Robust RGB-T Tracking via Graph Attention-Based Bilinear Pooling

Bin Kang, Dong Liang, Junxi Mei, Xiaoyang Tan, Quan Zhou, *Member, IEEE*, and Dengyin Zhang, *Member, IEEE*

***Abstract*— RGB-T tracker possesses strong capability of fusing two different yet complementary target observations, thus providing a promising solution to fulfill all-weather tracking in intelligent transportation systems. Existing convolutional neural network (CNN)-based RGB-T tracking methods often consider the multisource-oriented deep feature fusion from global viewpoint, but fail to yield satisfactory performance when the target pair only contains partially useful information. To solve this problem, we propose a four-stream oriented Siamese network (FS-Siamese) for RGB-T tracking. The key innovation of our network structure lies in that we formulate multidomain multilayer feature map fusion as a multiple graph learning problem, based on which we develop a graph attention-based bilinear pooling module to explore the partial feature interaction between the RGB and the thermal targets. This can effectively avoid uninformed image blocks disturbing feature embedding fusion. To enhance the efficiency of the proposed Siamese network structure, we propose to adopt meta-learning to incorporate category information in the updating of bilinear pooling results, which can online enforce the exemplar and current target appearance obtaining similar sematic representation. Extensive experiments on grayscale-thermal object tracking (GTOT) and RGBT234 datasets demonstrate that the proposed method outperforms the state-of-the-art methods for the task of RGB-T tracking.**

***Index Terms*— Bilinear pooling, meta-learning, RGB-T tracking, Siamese network.**

I. Introduction

With the flourishment of multimedia, thermal infrared camera has become economically affordable. Such camera can capture the thermal infrared radiation emitted by the targets with temperature above absolute zero, and hence is suitable for night surveillance. For this reason, two advantages have been identified in jointly using RGB and thermal infrared cameras.

1) Thermal infrared camera is skilled in resisting illumination change, which can offer strong support to RGB camera under poor light condition.
2) RGB camera would help solve the crossover challenge suffered in thermal infrared camera-based surveillance. Therefore, RGB-T tracking with both RGB and thermal features can effectively tackle the bad weather challenge [1].

Fig. 1. Challenging scenario in RGB234 dataset, where the car is seriously occluded. It is very difficult to discriminate the car from the background in thermal image. Besides, there also exists MB caused by RGB camera jitter. In this case, only a little useful yet partially matched information can be used for the complement of RGB and thermal target appearance. (a) RGB image. (b) Thermal image.

In RGB-T tracking, the RGB and thermal video sequences are obtained in pairs (see Fig. 1). The key idea is to exploit the complementarity of the RGB and thermal information for efficient multimodel fusion. To this end, many state-of-the-art methods have been developed over the past decade, which can be briefly categorized into three classes. The first class is the particle fusion-based RGB-T tracker, which requires effective representation of the appearance variation of the RGB and thermal targets for the estimation of particle fusion weights [2], [3]. The second one is to build multiple graph fusing model to effectively exploit the spatial relation between the RGB and the thermal target blocks [4], [5]. The third-class benefits from sparse representation, where the sparse codes and the correlation between two sparse representation models can be simultaneously estimated through solving the unified optimization problem [6]–[8]. All of aforementioned methods use handcraft feature for multimodel fusion. Compared with handcraft feature, deep convolutional neural networks (CNN) can extract the translation and light invariant deep semantic information for robust representation of the target. Thus, deep learning technology has appeared to have great potential in RGB-T tracking recently. For example, Zhu *et al.* [9]

Manuscript received November 20, 2020; revised July 21, 2021 and February 27, 2022; accepted March 16, 2022. This work was supported in part by the NSFC under Grant 62171232, Grant 61976115, Grant 61876093, and Grant 61872423; in part by the China Postdoctoral Science Foundation under Grant 2020M681684; and in part by the Jiangsu Postdoctoral Science Foundation under Grant 2021K278B. *(Bin Kang and Dong Liang contributed equally to this work.) (Corresponding author: Dong Liang.)*

Bin Kang and Dengyin Zhang were with the Jiangsu Key Laboratory of Broadband Wireless Communication and Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China. They are now with the School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China.

Dong Liang and Xiaoyang Tan are with the Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: liangdong@nuaa.edu.cn).

Junxi Mei and Quan Zhou are with the Department of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3161969.

Digital Object Identifier 10.1109/TNNLS.2022.3161969

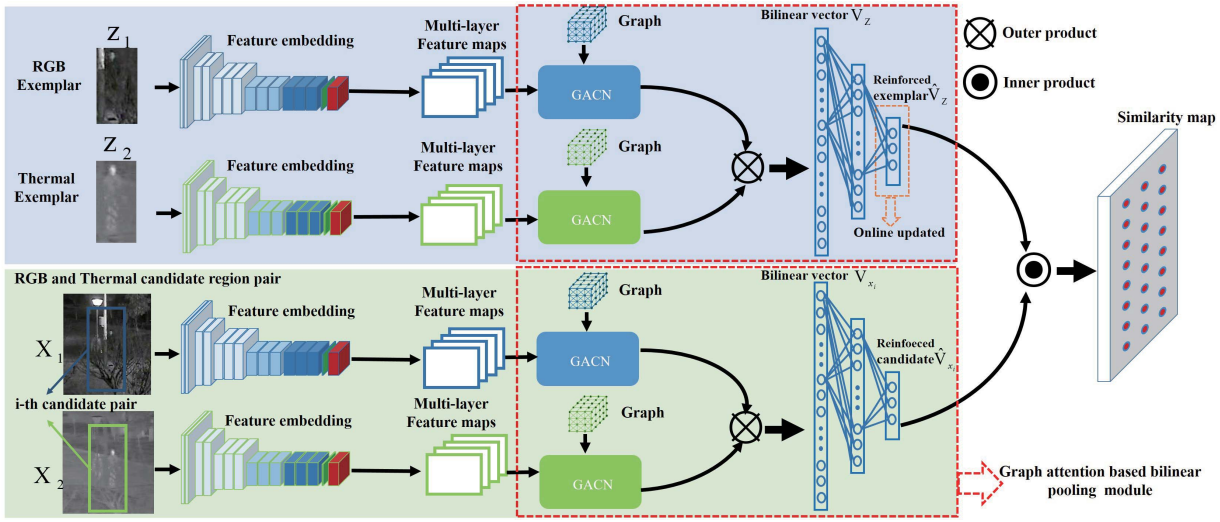


Fig. 2. Pipeline of our FS-Siamese, which is consisted of three components: 1) feature embedding; 2) graph attention-based bilinear pooling module for generating reinforced exemplar and candidate; 3) inner product for calculating the similarity between reinforced exemplar and each reinforced candidate within the search region. There have three scales for the online appearance changes. The first innovation of FS-Siamese lies in that we formulate a multiple graph learning problem to integrate GACN, outer product and fully convolutional layers into a unified end-to-end network structure. The second innovation is to design a meta-learning strategy for online updating reinforced exemplar.

proposed a dense CNN for RGB-T tracking, which can recursively aggregate informative features of two kinds of convolutional paths (RGB image-oriented convolutional path and thermal image-oriented convolutional path). In [10], a multiadapter convolutional network is proposed to simultaneously explore the complementarity property and achieve instance-aware feature learning in an end-to-end manner.

Existing CNN-based RGB-T trackers often consider the multilayer convolutional feature maps as the hierarchically holistic feature, ignoring the partial feature interaction between the RGB and thermal targets. This may obviously reduce tracking accuracy in challenging video pairs such as in Fig. 1, where only a little useful information can be extracted from RGB or thermal video sequences for representing pairwise targets. What's worse, the little useful information on the RGB and thermal targets may be partially matched or even unmatched in spatial domain. In this case, simply treating multiple deep features as holistic feature for multimodel fusion may result in inevitable negative effect. Part-feature-based RGB-T trackers such as [4], [5] can achieve partial information fusion according to the importance of different image blocks. However, those methods pay attention to the handcraft feature only, which could not be easily extended to multiple convolutional network fusion.

In this article, we propose a simple yet efficient four-stream oriented Siamese network (FS-Siamese) for RGB-T tracking as shown in Fig. 2, where the feature embedding of four streams can be divided into exemplar embedding pair and candidate embedding pair. Two embedding pairs can be fused, respectively, through the graph attention-based bilinear pooling module for generating the reinforced exemplar and reinforced candidate, which are used to produce the subsequent similarity map. Bilinear pooling has shown superior performance over traditional linear fusion strategy on the fusion of heterogeneous partial information in fine-grained recognition [11] and visual question answering [12]. Although bilinear pooling has won a certain performance gain, it could not discriminate the importance of the elements in the deep feature maps. This may give rise to unavoidable negative effect when facing challenging scenarios such as in Fig. 1. In view of these observations, we introduce coattention mechanism in the bilinear pooling to formulate multimodel pooling as a multiple graph learning problem. Based on the new problem formulation, we develop a graph attention-based bilinear pooling module to integrate two tasks, namely the exploration of partial feature interaction and the fusion of multisource oriented feature embeddings, into a unified end-to-end network structure.

Since the target appearance may dramatically change, it is necessary to introduce an effective strategy in updating the graph attention-based bilinear pooling module. The state-of-the-art updating strategies such as [13]–[15] only focus on exploring the temporal correlation between the current and previous target appearance, while ignoring a fact that online exploring the spatial relation between the target and its surrounding background is very important for locating the most similar candidate pairs. Considering this issue, we design a meta-learning-based updating strategy to effectively update the fully connected layer of the graph attention-based bilinear pooling module. This paves a way on utilizing category information to online update semantic representation of exemplar. The main contributions of this article are listed as follows.

1) We formulate the attention-based bilinear pooling as a multiple graph learning problem, based on which we integrate the graph attention network and outer product into a unified structure to highlight the discriminative local information in RGB and thermal targets. This can effectively eliminate the disturbance in target pair fusion.
2) Traditional multiple stream-oriented tracking networks only fuse the target regression results of different streams, without exploring the pairwise relation during the fusion of target embeddings. To overcome this limitation, we propose a four-stream oriented network structure using graph attention-based bilinear pooling for the effective fusion of multisource embedding pairs.

3) We adopt meta-learning to update the graph attention-based bilinear pooling, and thereby utilize the category information to online restrict the exemplar to learn a similar semantic representation as current tracking result, which is helpful for discriminating the reinforced exemplar and reinforced candidates.
4) Extensive experiments on grayscale-thermal object tracking (GTOT), RGBT234, CUB-200-2011, fine-grained visual classification (FGVC)-aircraft, and Cars datasets show that our graph attention-based bilinear pooling module not only can effectively fuse multidomain multilayer feature maps in RGB-T tracking, but also can be extended to other multimodel fusion tasks.

II. Related Works

A. Siamese Network in RGB Tracking

Siamese network has popular in RGB camera-based visual tracking due to its simple network structure and fast tracking speed. In Siamese network-based RGB tracking, Bertinetto *et al.* [16] is the pioneer who designs the Siamese network structure, where the tracking result is obtained by orderly calculating the similarity between the exemplar embedding and each candidate embeddings within the search region. The cross correlation is often used as the similarity measure. Following Bertinetto's work, the following studies have emerged which can be briefly divided into three scenarios.

1) The attention-based Siamese networks (e.g., [13], [17]) that effectively use the gradient of backward propagation and the channel attention mechanism to make the target appearance embedding concentrate on the informative subregion.
2) The local pattern-based Siamese Networks (e.g., [18]–[20]) that can explore the spatial relation between different target blocks.
3) The region proposal network (RPN)-based Siamese networks (e.g., [21]–[23]) that introduce RPN in Siamese network to avoid the time-consuming multiscale estimation step.

Unfortunately, all of aforementioned works cannot be easily extended to RGB-T tracking because of the following two main challenges.

1) The state-of-the-art RGB trackers have explored relation among different target blocks and introduced attention mechanism in the Siamese network, however those works are carried out in RGB domain. It is seen in Fig. 1 that there is a big gap between RGB and thermal images. Thus it not only requires the RGB-T tracker to explore the spatial correlation between target blocks in a single image domain, but also requires to overcome image gap challenge to effectively locate informative target blocks for exploring the partial feature interaction between two image domains.
2) We should not sacrifice the simple Siamese structure for exploring the partial feature interaction between RGB and thermal targets. Hence it requires to find a tradeoff between the complexity of the multimodel fusion and the efficiency of the network structure.

Similar to our network structure, the work in [24] introduced Siamese network in RGB-T tracking. However it still adopts linear fusion strategy to make RGB and thermal target feature complement with each other, ignoring theoretically study how the effective extract the common and the specific information between two image domains for unlinear fusion. Although the RGB tracker in [25] also has four embedding paths, it only fuses the tracking results of different streams, while ignoring the pairwise relation during feature embedding.

B. Bilinear Pooling

After the work in [26] that used multimodal compact bilinear pooling to explore the pairwise relation between two heterogeneous models, bilinear pooling has become an effective tool in visual question answering (VQA). Since the dimension of the output bilinear vector in [26] is high, Kim *et al.* [27] proposed a low-rank bilinear pooling to use two online estimated projection to project bilinear vector into a low-rank subspace, in which the redundant information in bilinear vector can be obviously reduced. Besides VQA, bilinear pooling has also been widely used in face recognition and fine-grained recognition, e.g., Chang *et al.* [28] proposed a Compound Rank-k Projections (CRP) algorithm for bilinear analysis, where the 2-D handcraft feature-based discriminant projections can be simultaneously learned in a collaborative way. Lin *et al.* [11] proposed a bilinear CNN model to use outer product to effectively fuse the pairwise fine-grained target information between two kinds of CNN networks. Wei *et al.* [29] used bilinear pooling to explore the partial feature interaction between two fine-grained models.

Unlike aforementioned works, we introduce graph attention CNN in bilinear pooling to simultaneously locate the informative target blocks in RGB-thermal image pairs. Although Gao *et al.* [30] has also used graph CNN in visual tracking, this work use predefined affinity matrix to build the graph. Our work is more challenging than the work in [30] because it requires to adaptively estimate multiple affinity matrices without any prior knowledge.

C. One Shot Learning

One shot learning aims to study the ability of using a single class example to recognize novel categories. The representative works include [31]–[33], and so on. Since meta-learning methods own the capability of learning to learn, it has become a useful tool for achieving one shot learning. For example, Gidaris and Komodakis [34] used meta-learning to learn the mapping function between classification weight and semantic feature vectors for one shot recognition. Li *et al.* [21] introduced meta-learning-based one-shot detection module in Siamese network for adaptive scale estimation. Similar to our work, Dong *et al.* [35] also introduced one-shot learning-based classification in visual tracking. However, this work could not realize real-time updating because it adopted triplet loss that may not guarantee a fast convergence. In contrast, we design a meta-learning strategy to rapidly learn optimal classification parameters without much experimental setting involved.

III. Proposed Approach

A. Overview

The network structure of our four-stream oriented Siamese network is shown in Fig. 2, where the network contains

four embedding streams. Two streams are used for embedding the target exemplar (target template) pair $\mathbf{Z}_1$ and $\mathbf{Z}_2$. And the other two streams are used for embedding the candidate pair ($\mathbf{x}_i^1$ and $\mathbf{x}_i^2$) within the search regions. After feature embedding, the exemplar embedding pair and the ith candidate embedding pair are, respectively, fused in a reinforcement way through graph attention-based bilinear pooling. This can yield a reinforced target appearance representation for the inner product calculation. It is noted that in traditional Siamese networks, the accuracy of the target location relies on the cross correlation between the exemplar and target candidates. In contrast, our network structure can give a more accurate similarity calculation result. The reason for that is we fully exploit the inherent partial feature interaction existing in the multisource embedding pair through adopting graph attention-based bilinear pooling module. Section III-B introduces the graph attention-based bilinear pooling module in detail. Before introduction, we summarize the notation of main mathematical symbols as follows.

1) $\mathbf{X}_t$ is the video frame at time t. In the proposed network, $\mathbf{X}_t^1$ and $\mathbf{X}_t^2$ means the video frame in RGB and thermal image domains, respectively. For simplicity, the subscript t can be omitted. The ith candidate pair extracted from $\mathbf{X}^1$ and $\mathbf{X}^2$ are denoted as $\mathbf{x}_i^1$ and $\mathbf{x}_i^2$.
2) $\mathbf{F}$ denotes the feature map tensor obtained from CNNs. The reshaped feature map tensor is denoted as $\tilde{\mathbf{F}}$ and the projection result of feature map tensor is $\hat{\mathbf{F}}$.
3) $\mathbf{V}$ is the bilinear pooling results. Notation $\mathbf{V}$ with a subscript in the follow up section indicates the inputs of Siamese networks. Specifically, the bilinear pooling of exemplar pair is denoted as $\mathbf{V}_z$, while the bilinear pooling of ith candidate pair is $\mathbf{V}_{x_i}$. The reduced dimensional vector of $\mathbf{V}_z$ and $\mathbf{V}_{x_i}$ are denoted as $\hat{\mathbf{V}}_z$ and $\hat{\mathbf{V}}_{x_i}$, respectively.
4) $\mathbf{Q}$ denotes similarity matrix in Siamese network structure.

B. Graph Attention-Based Bilinear Pooling

The deep CNN has acquired remarkable achievement in visible spectrum camera-based classification. However, for RGB-T tracking, the state-of-the-art network structures often use linear pooling, e.g., concatenation or element-wise addition, to fuse multilayer multichannel feature maps, which may not make the target fusion result sufficiently expressive to capture the complementary advantages among isolate targets. Above limitation arises from the fact that the deep feature maps are considered as holistic features, and the intrinsic elementwise interaction between different feature maps cannot be fully explored. Bilinear pooling is a promising module that can overcome the limitation of linear pooling because it uses outer product to explore pairwise correlation between feature channels. Suppose we have obtained two domain feature map tensors $\mathbf{F}^1 \in \mathbb{R}^{N\times K\times C}$ and $\mathbf{F}^2 \in \mathbb{R}^{N\times K\times C}$ (N and K are the length and width of a single feature map, and C indicates the number of the feature map channels). After using outer product to multiply the locations of the two tensors and pooling all products together, we can finally obtain the bilinear vector $\mathbf{V} \in \mathbb{R}^{C^2\times 1}$. Since a single element in feature map corresponds to a certain block in original images, if considering the target block as local pattern, the outer product in bilinear pooling can actually explore the structural relationship among local patterns in two image domains. In this way, we can use conditional partial information to represent the target appearance. Reformulating tensors $\mathbf{F}^1$ and $\mathbf{F}^2$ in matrix form $\tilde{\mathbf{F}}^1 \in \mathbb{R}^{NK\times C}$ and $\tilde{\mathbf{F}}^2 \in \mathbb{R}^{NK\times C}$, the bilinear pooling vector can be formulated as

$$\mathbf{V} = \text{bilinear}\big(\tilde{\mathbf{F}}^1, \tilde{\mathbf{F}}^2\big) = \text{vec}\big(\big(\tilde{\mathbf{F}}^1\big)^T \tilde{\mathbf{F}}^2\big) \tag{1}$$

where $\tilde{\mathbf{F}}^1 = [\tilde{\mathbf{f}}_1^1, \ldots, \tilde{\mathbf{f}}_i^1, \ldots, \tilde{\mathbf{f}}_C^1]$ and $\tilde{\mathbf{F}}^2 = [\tilde{\mathbf{f}}_1^2, \ldots, \tilde{\mathbf{f}}_i^2, \ldots, \tilde{\mathbf{f}}_C^2]$, and the $((j-1)C+i)$th element in vector $\mathbf{V}$ is denoted as $\mathbf{V}_{(j-1)C+i} = (\tilde{\mathbf{f}}_i^1)^T \tilde{\mathbf{f}}_j^2$. bilinear $(\cdot)$ indicates the bilinear operator. Each element in vector $\tilde{\mathbf{f}}_i^1$ (or $\tilde{\mathbf{f}}_j^2$) indicates the conditioned local pattern representation for an image block. Equation (1) implies each local pattern representation has equal importance, while ignoring a fact that the contribution of the columns in $\tilde{\mathbf{F}}^1$ and $\tilde{\mathbf{F}}^2$ for multimodel fusion are actually varied. Taking Fig. 1 as an example, there exist only a few image blocks that contain useful yet matched information in RGB-thermal pair. The uninformative image blocks would severely degrade the pooling performance. Thus it is of crucial importance to discriminate the contribution of image blocks. From this observation, we design a graph attention-based bilinear pooling module to exploit coattention mechanism [36]. Specifically, the element of $\mathbf{V}$ is reformulated as

$$\mathbf{V}_{(j-1)C+i} = \big(\tilde{\mathbf{f}}_i^1\big)^T \mathbf{W}_{ij} \tilde{\mathbf{f}}_j^2 \tag{2}$$

where the coattention weight matrix $\mathbf{W}_{ij}$ is aimed to indicate the correlation between elements in vectors $\tilde{\mathbf{f}}_i^1$ and $\tilde{\mathbf{f}}_j^2$. Based on this design, we can highlight those elements in $\tilde{\mathbf{f}}_i^1$ and $\tilde{\mathbf{f}}_j^2$ that yield informative yet matched information.

The motivation of this article is to integrate the target embedding, coattention weight matrix estimation, and feature embedding fusion into a unified end-to-end network structure. To achieve this purpose, the proposed graph attention-based bilinear pooling module combines graph attention convolutional network (GACN) and outer product together, which can effectively utilize message passing to locate the informative image block in both RGB and thermal images with low-computational complexity. The problem formulation for the proposed graph attention-based bilinear pooling module is described as follows.

Based on matrix decomposition, $\mathbf{W}_{ij}$ can be decomposed into

$$\mathbf{W}_{ij} = \mathbf{P}^T \mathbf{D}_{ij} \mathbf{P} \tag{3}$$

where $\mathbf{D}_{ij}$ is the diagonal matrix, which can be further decomposed into two diagonal matrices $\mathbf{D}_{ij} = (\mathbf{S}_i)^T \mathbf{S}_j$. Defining $\mathbf{D}_i = \mathbf{S}_i \mathbf{P}$, $\mathbf{D}_j = \mathbf{S}_j \mathbf{P}$. Based on this definition, taking (3) into (2), we can obtain

$$\mathbf{V}_{(j-1)C+i} = \big(\tilde{\mathbf{f}}_i^1\big)^T \big(\mathbf{D}_i\big)^T \mathbf{P}^T \mathbf{P} \mathbf{D}_j \tilde{\mathbf{f}}_j^2 = \big(\mathbf{P}_i \tilde{\mathbf{f}}_i^1\big)^T \big(\mathbf{P}_j \tilde{\mathbf{f}}_j^2\big). \tag{4}$$

From (4) we can see that $\mathbf{P}_i = \mathbf{P}\mathbf{D}_i$. Defining $\tilde{\mathbf{f}}_i^1 = \mathbf{P}^T \hat{\mathbf{f}}_i^1$, we can obtain

$$\mathbf{P}_i \tilde{\mathbf{f}}_i^1 = \big(\mathbf{P}\mathbf{D}_i \mathbf{P}^T\big) \hat{\mathbf{f}}_i^1. \tag{5}$$

$\mathbf{D}_i$ is the square matrix, it can be further decomposed. Based on this observation and suppose $\mathbf{P}$ is the eigenvector of Laplacian matrix, $(\mathbf{P}\mathbf{D}_i\mathbf{P}^T)\hat{\mathbf{f}}_i^1$ can be considered as the graph convolution. Similarly, $\mathbf{D}_j$ can also be updated using graph convolution. Based on above analysis, let $G(\hat{\mathbf{F}}^1, \hat{\mathbf{A}}^1)$ and $G(\hat{\mathbf{F}}^2, \hat{\mathbf{A}}^2)$ be the attributed graphs for the RGB and thermal feature map tensors, respectively, where the rows in $\hat{\mathbf{F}}^i$ $(i = 1, 2)$ are denoted as the nodes in the ith graph and $\hat{\mathbf{A}}^i$ is the adjacent matrix which encodes the pairwise similarity between nodes pairs. The bilinear pooling-based multiple graphs learning problem is formulated as

$$\mathbf{V} = \text{bilinear}\left(G\left(\hat{\mathbf{F}}^1, \hat{\mathbf{A}}^1\right), \quad G\left(\hat{\mathbf{F}}^2, \hat{\mathbf{A}}^2\right); \boldsymbol{\Theta}\right) \tag{6}$$

where graphs $G(\hat{\mathbf{F}}^1, \hat{\mathbf{A}}^1)$ and $G(\hat{\mathbf{F}}^2, \hat{\mathbf{A}}^2)$ can be learned by graph CNNs, $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}^1, \boldsymbol{\Theta}^2\}$ is defined as the parameter set of graph CNNs, bilinear$(\cdot)$ means the bilinear operator that uses outer product to dynamically aggregate two graph CNNs. Traditional graph CNN often use predefined adjacent matrices for the single graph learning. In (6), due to challenging factors such as occlusion and thermal crossover (TC) and so on, the graph nodes correlation in different image domains dynamically changes, thus it is extremely difficult to simultaneously predefine the suitable adjacent matrices $\hat{\mathbf{A}}^i$. Here, we build GACN for achieving graph learning without any prior knowledge. Specifically, we simplify (5) as

$$\mathbf{P}_i\tilde{\mathbf{f}}_i^1 = \sigma\left(\Sigma_{k\in\mathcal{N}(i)}\eta(i,k)\hat{\mathbf{f}}_k^1\right) \tag{7}$$

where $\eta(i, k)$ denotes the weight of the edge between nodes i and k, $\sigma(\cdot)$ is the activation function, $\mathcal{N}(i)$ denotes neighbor set of node i. Based on (6), we adaptively learn $\eta(i, j)$ to estimate $\mathbf{P}_i\tilde{\mathbf{f}}_i^1$. Similarly, $\mathbf{P}_j\tilde{\mathbf{f}}_j^2$ can be estimated in the same way. Similar to [37], the weight $\eta(i, k)$ for $G(\hat{\mathbf{F}}^1, \hat{\mathbf{A}}^1)$ is calculated by

$$\eta(i,k) = \frac{\exp\left(\text{LeakyReLU}\left(\boldsymbol{\beta}^T\left[\mathbf{U}\hat{\mathbf{f}}_i^1 \| \mathbf{U}\hat{\mathbf{f}}_k^1\right]\right)\right)}{\Sigma_{s\in\mathcal{N}(i)}\exp\left(\text{LeakyReLU}\left(\boldsymbol{\beta}^T\left[\mathbf{U}\hat{\mathbf{f}}_i^1 \| \mathbf{U}\hat{\mathbf{f}}_s^1\right]\right)\right)} \tag{8}$$

where $\boldsymbol{\beta}$ denotes the parameter vector of the single-layer feedforward neural network and $\mathbf{U}$ is the parameter matrix that indicates the relation between $\tilde{\mathbf{A}}$ and $\hat{\mathbf{A}}$. Different from traditional graph attention methods, $\mathbf{U}$ is to learn the normalized row-wise representation of coattention matrix, which can use pairwise information from the other image domain to give a restriction to the estimation of attention weights, avoiding static attention [38] drawback. $\|$ is the concatenation operator, and LeakyReLU$(\cdot)$ is the activation function.

C. Updating Strategy

In this article, we would like to reformulate the updating of graph attention-based bilinear pooling results as a one shot learning problem. This intuition is derived from an observation: the tracking result of the current frame is actually the positive sample. Those candidates that have less similarity with exemplar can be considered as the negative samples. No matter what dramatic changes the current candidates have suffered, the exemplar and current tracking result should still have the same category. Based on this observation, we can incorporate the category information in the online updating of $\hat{\mathbf{V}}_z$ ($\hat{\mathbf{V}}_z$ is the fully connected layer after yielding the bilinear vector of exemplar pair). Specifically, we define the kth classification score of $\hat{\mathbf{V}}_z$ at the first frame as s_k, where $s_k = (\hat{\mathbf{V}}_z)^T\mathbf{M}_k$, with $\mathbf{M}_k$ denoting the weight vector for the kth classification. Inspired by Gidaris and Komodakis [34], we introduce parameter vector $\boldsymbol{\phi}$ in the classification. In this case, the kth classification score after the first frame is changed as $s_k = (\hat{\mathbf{V}}_z \odot \boldsymbol{\phi})^T\mathbf{M}_k$, where $\odot$ denotes Hadamard product. Based on this definition, we can adopt meta-learning to online learn $\boldsymbol{\phi}$ for the fine-tuning of $\hat{\mathbf{V}}_z$. The detailed fine-tuning process is achieved using the category information to enforce $\hat{\mathbf{V}}_z \odot \boldsymbol{\phi}$ to be similar to the bilinear vector of positive candidate pair. Aforementioned strategy is helpful for enhancing the capability of discriminating the exemplar and background.

To achieve meta-learning, we define the ith candidate pooling result $\hat{\mathbf{V}}_{x_i}$ that is most similar to $\hat{\mathbf{V}}_z$ as positive sample $\mathbf{c}_1$, while the jth candidate pooling result $\hat{\mathbf{V}}_{x_j}$ that has lowest similarity with $\hat{\mathbf{V}}_z$ is defined as the negative example $\mathbf{c}_2$. The loss for the online training of $\boldsymbol{\phi}$ is defined as

$$J(\boldsymbol{\phi}) = -\log\mathcal{P}\left(y = 1|\hat{\mathbf{V}}_z\right) \tag{9}$$

where $\mathcal{P}(y = 1\,|\,\hat{\mathbf{V}}_z)$ is defined as

$$\mathcal{P}\left(y = 1\,|\,\hat{\mathbf{V}}_z\right) = \frac{\exp\left(-\|f_{\boldsymbol{\phi}}\left(\hat{\mathbf{V}}_z\right) - \mathbf{c}_1\|^2\right)}{\sum_{k=1}^{2}\exp\left(-\|f_{\boldsymbol{\phi}}(\hat{\mathbf{V}}_z) - \mathbf{c}_k\|^2\right)}. \tag{10}$$

It should be noted that the motivation of the meta-learning strategy in [34] and our meta-learning are quite different, i.e., [34] adopts meta-learning to train parameter vector $\boldsymbol{\phi}$ for fine-tuning classification weight matrix $\mathbf{M}_k$. Its aim is to recognize new categories. In contrast, we adopt $\boldsymbol{\phi}$ to fine-tuning semantic representation for template updating.

D. Inner Product-Based Logistical Loss

As it is shown in Fig. 2, the outputs from two graph attention bilinear pooling modules are defined as bilinear vectors $\mathbf{V}_z$ and $\mathbf{V}_{x_i}$. The dimension of $\mathbf{V}_z$ and $\mathbf{V}_{x_i}$ are all 65536, thus we apply two fully connected layers after obtaining $\mathbf{V}_z$ and $\mathbf{V}_{x_i}$. This can reduce the dimension of $\mathbf{V}_z$ and $\mathbf{V}_{x_i}$ to 256, making them yield dense feature representation. The final outputs of the two graph attention bilinear pooling modules are $\hat{\mathbf{V}}_z$ and $\hat{\mathbf{V}}_{x_i}$. Since the exemplar and candidate pooling results are not the matrices as that in traditional Siamese network, we use inner product to measure the similarity between $\hat{\mathbf{V}}_z$ and $\hat{\mathbf{V}}_{x_i}$. Defining $Q(\hat{\mathbf{V}}_z, \hat{\mathbf{V}}_{x_i})$ as a similarity score in similarity map, the final similarity map is represented as

$$\mathbf{Q} = \begin{bmatrix} Q_1 & Q_2 & \cdots & Q_{\sqrt{k}} \\ Q_{\sqrt{k}+1} & Q_{\sqrt{k}+2} & \cdots & Q_{2\sqrt{k}} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & Q_k \end{bmatrix} \tag{11}$$

where for the sake of simple expression, let $Q(\hat{\mathbf{V}}_z, \hat{\mathbf{V}}_{x_i}) = Q_i$, and Q_i is the ith element in matrix $\mathbf{Q}$. The point with highest similarity score indicates the location of the target. After locating the highest similarity score, we can use interpolation to find the bounding box of the target in the search area.

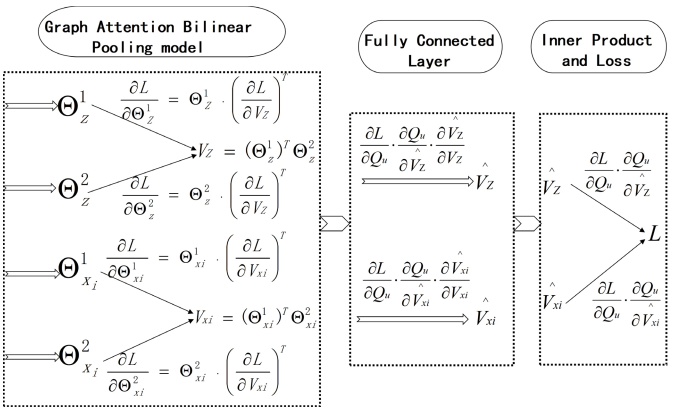


Fig. 3. Detailed chain rule.

Similar to traditional Siamese network, we adopt the logistic loss to train the network with positive and negative sample pairs. The loss function is defined as

$$L = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \log\left(1 + \exp\left(-Y_i Q_i\right)\right) \tag{12}$$

where $\mathcal{D}$ is a set that contains all the shifting positions on the search image. Q_i is the similarity score of the ith reinforced exemplar-candidate pair and Y_i is the corresponding ground truth label.

The parameters of the whole network architecture can be trained by back-propagating the gradients of the loss function. The output feature maps from GACN are defined as $\mathbf{\Theta}_z^1$, $\mathbf{\Theta}_z^2$, $\mathbf{\Theta}_{x_i}^1$ and $\mathbf{\Theta}_{x_i}^2$. The detailed chain rule for our networks is shown in Fig. 3.

IV. Experimental Result

To test the efficiency of our network structure, we have carried out extensive experiments on two widely-used RGB-T datasets: GTOT [6] and RGBT234 [39]. In these experiments, we not only test the quantitative tracking performance but also utilize serious ablation studies to test the effectiveness of the graph attention-based bilinear pooling module and the updating strategy. Compared with state-of-the-art methods, our FS-Siamese network can yield excellent performance on both datasets as shown below.

A. Datasets and Evaluation Matrices

GTOT Dataset contains 50 grayscale-thermal video pairs with seven kinds of challenges: Occlusion (OCC), large scale variation (LSV), fast motion (FM), low illumination (LI), TC, small object (SO), and deformation (DEF).

RGBT234 Dataset contains 234 grayscale-thermal video pairs with 12 kinds of challenges: scale variation (SV), FM (Fast Motion), LI, TC, DEF, none occlusion (NO), partial occlusion (PO), heavy occlusion (HO), motion blur (MB), camera moving (CM), low resolution (LR), and background clutter (BC). The total frame number in this dataset is 210K and the maximum number of frames in a single sequence is 8K.

Evaluation Matrices: Referring to [40], the quantitative evaluation is carried out by three objective measures, namely position plot, success plot, and success rate.

1) ***Precision Plot*** indicates accumulated position errors under different location error thresholds, where the position error is defined as the Euclidean distance between the central location of the tracked bounding box and the manually labeled ground truth.
2) ***Success Rate*** is defined as the number of video frames when the overlap score is larger than 0.5. The overlap score is defined as $\text{area}(B_T \cap B_G)/\text{area}(B_T \cup B_G)$, where B_T denotes the bounding box of the tracked target in current frame, and B_G is the corresponding ground truth.
3) ***Success Plot*** reflects the accumulated success rates versus different overlap thresholds.

B. Implementation Details

In our FS-Siamese, we use VGG-16 as backbone for feature embedding. The bounding box of the first frame is predefined as the exemplar and the size of exemplar pair z_1 and z_2 are 112×112. The search regions X_1 and X_2 are resized to 224×224. Inspired by Qi *et al.* [41], we use the feature maps from four convolutional layers (9, 10, 12, 13th layers) to carry out graph attention-based bilinear pooling, where all feature maps are resized to 14×14. The number of feature maps in each layer is 512. We concatenate four-layer feature maps together to build the graphs for exemplar and candidate pairs. Specifically, there are two kinds of graphs for the bilinear pooling of the exemplar pair (or candidate pair), and we consider each grid of the concatenated feature maps as the node of the undirected graph [42], and the total number of the nodes in a single graph is set to 196. The size of the bilinear vector is $65\,536 \times 1$. We adopt two fully convolutional layers to reduce the size of bilinear vector. The size of the first fully convolutional layer is 512 and the size of the final convolutional layer is 256. The size of the similarity map is 17×17.

We adopt the adaptive momentum (ADAM) optimizer with learning rate of 0.01. The weight decay is set to $5e-4$. The model is trained for 50 epochs with a batch size of 64. In the training process, we first use videos from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC2015) dataset [43] to train the FS-Siamese. Then, we fine-tune the two thermal embedding paths using the first five frames of thermal video sequences in RGBT234. Exponential linear units (ELU) [44] is used in two fully convolutional layers after bilinear vector as the nonlinear activation function. In online tracking, we follow [16] to choose three scales for the current target appearance with scale factors of $1.05\{-1, 0, 1\}$. We update the scale by linear interpolation with a factor of 0.68 to provide damping.

Baselines: Existing trackers mainly focus on using RGB video sequence to carry out visual tracking. By comparison, our tracking method adopts RGB-thermal video pairs to exploit the complementarity of the RGB and thermal targets. The selected competitors include: discriminative scale space tracker (DSST) [45], multi-task sparse learning (MTT) [46], multiple experts using entropy minimization (MEEM) [47], spatially ordered and weighted patch (SOWP) [48], inverse nonnegative local coordinate factorization (INLCF) [49], kernelized correlation filter (KCF) [50], continuous convolution

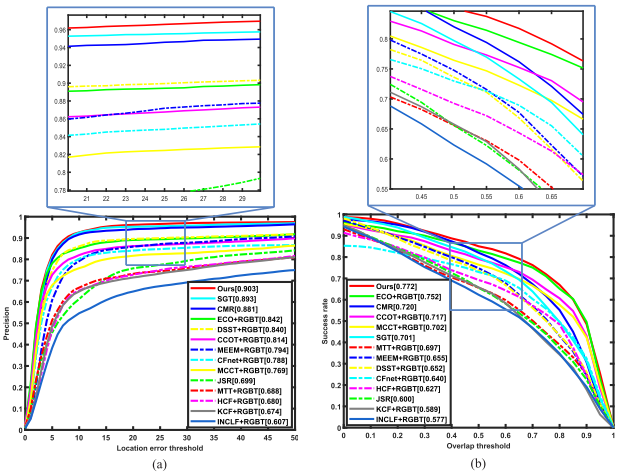


Fig. 4. Overall tracking performance on GTOT dataset: (a) precision plot and (b) success plot. The distance precision score and AUC score are shown in the legend of precision and success plots, respectively. This can indicate the performance of different trackers in precision plot and success plot more clearly.

operators for visual tracking (C-COT) [51], efficient convolution operator (ECO) [52], multi-cue correlation filters for robust visual tracking (MCCT) [53], hierarchical convolutional feature (HCF) [54], MDnet [55], CFnet [56], joint sparse representation (JSR) [57], channel and spatial reliability (CSR) [6], cross-modal ranking (CMR) [58], sparse representation regularized graph tracking (SGT) [4], graph convolutional tracking (GLT) [39], cross-modal pattern-propagation (CMPP) [59], and self-SDCT+RGB [60]. In those competitors, JSR, CSR, CMR, SGT, GLT, and CMPP are the state-of-the-art RGB-T trackers. Beside of RGB-T trackers, other competitors are RGB camera-based trackers. It should be noted that all of RGB camera-based competitors are extended to RGB-thermal version for fair comparison. Specifically, we stack the RGB and thermal features into a single vector for traditional handcraft-based RGB trackers (e.g., MTT, MEEM, and INLCF). Meanwhile, we consider the thermal video sequence as an extra channel in the correlation filter and deep learning-based trackers (e.g., C-COT, ECO, MCCT, MDnet, HCF, and CFnet). The original RGB trackers that have been extended to RGB-T version are given annotation "+**RGBT**." Since ECO is a representative RGB tracker, we also compare our method with this method to test the tracking performance between two-model fusion and a single model-based trackers. The original RGB tracker is given annotation "+**RGB**."

C. Quantitative Tracking Experiments

1) GTOT Dataset:

a) Overall performance: The overall tracking performance on GTOT dataset is shown in Fig. 4. We can clearly see that our method gives the best precision performance. Specially, the distance precision score of our method is higher than ECO-RGBT by over 5%. Since ECO-RGBT involves thermal information, its distance precision score is slight higher than ECO-RGB. The tracking performance in Fig. 4(a) can verify the effectiveness of the proposed fusion module. It is seen from Fig. 4(b) that our method also gives the highest the area under curve (AUC) score. Especially, the AUC score of our method is higher than top RGB-T tracker CMR by over 1%. This can illustrate that our method can use an appropriate bounding box scale to locate the target.

b) Attribute-based performance: The position error only measures the distance between key pixels, which could not reflect the scale of the target. Comparing with position error, the overlap score often gives more comprehensive evaluation on tracking methods because it can evaluate the scale of target bounding box. Thus, we use averaged overlap score to evaluate the tracking performance of different method over seven challenging factors (As seen from Fig. 5), the averaged overlap score of our method in OCC, LSV, LI, and DEF scenarios are higher than other 13 methods. This result can validate our advantage that the graph attention-based bilinear pooling module can explore the partial feature interaction between the RGB and thermal targets. Beside OCC, LSV, LI, and DEF, our method still gives top-2 overlap score in other three attributes, which can illustrate that our method can locate the target with an appropriate bounding box in various challenging scenarios.

2) RGBT234 Dataset:

a) Overall performance: The overall tracking performance on RGBT234 dataset is shown in Fig. 6. RGBT234 contains much more video pairs and involves more challenging factors than GTOT dataset. Thus, it can give a comprehensive and convincing testing on the tracking performance. From Fig. 6(a), we could clearly see that the distance precision score of our method is obviously higher than other 13 comparing methods. Similarly, our method also wins the first place in the successful plot [see Fig. 6(b)]. Especially, the AUC score of our method is higher than well-known deep learning and correlation filter-based trackers such as multi-domain network (MDNet)+RGBT and ECO+RGBT by over 1.5%. This can give a strong support to validate the effectiveness of the proposed network structure. Besides this test, we also give the overall tracking performance on RGBT210 dataset in Fig. 7. RGBT234 dataset is the extension of RGBT210 dataset. From Fig. 7, we could clearly see that our method still win the top place when comparing with other methods.

b) Attribute-based performance: The precision plots over 12 challenging factors are shown in Table I. From this test we could clearly see that our method wins the first place in most challenging factors. Specifically, HO is very challenging because there only a few useful information can be extracted from the RGB and thermal targets. Due to this reason, state-of-the-art tracking methods such as ECO, CMR, and GLT give poor tracking performance in this scenario. Different from traditional methods, the success rate of our method is higher than top method CMPP over 10%. Beside of HO, BC, CM, Fast Motion (FM), LI, and PO are often be considered as the challenging scenarios that can be used as the representative tests to verify the tracking accuracy. Obviously, our method can also enhance the success rate over 6% when comparing with CMPP. Since the thermal target appearance would be seriously disturbed in TC, pooling module may suffer more negative effect when exploring the block relation. Due to this reason, the holistic deep feature-oriented tracker (ECO+RGBT) gives the best success rate. From the testing results in Table I, we can give a confident conclusion that our method can effectively use graph attention-based bilinear pooling module to enhance the tracking performance in challenging scenarios.

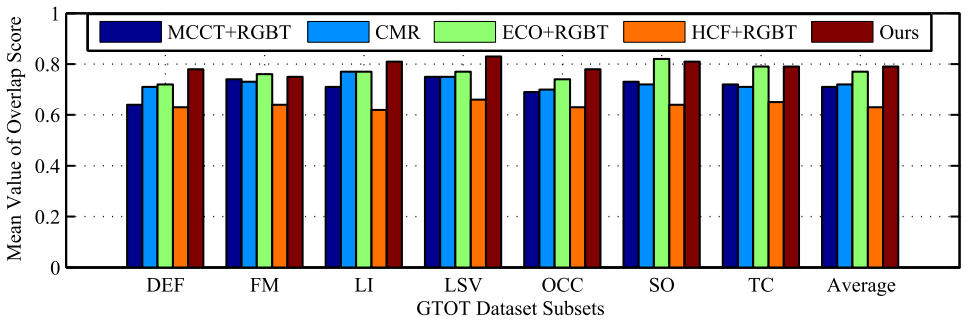


Fig. 5. Mean value of overlap score over different video subsets in GTOT dataset. We select top5 methods in overall tracking performance as competitors.

TABLE I

MEAN VALUE OF SUCCESS RATE OVER DIFFERENT VIDEO SUBSETS IN RGB-234 DATASET. THE BEST TWO RESULTS ARE DENOTED AS AND RED AND BLUE

Attr. \ Meth.	Ours	ECO+RGBT	ECO+RGB	GLT	CFNet+RGBT	CMR	DSST+RGBT	CSR	KCF+RGBT	C-COT+RGBT	MDNet+RGBT	MEEM+RGBT	SOWP+RGBT	SGT	CMPP	self-SDCT+RGB
BC	56.1	52.0	49.9	50.7	28.8	37.6	42.5	34.1	37.5	50.2	48.5	52.8	50.7	50.3	53.8	44.3
CM	57.3	50.8	47.0	43.1	26.9	37.5	33.6	31.8	30.8	49.7	45.6	41.6	42.6	42.9	54.1	37.6
DEF	53.2	46.6	46.9	41.2	28.9	40.1	31.4	33.6	31.1	46.4	47.4	41.8	42.1	43.6	54.1	35.4
FM	54.6	45.6	45.3	41.9	28.6	42.2	30.3	34.7	27.3	45.3	47.6	46.3	45.5	45.2	50.8	36.3
HO	58.1	50.7	49.7	43.6	26.5	39.5	34.1	32.5	30.7	50.8	48.2	45.2	44.9	45.0	50.3	37.0
LI	58.6	52.6	54.0	48.2	32.9	34.4	43.2	34.7	39.1	53.9	43.8	48.1	49.6	45.8	58.4	48.3
LR	63.2	58.4	56.3	58.6	35.4	48.2	53.3	45.4	49.0	64.9	58.6	58.8	57.3	58.8	57.1	56.5
MB	54.5	55.2	49.5	43.3	23.6	37.9	34.1	29.5	29.9	49.9	46.7	37.8	43.2	41.1	54.1	36.1
NO	71.3	66.7	66.4	45.7	50.0	43.1	34.1	47.1	34.9	66.0	59.7	41.3	42.9	47.0	67.8	37.2
PO	62.7	62.2	61.6	50.7	42.6	45.5	43.0	43.1	40.8	62.1	57.6	49.2	52.2	51.3	60.1	43.5
SC	59.7	61.2	58.6	37.3	40.7	38.8	30.7	41.4	28.7	60.4	55.0	36.3	36.8	40.0	57.2	35.5
TC	67.2	70.0	62.4	51.5	41.1	55.0	34.2	37.6	34.7	71.9	59.5	51.9	51.6	57.2	58.3	38.8
Average	59.7	56.0	54.0	46.3	33.8	41.7	37.0	37.1	34.5	56.0	51.5	45.9	46.6	47.4	57.5	40.5

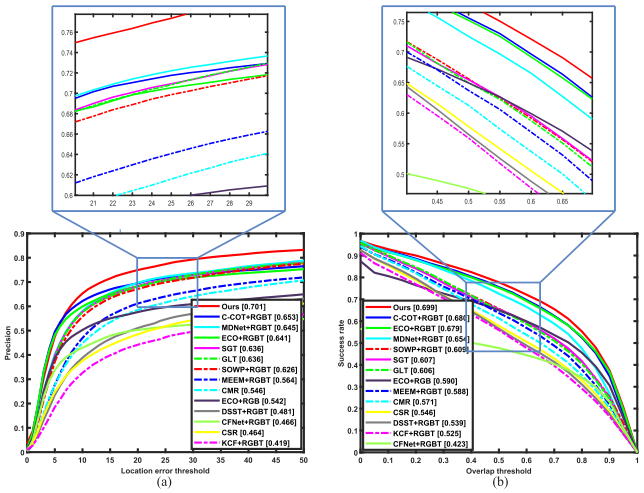


Fig. 6. Overall tracking performance on RGBT234 dataset: (a) precision plot and (b) success plot. The distance precision score and AUC score are shown in the legend of precision and success plots, respectively. This can indicate the performance of different trackers in precision plot and success plot more clearly.

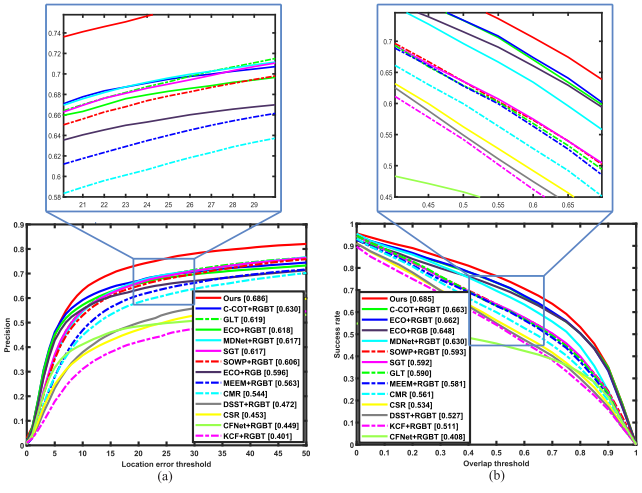


Fig. 7. Overall tracking performance on RGBT210 dataset: (a) precision plot and (b) success plot. The distance precision score and AUC score are shown in the legend of precision and success plots, respectively. This can indicate the performance of different trackers in precision plot and success plot more clearly.

D. Qualitative Tracking Experiments

Here we show the qualitative tracking performance in Fig. 8, where three video sequences are randomly selected from each scenario. The moving target is often occluded by the tree trunk in diamond sequence. State-of-the art methods often lose the target after serious occlusion. From Fig. 8(a), we see that our method can still follow the target no matter the partial or HO. The target and adjacent pedestrians move together, causing serious BC in Fig. 8(b). In this scenario, our method can do the same as ECO-RGBT that gives a good tracking performance. In kite sequence, other methods would begin to drift in some extent after the 300th frame, while our method can still track the kite in whole video frames [Fig. 8(c)]. It contains severe haze in Fig. 8(e). Besides this challenging factor, it also involves occlusion and BC in Fig. 8(e). From this test, we observe that our method can still use an appropriate

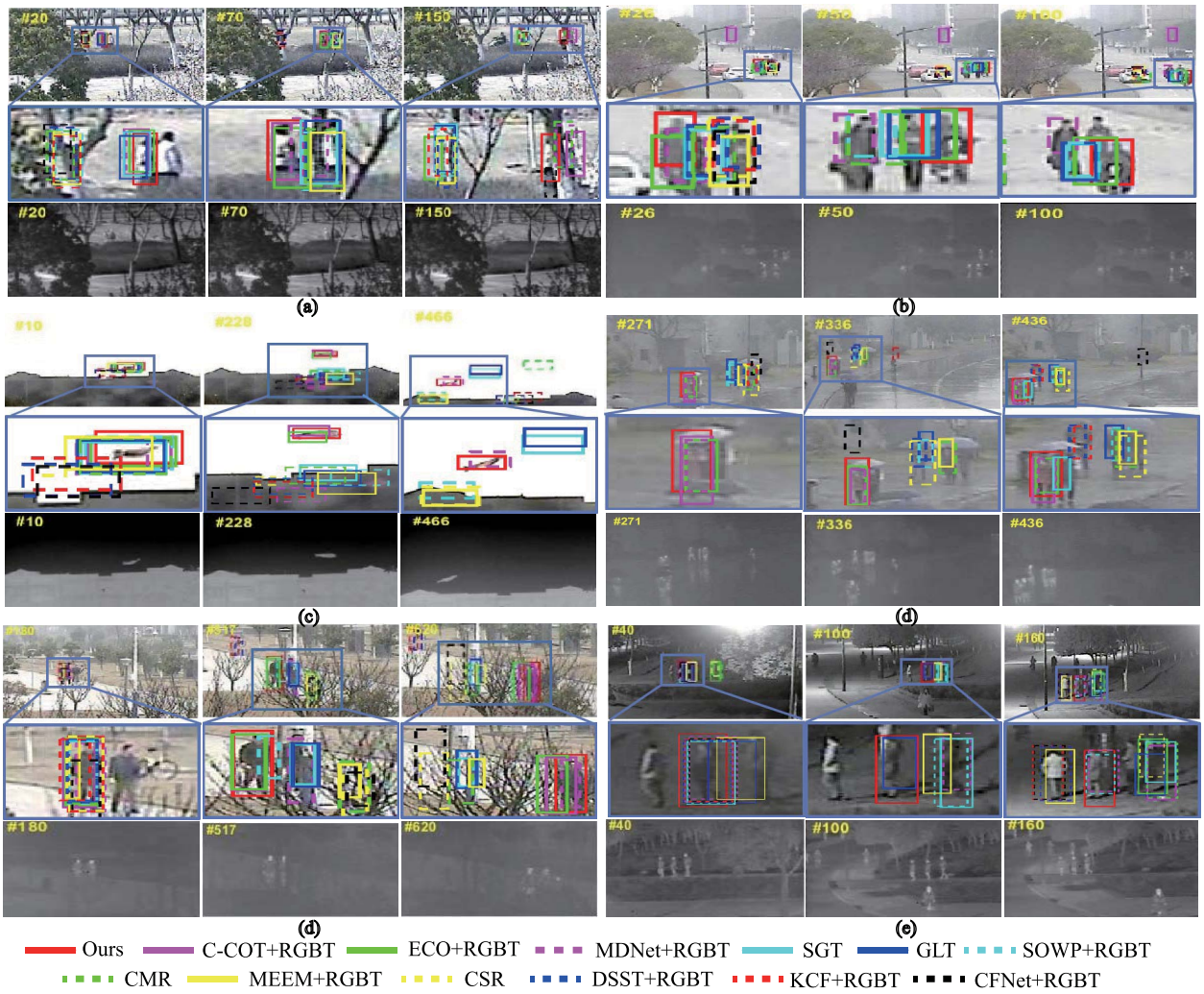


Fig. 8. Qualitative results on six video pairs: (a) diamond video pair; (b) Elecbike3 video pair; (c) Kite4 video pair; (d) Manafterrain video pair; (e) Fog video pair; and (f) Nightthreepeople video pair.

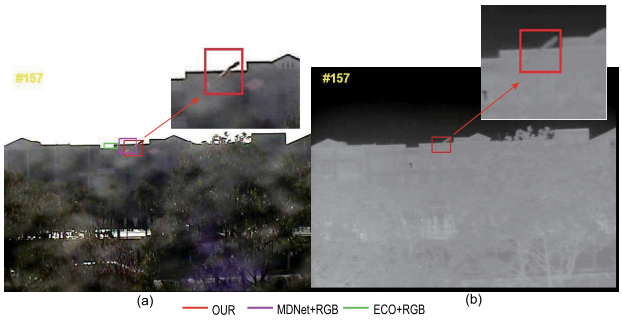


Fig. 9. Qualitative example of RGB-T tracking result in kite sequence pair. Tracking result of (a) RGB image and (b) thermal image.

bounding box to locate the target appearance, while the scale of CMR has dramatically changed when facing occlusion. Kite sequence is a very challenging sequence because the target is really small. Fig. 8(d) and (f) suffer LI in raining and night scenarios. From the two examples, we can see that our method can effectively use the thermal information to complement the RGB sequences.

TABLE II
DETAILED ABLATION SETTING

Backbone	Outer product	GACN	Updating module	Methods
VGG-16	√	√	√	**Ours**
VGG-16	√	√		**Ours I**
VGG-16	√			**Ours II**

Detailed Discussion: Here we take kite sequence pair as example to show the advantage of our method in more detail. Kite sequence is very challenging because the target is small with a long tail. From the local enlarged image in Fig. 9, we can see that the tail of the kite is immersed in the background in both RGB and thermal images. Especially in thermal image, there exists only a few pieces of useful information for the representation of the kite appearance. From the comparison between top-3 tracking results in Fig. 9, we can get a conclusion that our method can effectively use the partial information in RGB and thermal target to guarantee the tracking accuracy.

TABLE III
FPS PERFORMANCE ON DIFFERENT RGB CAMERA-BASED TRACKERS

Tracker	Our	DSST	CFnet	KCF	MEEM	MTT	INCLF	MCCT	SWOP	HCF	MDnet	MCPF	SimaFC
Compiler	python	matlab	python	matlab	matlab & C++	matlab	matlab	matconvnet	matlab	python	python	matconvnet	python
FPS	9.3	10.6	29.5	40.2	33.2	4.3	2.5	2.9	4.9	7.8	0.8	0.2	31.5
Success rate	0.701	0.481	0.466	0.419	0.564	0.558	0.545	0.642	0.609	0.583	0.654	0.625	0.591

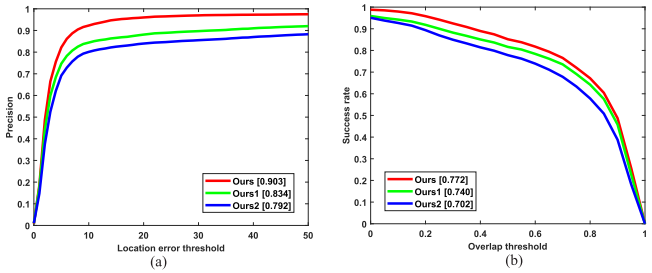


Fig. 10. Ablation test on GTOT dataset: (a) precision plot and (b) success plot.

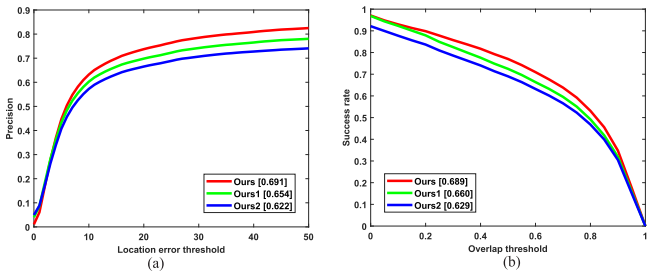


Fig. 11. Ablation test on RGBT234 dataset: (a) precision plot and (b) success plot.

E. Ablation Study

1) Ablation on Graph Attention-Based Bilinear Pooling Module: Graph attention-based bilinear pooling module is the core in our FS-Siamese network, which mainly contains three components: GACN, outer product, and updating module. In this test, we carry out ablation study on GTOT and RGBT234 datasets to show the effectiveness of different components. The detailed experiment setting is shown in Table II. From Figs. 10 and 11 we can see that the precision and success plots on two datasets indicate the effectiveness of our graph attention-based bilinear pooling module.

2) Effectiveness of GACN: GACN is the key point in graph attention-based bilinear pooling, which can highlight the important image blocks through exploring partial feature interaction. In this section, we design a fine-grained classification test to show the effectiveness of GACN. Specifically, we add GACN at the end of Conv layers of multi-attention (MA)-CNN network [61]. In this way, the target embedding would pay more concentration on informative target block. The estimated subregion masks in Fig. 12(c) and (d) can indicate the effectiveness of GACN. For example, although the resolution of the local enlarged images in the first and third rows are low, MA-CNN+GACN can still locate the informative subregion [see Fig. 12(c)]. In contrast, the original method may involve uninformative background noise in the mask [see Fig. 12(d)].

3) Generality of Graph Attention-Based Bilinear Pooling Module: B-CNN [11] is a well-known method in fine-grained recognition that can use bilinear pooling to fuse

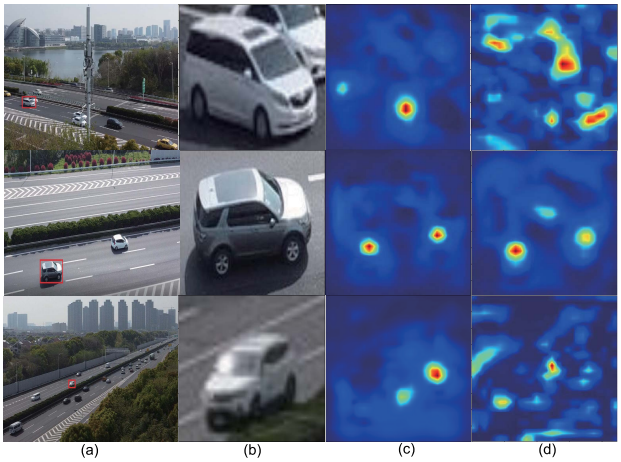


Fig. 12. Informative subregion location test. (a) Testing image, (b) local enlarged image, (c) estimate masks using MA-CNN+GACN, and (d) estimate masks using MA-CNN. MA-CNN can divide the multiple feature channels into four clutter for generating four subregion masks. Images (b) and (c) are obtained from the second clutter.

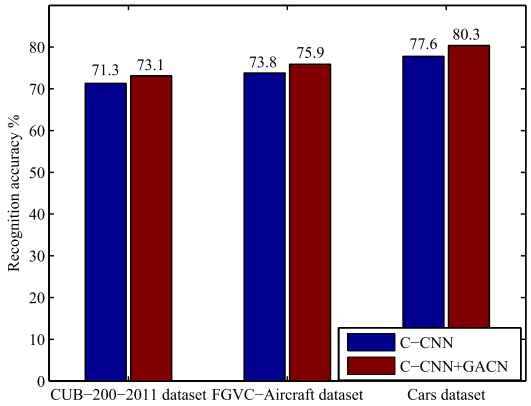


Fig. 13. Fine-grained recognition for testing the generality of GACN.

feature maps of two network structures. Here we extend our graph attention-based bilinear pooling module to this method, namely, "B-CNN+GACN," for verifying the generality of the key innovation of FS-Siamese. The testing is carried out on three fine-grained recognition datasets: CUB-200-2011, FGVC-aircraft, and Cars. From Fig. 13 we could clearly see that B-CNN+GACN can obviously enhance the recognition accuracy over 3% when comparing when original B-CNN methods.

F. Tracking Speed

In this test, we use the number of frames per second (FPS) as the objective measure for evaluating the online tracking speed. The FPS testing is carried out on computer workstation with single NVIDIA GTX1060Ti GPU. It should be noted that this test is carried out on RGBT234 dataset. Table III gives the tracking speed comparison between our method and RGB camera-based tracking methods. Since our method

TABLE IV
FPS PERFORMANCE ON DIFFERENT RGB-T TRACKERS

Tracker	Our	L1-PF	CSR	SGT	LGMG [5]	MCSR [7]
Compiler	python	matlab&C++	matlab	matlab	matlab	matlab
FPS	9.3	7.0	0.9	2.3	1.3	0.3
Success rate	0.701	0.431	0.463	0.636	0.698	0.688

1) involves two more VGG-16-based feature embedding paths and 2) achieves updating at every frame when comparing with traditional SimaFC, this may reduce the tracking speed in some extent. We have also shown speed comparison between our method and the state-of-the-art RGB-T trackers in Table IV. Clearly, our tracking speed surpasses traditional handcraft based RGB-T trackers.

V. CONCLUSION

In this article, we have proposed a four-stream oriented Siamese network (FS-Siamese) to effectively fuse RGB and thermal information. Our network has benefited from the proposed graph attention-based bilinear pooling module that can adopt coattention mechanism to explore the partial feature interaction between the RGB and the thermal targets. Besides, we have adopted meta-learning to update the bilinear pooling result, which can perform online updating on the spatial relation between the target and its surrounding background. Extensive experiments on GTOT and RGBT234 datasets indicated that the proposed FS-Siamese network can give a superior performance as compared to the state-of-the-art RGB and RGB-T trackers.

REFERENCES

[1] Y. Choi *et al.*, "KAIST multi-spectral day/night data set for autonomous and assisted driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 934–948, Mar. 2018.

[2] A. Leykin and R. I. Hammoud, "Pedestrian tracking by fusion of thermal-visible surveillance videos," *Mach. Vis. Appl.*, vol. 21, no. 4, pp. 587–595, Jun. 2010.

[3] M. Talha and R. Stolkin, "Particle filter tracking of camouflaged targets by adaptive fusion of thermal and visible spectra camera data," *IEEE Sensors J.*, vol. 14, no. 1, pp. 159–166, Jan. 2014.

[4] 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.

[5] C. Li, C. Zhu, J. Zhang, B. Luo, X. Wu, and J. Tang, "Learning local-global multi-graph descriptors for RGB-T object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2913–2926, Oct. 2019.

[6] 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.

[7] X. Lan, M. Ye, R. Shao, B. Zhong, P. C. Yuen, and H. Zhou, "Learning modality-consistency feature templates: A robust RGB-infrared tracking system," *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9887–9897, Dec. 2019.

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

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

[10] 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.

[11] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1449–1457.

[12] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1821–1830.

[13] Q. Wang, Z. Teng, J. Xing, J. Gao, and S. Maybank, "Learning attentions: Residual attentional Siamese network for high performance online visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4854–4863.

[14] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware Siamese networks for visual object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 101–117.

[15] L. Zhang, A. Gonzalez-Garcia, J. V. D. Weijer, M. Danelljan, and F. S. Khan, "Learning the model update for Siamese trackers," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jan. 2019, pp. 4010–4019.

[16] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional Siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 850–865.

[17] X. Li, C. Ma, B. Wu, Z. He, and M.-H. Yang, "Target-aware deep tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1369–1378.

[18] Y. Zhang, L. Wang, J. Qi, D. Wang, M. Feng, and H. Lu, "Structured Siamese network for real-time visual tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 351–366.

[19] Z. Liang and J. Shen, "Local semantic Siamese networks for fast tracking," *IEEE Trans. Image Process.*, vol. 29, pp. 3351–3364, 2020.

[20] M. Gao, L. Jin, Y. Jiang, and B. Guo, "Manifold Siamese network: A novel visual tracking convnet for autonomous vehicles," *IEEE Trans. Intell. Transport. Syst.*, vol. 21, no. 4, pp. 1612–1623, Apr. 2020.

[21] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with Siamese region proposal network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8971–8980.

[22] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of Siamese visual tracking with very deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4282–4291.

[23] H. Fan and H. Ling, "Siamese cascaded region proposal networks for real-time visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7952–7961.

[24] P. Jingchao, Z. Haitao, H. Zhengwei, Z. Yi, and W. Bofan, "Siamese infrared and visible light fusion network for RGB-T tracking," 2021, *arXiv:2103.07302*.

[25] A. He, C. Luo, X. Tian, and W. Zeng, "A twofold Siamese network for real-time object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4834–4843.

[26] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodel compact bilinear pooling for visual question answering visual grounding," in *Proc. Empirical Methods Natural Lang. Process.*, 2016, pp. 1–12.

[27] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–14.

[28] X. Chang, F. Nie, S. Wang, Y. Yang, X. Zhou, and C. Zhang, "Compound rank-k projections for bilinear analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 7, pp. 1502–1513, Jul. 2016.

[29] X.-S. Wei, P. Wang, L. Liu, C. Shen, and J. Wu, "Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 6116–6125, Dec. 2019.

[30] J. Gao, T. Zhang, and C. Xu, "Graph convolutional tracking," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4649–4659.

[31] O. Vinyals *et al.*, "Matching networks for one shot learning," in *Proc. Conf. Workshop Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 3630–3638.

[32] L. Bertinetto, J. F. Henriques, J. Valmadre, P. Torr, and A. Vedaldi, "Learning feed-forward one-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1–9.

[33] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Conf. Workshop Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 4077–4087.

[34] S. Gidaris and N. Komodakis, "Dynamic few-shot visual learning without forgetting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4367–4375.

[35] X. Dong, J. Shen, D. Wu, K. Guo, X. Jin, and F. Porikli, "Quadruplet network with one-shot learning for fast visual object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3516–3527, Jul. 2019.

[36] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Conf. Workshop Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 1–9.

[37] V. Petar, C. Guillem, C. Arantxa, R. Adriana, L. Pietro, and B. Yoshua, "Graph attention networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–12.

[38] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?" in *Proc. ICLR*, 2021, pp. 1–26.

[39] 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.

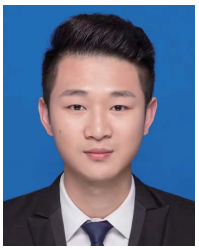
[40] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.
[41] Y. Qi *et al.*, "Hedging deep features for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 5, pp. 1116–1130, May 2019.
[42] Z. Cui, Y. Cai, W. Zheng, C. Xu, and J. Yang, "Spectral filter tracking," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2479–2489, May 2019.
[43] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[44] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *Proc. Int. Congerence Learn. Represent. (ICLR)*, 2016, pp. 1–14.
[45] 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. 1–5.
[46] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via structured multi-task sparse learning," *Int. J. Comput. Vis.*, vol. 101, no. 2, pp. 367–383, Jan. 2013.
[47] 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.
[48] H. U. Kim, D. Y. Lee, J. Y. Sim, and C. S. Kim, "Sowp: Spatially ordered and weighted patch descriptor for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2016, pp. 3011–3019.
[49] F. Liu, T. Zhou, C. Gong, K. Fu, L. Bai, and J. Yang, "Inverse nonnegative local coordinate factorization for visual tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1752–1764, Aug. 2017.
[50] 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.
[51] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 472–488.
[52] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6638–6646.
[53] N. Wang, W. Zhou, Q. Tian, R. Hong, M. Wang, and H. Li, "Multi-cue correlation filters for robust visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4844–4853.
[54] C. Ma, J. Huang, X. Yang, and M. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2015, pp. 3074–3082.
[55] 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.
[56] 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.*, Jul. 2017, pp. 2805–2813.
[57] H. Liu and F. Sun, "Fusion tracking in color and infrared images using joint sparse representation," *Inf. Sci.*, vol. 55, no. 3, pp. 590–599, 2012.
[58] 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.*, 2018, pp. 808–823.
[59] C. Wang *et al.*, "Cross-modal pattern-propagation for RGB-T tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7064–7073.
[60] D. Yuan, X. Chang, P. Huang, and Q. Liu, "Self-supervised deep correlation tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 976–985, 2020.
[61] H. Zheng, J. Fu, M. Tao, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5209–5217.

Bin Kang received the Ph.D. degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2016.

He is currently an Associate Professor with the College of Internet of Things, Nanjing University of Posts and Telecommunications. His research interests include computer vision and pattern recognition.

Dong Liang received the B.S. degree in telecommunication engineering and the M.S. degree in circuits and systems from Lanzhou University, Lanzhou, China, in 2008 and 2011, respectively, and the Ph.D. degree from the Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan, in 2015.

He is currently an Associate Professor with the Pattern Recognition and Neural Computing Laboratory, College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His research interests include computer vision and pattern recognition.



Junxi Mei received the B.S. degree in communication engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2021.

His current research interests include machine learning for vision, object tracking and recognition.

Xiaoyang Tan received the Ph.D. degree from the Nanjing University, Nanjing, China, in 2005.

He was a Post-Doctoral Researcher with the LEAR Team, INRIAR Rhone-Alpes, Grenoble, France, from 2006 to 2007. He is currently a Professor with the Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing. He has authored or coauthored more than 50 conference papers and journal articles. His research interests include deep learning, reinforcement learning, and Bayesian learning.

Dr. Tan's colleagues were awarded the IEEE Signal Processing Society Best Paper in 2015.

Quan Zhou (Member, IEEE) received the Ph.D. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2013.

He is currently an Associate Professor with the Nanjing University of Posts and Telecommunications, Nanjing, China. He has published more than 30 articles in top journals (e.g., IEEE TRANSACTIONS ON IMAGE PROCESSING, *Pattern Recognition (PR)*, *English Lingua (EL)*, and *Sensors*) in image processing and computer vision. His research topics include image labeling and scene understanding, visual attention and saliency detection, and face identification and recognition.

Dr. Zhou currently serves as a Reviewer for a series of SCI journals, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON COMPUTERS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and *Neurocomputing*.

Dengyin Zhang (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 1986, 1989, and 2004, respectively.

He was a Visiting Scholar with the Digital Media Laboratory, Umea University, Umeå, Sweden, from 2007 to 2008. He is currently a Professor with the School of Internet of Things, Nanjing University of Posts and Telecommunications. His research interests include signal and information processing, networking technique, and information security.