagation module to explore and utilize relationship between modalities, and uses the long-term context propagation module to transfer the information on time series. Although CMPP obtains robust representations, it also introduces a large amount of network parameters. In contrast, we use a two-layer feature aggregation module to transfer the complementary information between the two modalities, which can reduce the computational cost. Second, our M^{5}L is an end-to-end network, but CMPP requires a two-stage training procedure. Therefore, CMPP is much slower than M^{5}L, specifically, CMPP only reaches 1.5 FPS, but M^{5}L achieves 9.75 FPS.

2) Comparison With Siamese-Based Trackers: Both Siamese based trackers and the proposed one use the initial position of object in the first frame for tracking. The major difference between them is the scheme of locating the object. Siamese trackers use a large amount of training data to obtain a powerful tracking model, thus do not require updating the model in tracking process. Therefore, they often have a high tracking speed, but need a large amount of training data. Our tracker is based on RT-MDNet [32]. We need to update our model to adapt appearance changes of object, but our model can be trained easily and does not require too many training data. We also do comparison experiments with some

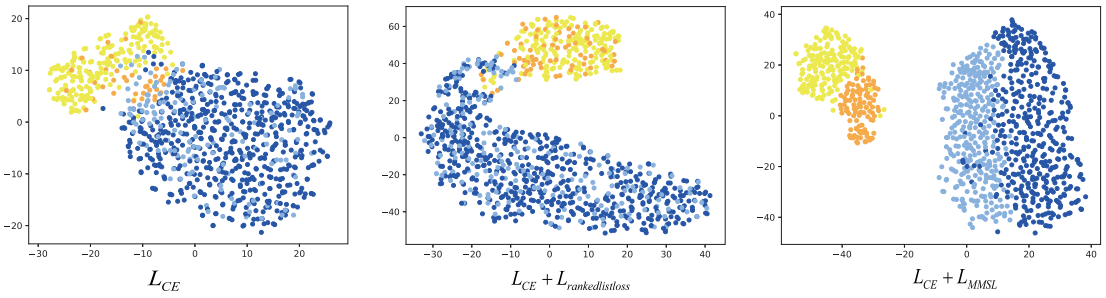


Fig. 9. Feature embedding visualization of samples. Blue points represent normal negative samples. The lighter blue ones are hard negative samples. Yellow points are normal positive samples and orange points represent hard positive samples.

TABLE V
EVALUATION ON DIFFERENT NUMBER OF SAMPLES WITH DIFFERENT DEEP METRIC LEARNING METHODS, WHICH ARE TRAINED ON GTOT AND TESTED ON RGBT234

Number of samples	Evalutions	RT-MDNet+Margin Loss	RT-MDNet+Ranked List Loss	M^{5}L
140	Training Time	**0.89h**	1.04h	1.34h
	PR(Testing Phase)	0.739	0.739	**0.763**
	SR(Testing Phase)	0.485	0.484	**0.524**
	Speed	**18.24fps**	16.98fps	11.32fps
200	Training Time	**1.11h**	1.16h	1.47h
	PR(Testing Phase)	0.742	0.755	**0.776**
	SR(Testing Phase)	0.493	0.496	**0.531**
	Speed	**17.28fps**	16.43fps	10.97fps
260	Training Time	**1.17h**	1.21h	1.89h
	PR(Testing Phase)	0.757	0.762	**0.795**
	SR(Testing Phase)	0.499	0.503	**0.542**
	Speed	**15.92fps**	15.36fps	9.75fps
320	Training Time	**1.24h**	1.37h	2.11h
	PR(Testing Phase)	0.755	0.756	**0.784**
	SR(Testing Phase)	0.503	0.505	**0.539**
	Speed	**14.64fps**	14.38fps	9.11fps

Siamese based RGBT trackers, including SiameseFC [52] and SiameseFT [53]. The results are shown in TableVI. From the results, we can see our tracker achieves 17.8%/ 10.6% higher in PR/SR over SiameseFC on RGBT234 dataset and 23.2%/ 16.5% higher on GTOT dataset, respectively. Compared with SiameseFT, our tracker obtain 10.7%/ 5.6% higher in PR/SR on RGBT234 dataset, respectively. On GTOT dataset, our tracker achieves 13.8%/ 8.7% higher in PR/SR, respectively. However, our network speed is 9.75fps, which is lower compared with SiameseFC's 25fps and SiameseFT's 28fps. This is due to M^{5}L takes some overlapping samples in the course of tracking, which costs more time. But our network outperforms the Siamese based RGBT trackers in overall performance.

3) Mining Hard Samples: As demonstrated in Section **Feature embedding module**, the multi-modal multi-margin structural loss(MMSL) mines the samples from a certain interval decided by α, β, m. Specifically, we first identify the hard samples as shown in Eq. (3) and pull hard positive samples into the region$[\alpha - \beta, \alpha]$ by hard-positive margin, referred in Eq. (5) and Eq. (6). Simultaneously, the optimization can pull the hard negative samples into the region$[\alpha + m, \alpha + m + \beta]$ by hard-negative margin, referred in Eq. (5) and Eq. (6). Hence, we conduct some experiments on RGBT234 dataset to analyze the effects of pre-defined parameters.

TABLE VI
COMPARISONS BETWEEN M^{5}L, SIAMESEFC AND SIAMESEFT

Method	RGBT234		GTOT		Speed
	PR	SR	PR	SR	
(a) $SiameseFC$	0.617	0.436	0.664	0.545	25fps
(b) $SiameseFT$	0.688	0.486	0.758	0.623	**28fps**
(c) M^5L	**0.795**	**0.542**	**0.896**	**0.710**	9.75fps

TABLE VII
INFLUENCE OF POSITIVE-NEGATIVE MARGIN m. PR/SR RESULTS(%) ARE LISTED, HERE $\alpha = 1.6$, $\beta = 0.1$

α=1.6, β=0.1	Precision Rate	Success Rate
m = 0	0.738	0.499
m = 0.2	**0.795**	**0.542**
m = 0.4	0.768	0.513

TABLE VIII

INFLUENCE OF β WHICH CONTROLS THE SIZE OF HARD-POSITIVE MARGIN AND HARD NEGATIVE MARGIN. PR/SR RESULTS(%) ARE SHOWN, HERE $\alpha = 1.6$, $m = 0.2$

α=1.6, m=0.2	Precision Rate	Success Rate
β = 0	0.765	0.519
β = 0.1	**0.795**	**0.542**
β = 0.2	0.770	0.517
β = 0.3	0.764	0.508

TABLE IX

COMPARISONS WITH OTHER METHODS OF CHOOSING HYPER PARAMETERS

Methods	Parameters	RGBT234		GTOT	
		PR	SR	PR	SR
Random Search	$\alpha = 1.7, \beta = 0.3, m = 0.2$	0.773	0.531	0.860	0.691
Grid Search	$\alpha = 1.6, \beta = 0.1, m = 0.2$	**0.795**	**0.542**	**0.896**	**0.710**

TABLE X

PERFORMANCE EVALUATION WITHOUT ATTENTION-BASED FUSION(AF)/ MULTI-MODAL MULTI-MARGIN STRUCTURAL LOSS(MMSL) ON DATASETS RGBT234 AND GTOT RESPECTIVELY

Method	RGBT234		GTOT	
	PR	SR	PR	SR
(a) M^5L	**0.795**	**0.542**	**0.896**	**0.710**
(b) $M^5L - ABFM$	0.787	0.530	0.871	0.679
(c) $M^5L - MMSL$	0.776	0.522	0.859	0.678
(d) $M^5L - FA$	0.785	0.540	0.870	0.692
(e) $RT - MDNet + RGBT$	0.734	0.483	0.829	0.667

4) Influence of Positive-Negative Margin m: First, we fix $\alpha = 1.6$ and $\beta = 0.1$. From Table VII, it is easily seen that when $m = 0.2$, our multi-modal multi-margin structural loss performs much better than other margins, i.e., 3% higher in PR/SR. Therefore, it is important for multi-modal multi-margin structural loss to select a proper interval. More details are presented in Table VII.

5) Influence of β: To study the effect of β, we fix $\alpha = 1.6$ and $m = 0.2$. From Table VIII, we verify our idea that we optimize the features of hard positive and negative samples into the regions by the fixed hard-positive margin amd hard-negative margin, aiming at remaining the structural relations of hard samples well to improve the performance of tracking.

We also compare the result with different hyper parameter selection methods. We can observe clearly from the Table IX that grid search selects the most suitable hyper parameters, which achieve the best results against another selection method.

6) Qualitative Evaluation: We make the qualitative evaluation at the level of inter-class and intra-class. Fig. 9 demonstrates the feature embedding visualization of samples with different loss functions. The cross entropy(CE) loss function produces two clusters including positive cluster and negative cluster, which shows the boundary is in mess and the hard samples are distributed everywhere in each cluster. Then we adopt CE loss and ranked list loss to enforce a margin between positive and negative samples, however, some hard samples are difficult to be distinguished, which reduces the classification accuracy for hard samples. We combine CE loss and MMSL to make hard samples more compact in the clusters and constrain a margin between positive and negative samples, which improves the robustness of classifying hard samples.

TABLE XI

RESULTS EVALUATED ON TWO RGBT DATASETS RGBT234 AND GTOT. WE MAKE COMPARISONS BETWEEN M^{5}L, M^{5}L W/O L_{cross}, M^{5}L W/O L_{rgb}, L_t AND THE BASELINE

Method	RGBT234		GTOT	
	PR	SR	PR	SR
(a) $RT - MDNet + RGBT$	0.734	0.483	0.829	0.667
(b) M^5L w/o L_{cross}	0.776	0.531	0.874	0.701
(c) M^5L w/o L_{rgb}, L_t	0.758	0.512	0.856	0.683
(d) M^5L	**0.795**	**0.542**	**0.896**	**0.710**

TABLE XII

RESULTS OF DIFFERENT TRACKERS COMBINED WITH MMSL

Method		PR	SR	Speed
MDNet	_	0.710	0.490	**1.27fps**
	+RGBT	0.722	0.495	0.82fps
	+RGBT+MMSL	**0.756**	**0.511**	0.65fps
VITAL	_	0.742	0.507	**1.36fps**
	+RGBT	0.758	0.514	0.95fps
	+RGBT+MMSL	**0.772**	**0.526**	0.73fps

7) Influence of Number of Samples: With more samples involved, it is expected that the performance will be improved to some extent, we have done the related experiments as shown as Table V. Specifically, we evaluate the training time, PR/SR and the tracking speed with different numbers of samples. We can see the PR and SR can be improved to a certain extent, meanwhile, the training time and the prediction time also increase with more samples involved. At last, we choose the number '260', which is a compromise among speed, precision and training time. We take RT-MDNet as our baseline and combine RT-MDNet with standard deep metric learning methods that are Margin Loss [54], Ranked List Loss [11] respectively. We can see more details from Table V.

8) Ablation Study: To justify main components of the proposed network, we evaluate the attention-based fusion module, the feature aggregation module and the feature embedding module on GTOT and RGBT234 datasets. We successively remove the attention-based fusion module (M^{5}L-ABFM), the feature embedding module (M^{5}L-MMSL) and the feature aggregation module(M^{5}L-FA). Table X presents the results on RGBT234 and GTOT respectively. The superior performance of M^{5}L versus M^{5}L-ABFM, M^{5}L-MMSL and our baseline RT-MDNet verifies the effectiveness of ABFM and MMSL for predicting modality attentions and separating positive and negative samples respectively. M^{5}L-FA verifies the effectiveness of the feature aggregation module to deliver complementary information. Table XI illustrates L_{cross} and L_{rgb}, L_t are all beneficial for RGBT tracking. To justify the effectiveness of

MMSL, we also combine our MMSL with other trackers and conduct the experiments on RGBT234 as shown as Table. XII. These results demonstrate the generalization of our MMSL is good.

9) Runtime Analysis: Ultimately, we analyze the runtime of our M^{5}L versus the baseline RT-MDNet and RT-MDNet+RGBT. Our implementation is on PyTorch0.4.0, python2.7 with NVIDIA GeForce GTX 1080Ti GPU and 4.2GHz Intel Core i7-8700k. The runtime of RT-MDNet, RT-MDNet+RGBT and M^{5}L on RGBT234 is 25fps, 15fps and 9fps respectively. However, their performances are far behind our M^{5}L, meanwhile, M^{5}L is much faster than MDNet.

V. Conclusion

In this paper, we have proposed a novel deep metric learning framework for RGBT tracking. We adopt an attention-based fusion strategy to adaptively aggregate features of different modalities, and also design an effective feature embedding module with a new deep metric loss called multi-modal multi-margin structural loss to distinguish hard samples and remain structural relations of samples including normal positive and negative samples, hard positive and negative ones in each modality, and an improved triplet loss [13] is further adopted to decrease the difference of two modalities, which could boost the tracking performance. Extensive experiments have suggested that our tracker achieves the superior tracking performance against other state-of-the-art trackers.

References

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

[2] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: Benchmark and baseline," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106977.

[3] C. Li, X. Sun, X. Wang, L. Zhang, and J. Tang, "Grayscale-thermal object tracking via multitask Laplacian sparse representation," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 4, pp. 673–681, Apr. 2017.

[4] X. Lan, M. Ye, and S. Zhang, "Robust collaborative discriminative learning for RGB-infrared tracking," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7008–7015.

[5] C. Li, N. Zhao, Y. Lu, C. Zhu, and J. Tang, "Weighted sparse representation regularized graph learning for RGB-T object tracking," in *Proc. ACM Int. Conf. Multimedia*, 2017, pp. 1856–1864.

[6] C. Li, A. Lu, A. Zheng, Z. Tu, and J. Tang, "Multi-adapter rgbt tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Oct. 2019, pp. 2262–2270.

[7] Y. Gao, C. Li, Y. Zhu, J. Tang, T. He, and F. Wang, "Deep adaptive fusion network for high performance RGBT tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Dec. 2019, pp. 1–9.

[8] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[9] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1857–1865.

[10] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4004–4012.

[11] X. Wang, Y. Hua, E. Kodirov, G. Hu, R. Garnier, and N. M. Robertson, "Ranked list loss for deep metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5207–5216.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[13] R. Chu, Y. Sun, Y. Li, Z. Liu, C. Zhang, and Y. Wei, "Vehicle re-identification with viewpoint-aware metric learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8282–8291.

[14] Y. Zhu, C. Li, B. Luo, and J. Tang, "FANet: Quality-aware feature aggregation network for robust RGB-T tracking," 2018, *arXiv:1811.09855*.

[15] L. Zhang, M. Danelljan, A. Gonzalez-Garcia, J. van de Weijer, and F. S. Khan, "Multi-modal fusion for end-to-end RGB-Ttracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Jul. 2019, pp. 1–10.

[16] A. Lu, C. Li, Y. Yan, J. Tang, and B. Luo, "RGBT tracking via multi-adapter network with hierarchical divergence loss," *IEEE Trans. Image Process.*, vol. 30, pp. 5613–5625, 2021.

[17] C. Li, X. Wu, N. Zhao, X. Cao, and J. Tang, "Fusing two-stream convolutional neural networks for RGB-T object tracking," *Neurocomputing*, vol. 281, pp. 78–85, Mar. 2018.

[18] Y. Zhu, C. Li, L. Bin, J. Tang, and X. Wang, "Dense feature aggregation and pruning for rgbt tracking," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 465–472.

[19] Li, Z. Xiang, J. Tang, B. Luo, and F. Wang, "RGBT tracking via noise-robust cross-modal ranking," *IEEE Trans. Neural Netw. Learn. Syst.*, 2021, doi: 10.1109/TNNLS.2021.3067107.

[20] C. Li, C. Zhu, Y. Huang, J. Tang, and L. Wang, "Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 808–823.

[21] L. Zhang, M. Danelljan, A. Gonzalez-Garcia, J. van de Weijer, and F. S. Khan, "Multi-modal fusion for end-to-end RGB-T tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–10.

[22] C. Wang *et al.*, "Cross-modal pattern-propagation for RGB-T tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 7064–7073.

[23] S. Zhai, P. Shao, X. Liang, and X. Wang, "Fast RGB-T tracking via cross-modal correlation filters," *Neurocomputing*, vol. 334, pp. 172–181, Mar. 2019.

[24] X. Zhang, P. Ye, S. Peng, J. Liu, and G. Xiao, "DSiamMFT: An RGB-T fusion tracking method via dynamic Siamese networks using multi-layer feature fusion," *Signal Process., Image Commun.*, vol. 84, May 2020, Art. no. 115756.

[25] Y. Kuai, D. Li, and Q. Qian, "Learning a twofold Siamese network for RGB-T object tracking," *J. Circuits, Syst. Comput.*, vol. 30, no. 5, Apr. 2021, Art. no. 2150089.

[26] M. Feng, K. Song, Y. Wang, J. Liu, and Y. Yan, "Learning discriminative update adaptive spatial-temporal regularized correlation filter for RGB-T tracking," *J. Vis. Commun. Image Represent.*, vol. 72, Oct. 2020, Art. no. 102881.

[27] P. Zhang, J. Zhao, D. Wang, H. Lu, and X. Yang, "Jointly modeling motion and appearance cues for robust RGB-T tracking," 2020, *arXiv:2007.02041*.

[28] X. Wang, G. Hua, and T. X. Han, "Discriminative tracking by metric learning," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 200–214.

[29] J. Hu, J. Lu, and Y.-P. Tan, "Deep metric learning for visual tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 11, pp. 2056–2068, Nov. 2016.

[30] X. Dong and J. Shen, "Triplet loss in Siamese network for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 459–474.

[31] X. Li, C. Shen, A. Dick, Z. Zhang, and Y. Zhuang, "Online metric-weighted linear representations for robust visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 931–950, May 2016.

[32] I. Jung, J. Son, M. Baek, and B. Han, "Real-time mdnet," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 83–88.

[33] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jun. 2009, pp. 248–255.

[35] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[36] X. Lan, M. Ye, S. Zhang, H. Zhou, and P. C. Yuen, "Modality-correlation-aware sparse representation for RGB-infrared object tracking," *Pattern Recognit. Lett.*, vol. 130, pp. 12–20, Jul. 2020.

[37] Q. Xu, Y. Mei, J. Liu, and C. Li, "Multimodal cross-layer bilinear pooling for RGBT tracking," *IEEE Trans. Multimedia*, early access, Jan. 28, 2021, doi: 10.1109/TMM.2021.3055362.

[38] Y. Zhu, C. Li, J. Tang, and B. Luo, "Quality-aware feature aggregation network for robust RGBT tracking," *IEEE Trans. Intell. Vehicles*, vol. 6, no. 1, pp. 121–130, Mar. 2021.

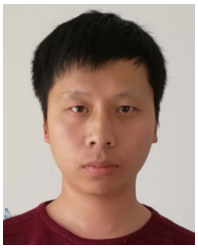
[39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[40] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4293–4302.
[41] C. Li, L. Liu, A. Lu, Q. Ji, and J. Tang, "Challenge-aware RGBT tracking," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 222–237.
[42] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.
[43] W. Zhong, H. Lu, and M.-H. Yang, "Robust object tracking via sparsity-based collaborative model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1838–1845.
[44] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–4.
[45] M. Danelljan, F. S. Khan, M. Felsberg, and J. Van De Weijer, "Adaptive color attributes for real-time visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2014, pp. 1090–1097.
[46] S. Hare *et al.*, "STRUCK: Structured output tracking with kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2096–2109, Oct. 2016.
[47] Z. Zhang and H. Peng, "Deeper and wider Siamese networks for real-time visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4591–4600.
[48] S. Pu, Y. Song, C. Ma, H. Zhang, and M.-H. Yang, "Deep attentive tracking via reciprocative learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1–5.
[49] H.-U. Kim, D.-Y. Lee, J.-Y. Sim, and C.-S. Kim, "SOWP: Spatially ordered and weighted patch descriptor for visual tracking," in *ICCV*, 2015, pp. 3011–3019.
[50] J. Zhang, S. Ma, and S. Sclaroff, "MEEM: Robust tracking via multiple experts using entropy minimization," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 188–203.
[51] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2805–2813.
[52] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional Siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 850–865.
[53] X. Zhang, P. Ye, S. Peng, J. Liu, K. Gong, and G. Xiao, "SiamFT: An RGB-infrared fusion tracking method via fully convolutional Siamese networks," *IEEE Access*, vol. 7, pp. 122122–122133, 2019.
[54] J. Deng, J. Guo, J. Yang, N. Xue, I. Cotsia, and S. P. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jun. 9, 2021, doi: 10.1109/TPAMI.2021.3087709.

Zhengzheng Tu received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2007 and 2015, respectively. She is currently an Associate Professor with the School of Computer Science and Technology, Anhui University. Her current research interests include computer vision and deep learning.

Chun Lin received the M.S. degree from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2021. His research interests include computer vision and deep learning.

Wei Zhao received the B.S. degree from Hefei University in 2019. He is currently pursuing the M.S. degree with the School of Computer Science and Technology, Anhui University, Hefei, China. His current research interests include computer vision and deep learning.



Chenglong Li received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2016, respectively. From 2014 to 2015, he worked as a Visiting Student with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He was a Postdoctoral Research Fellow with the National Laboratory of Pattern Recognition (NLPR), Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation, Chinese Academy of Sciences (CASIA), China. He is currently an Associate Professor with the School of Computer Science and Technology, Anhui University. His research interests include computer vision and deep learning.

Jin Tang received the B.Eng. degree in automation and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 1999 and 2007, respectively. He is currently a Professor with the School of Computer Science and Technology, Anhui University. His current research interests include computer vision, pattern recognition, machine learning, and deep learning.