# Multiple Stream Oriented Siamese Network for RGB-T Tracking

Yimo Wang*, Songlin Du* Quan Zhou†, and Bin Kang‡§
*School of Automation
Southeast University, Nanjing, China
†College of Communications and Information Engineering
Nanjing University of Posts and Telecommunications, Nanjing, China
‡College of Internet of Things
Nanjing University of Posts and Telecommunications, Nanjing, China
§Key Laboratory of Broadband Wireless Communications and Sensor Network Technology
Ministry of Education (NUPT), Nanjing, China

2021 13th International Conference on Wireless Communications and Signal Processing (WCSP) | 978-1-6654-0785-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/WCSP52459.2021.9613242

***Abstract*—RGB-T tracker owns the capability of fusing two different yet complementary target observations, thus it will become a promising technology to fulfill all-weather tracking. Existing convolutional neural network based tracking methods often consider the multi-source oriented deep feature fusion from global viewpoint, leading to inevitable negative effect when feature maps of the target pair only contain little useful information. To solve this problem, we propose a four-stream oriented Siamese network, named as FS-Siamese, for RGB-T tracking. In particular, we introduce co-attention mechanism in bilinear pooling to explore the partial feature interaction between the RGB and thermal targets. This can effectively avoid uninformed image blocks disturbing feature embedding fusion. To enhance the efficiency of our Siamese network, we also propose an inner product based logistical loss for training the feature embedding and the graph convolutional neural network based bilinear pooling in an end-to-end manner. Extensive experiments on GTOT datasets demonstrate that the proposed method achieves state-of-the-art performance in the task of RGB-T tracking.**

## I. Introduction

With the rapid development of Internet of Things, 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. Jointly using RGB and thermal infrared cameras involve two advantages: 1) Thermal infrared camera is skilled in resisting illumination change; 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.

The key point of RGB-T tracking is to exploit the complementarity of the RGB and thermal information for efficient multi-model fusion. To this end, state-of-the-art methods can be briefly categorized into two classes. The first one is to build multiple graph fusing model to effectively exploit the spatial relation between the RGB and thermal target blocks [1]. The second 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 [2]. All of aforementioned methods use handcraft feature for multi-model fusion. Compared with handcraft feature, deep convolutional neural networks can extract the translation and light invariant deep semantic information for robust representation of the target. Thus deep learning technology has give promising performance gain in RGB-T tracking recently. For example, the authors of [3] proposed a dense convolutional neural network for RGB-T tracking, which can recursively aggregate informative features of two kinds of convolutional paths.

Existing CNN based RGB-T trackers often consider the multi-layer 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. For example, in a darkness scenario, the tracking target pass through a bush, the target may be partially occluded. Moreover, the appearance is also obscurious in the darkness. In this case, directly fusing holistic features between thermal and RGB targets would obviously degrade the tracking accuracy because of the disturbance of the uninformative target appearance. Based on above observation, a promising solution is to achieve part-feature based thermal and RGB deep feature fusion. Based on this consideration, we propose a simple four-stream oriented Siamese network (FS-Siamese) for RGB-T tracking, 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 co-attention based bilinear pooling module. Choosing bilinear pooling as fusion model due to the fact that this model can use outer product to explore pairwise correlations between the feature channels, which is helpful for highlighting the importance of informative image blocks in RGB and thermal targets. The main contributions of this paper are two folds: (1) Based on the theoretical analysis of co-attention mechanism, we integrate graph convolutional neural network and bilinear pooling into a unified model, which can dynamically integrate informative image blocks in RGB and thermal image domains. (2) We design a inner operator based function for generating the similarity maps of the Siamese network, which can effec-

978-1-6654-0785-4/21/$31.00 ©2021 IEEE

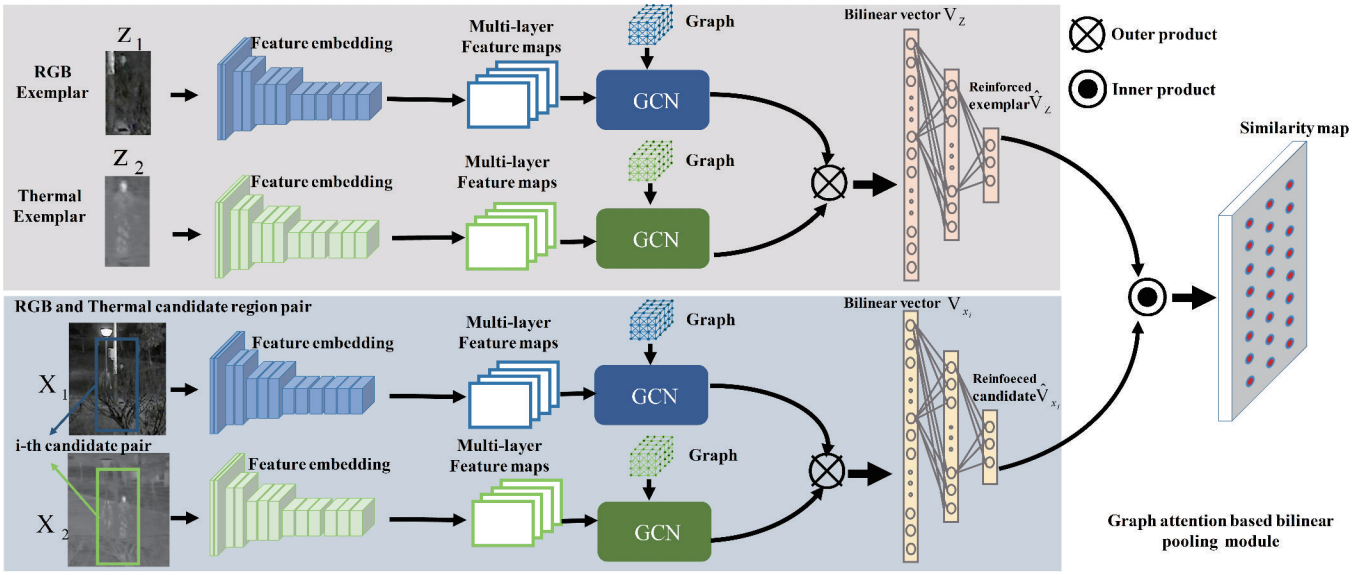



Fig. 1. The pipeline of our FS-Siamese, which is consisted of three components: (1) feature embedding, (2) GCN based bilinear pooling module for generating reinforced exemplar and candidate, (3) inner product for calculating the similarity between reinforced exemplar and each reinforced candidate within the search region.

tively exclude the disturbance of the background.

## 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* [4] is the pioneer who designes 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.* [5]) 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 patten based Siamese Networks (*e.g.* [6]) that can explore the spatial relation between different target blocks; 3) The RPN based Siamese networks (*e.g.* [7]) that introduce region proposal network in Siamese network to avoid the time-consuming multi-scale estimation step.

### B. Bilinear pooling

After the work in [8] that used multi-modal compact bilinear pooling to explore the pairwise relation between two heterogeneous models, bilinear pooling has become an effective tool in VQA (Visual Question Answering). Since the dimension of the output bilinear vector in [8] is high, Kim *et.al* [9] 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 fine-grained recognition, *e.g.* Lin *et.al* [10] 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* [11] used bilinear pooling to explore the partial feature interaction between two fine-grained models.

## III. Our approach

### A. Overview

The proposed network structure of our four-stream oriented Siamese network is shown in Fig. 1, where the network contains four embedding streams. Two streams are used for embedding the target exemplar (target template) pair $z_1$ and $z_2$. And the other two streams are used for embedding the candidate pair within the search regions $X_1$ and $X_2$. After feature embedding, the exemplar embedding pair and the $i$-th 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 multi-source embedding pair through adopting graph attention based bilinear pooling module. The following introduces the graph attention based bilinear pooling module in detail.

### B. Graph attention based bilinear pooling

The deep convolutional neural network has acquired remarkable achievement in visible spectrum camera based clas-

sification. 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 multi-layer multi-channel feature maps, which may not make the target fusion result sufficiently expressive to capture the complementary advantages among isolate targets. Above limitation arises from a fact that the deep feature maps are considered as holistic features, and the intrinsic elementwise interaction between different feature maps can not 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 correlations between the feature channels. Suppose we have obtained two feature map tensors $\mathbf{A} \in \mathbb{R}^{N\times K\times C}$ and $\mathbf{B} \in \mathbb{R}^{N\times K\times C}$ ($N$ 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 $\mathbf{u} \in \mathbb{R}^{C^2\times 1}$. Since a single element in feature map corresponds to an 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 $\mathbf{A}$ and $\mathbf{B}$ in matrix form $\tilde{\mathbf{A}} \in \mathbb{R}^{NK\times C}$ and $\tilde{\mathbf{B}} \in \mathbb{R}^{NK\times C}$, the bilinear pooling vector can be formulated as

$$\mathbf{u} = vec(\tilde{\mathbf{A}}^T\tilde{\mathbf{B}}). \tag{1}$$

where $\tilde{\mathbf{A}} = [\tilde{\mathbf{a}}_1, ..., \tilde{\mathbf{a}}_i, ..., \tilde{\mathbf{a}}_C]$ and $\tilde{\mathbf{B}} = [\tilde{\mathbf{b}}_1, ..., \tilde{\mathbf{b}}_j, ..., \tilde{\mathbf{b}}_C]$, and the $((j-1)C+i)$th element in vector $\mathbf{u}$ is denoted as $\mathbf{u}_{(j-1)C+i} = \tilde{\mathbf{a}}_i^T\tilde{\mathbf{b}}_j$. The element in vector $\tilde{\mathbf{a}}_i$ (or $\tilde{\mathbf{b}}_j$) indicates the conditioned local pattern representation for an image block. Eq. (1) considers each local pattern representation to have equal importance, while ignoring a fact that the contribution of the columns in $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$ for multi-model fusion are actually varied. To highlight the importance image blocks, we adopt co-attention mechanism [12] to reformulate the element of $\mathbf{u}$ as

$$\mathbf{u}_{(j-1)C+i} = \tilde{\mathbf{a}}_i^T\mathbf{W}_{ij}\tilde{\mathbf{b}}_j, \tag{2}$$

where the co-attention weight matrix $\mathbf{W}_{ij}$ is aimed to indicate the correlation between elements in vectors $\tilde{\mathbf{a}}_i$ and $\tilde{\mathbf{b}}_j$. The motivation of this paper is to integrate the target embedding, co-attention weight matrix estimation and feature embedding fusion into a unified end-to-end network structure. To achieve this purpose, the graph convolutional neural network and outer product can be combined together, which can effectively utilize message passing to locate the informative image block in both RGB and thermal images with low computational complexity. The theoretical analysis of above motivation 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. This diagonal matrix can be further decomposed into

$$\mathbf{D}_{ij} = (\mathbf{D}_i)^T\mathbf{D}_j, \tag{4}$$

Defining $\mathbf{Z}_i = \mathbf{D}_i\mathbf{P}$, $\mathbf{Z}_j = \mathbf{D}_j\mathbf{P}$ and taking Eq. (4) into Eq. (3), we can obtain

$$\mathbf{u}_{(j-1)C+i} = (\tilde{\mathbf{a}}_i)^T(\mathbf{Z}_i)^T\mathbf{P}^T\mathbf{P}\mathbf{Z}_j\tilde{\mathbf{b}}_j = (\mathbf{P}_i\tilde{\mathbf{a}}_i)^T(\mathbf{P}_j\tilde{\mathbf{b}}_j), \tag{5}$$

From Eq.(5) we can see that $\mathbf{P}_i\tilde{\mathbf{a}}_i = \mathbf{P}\mathbf{Z}_i\tilde{\mathbf{a}}_i$. Defining $\tilde{\mathbf{a}}_i = \mathbf{P}^T\hat{\mathbf{a}}_i$, we can obtain

$$\mathbf{P}_i\tilde{\mathbf{a}}_i = (\mathbf{P}\mathbf{Z}_i\mathbf{P}^T)\hat{\mathbf{a}}_i. \tag{6}$$

Supposing $\mathbf{P}$ is the eigenvector of Laplacian matrix, $\hat{\mathbf{a}}_i$ is the projection of $\tilde{\mathbf{a}}_i$. Since square matrix $\mathbf{Z}_i$ can be further SVD decomposed, thus $(\mathbf{P}\mathbf{Z}_i\mathbf{P}^T)\hat{\mathbf{a}}_i$ can be considered as the graph convolution, where $\mathbf{Z}_i$ can be iteratively updated in convolutional manner. Similarly, $\mathbf{Z}_j$ can also be updated using graph convolution. Based on above analysis, we formulate the rows in $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$ as the nodes in graphs $\mathcal{G}_1(\upsilon, \varepsilon)$ and $\mathcal{G}_2(\upsilon, \varepsilon)$, respectively. In this way, Eq. (2) can be reformulated as the operator for dynamically pooling two graph convolutional neural networks.

### C. Inner product based logistical loss

As it is shown in Fig. 1, 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 $\mathbf{V}_{x_i}$ is 65536, thus we add 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

$$Q(Z, X) = \begin{bmatrix} Q_1 & Q_2 & ... & Q_{\sqrt{k}} \\ Q_{\sqrt{k}+1} & Q_{\sqrt{k}+2} & ... & Q_{2\sqrt{k}} \\ ... & ... & ... & ... \\ ... & ... & ... & Q_k \end{bmatrix}, \tag{7}$$

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 $i$-th element in matrix $Q(Z, X)$. 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.

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

$$L(Y, Q(Z, X)) = \frac{1}{|\mathcal{D}|}\sum_{u\in\mathcal{D}} log\big(1 + exp(-Y_uQ_u)\big), \tag{8}$$

where $\mathcal{D}$ is a set that contains all the shifting positions on the search image. $Q_u$ is the similarity score of the $u$-th reinforced exemplar-candidate pair and $Y_u$ is the corresponding ground truth label. Based on above loss function, the final loss of the network is defined as:

$$arg \min_{Q} \quad L\big(Y, Q(Z, X)\big). \tag{9}$$

## IV. Experimental Result

### A. 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 $112 \times 112$. The search regions $X_1$ and $X_2$ are resized to $224 \times 224$. Inspired by [13], we use the feature maps from 4 convolutional layers (9, 10, 12, 13-th layers) to carry out GCN based bilinear pooling, where all feature maps are resized to $14 \times 14$. We adopt the 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.

### B. Baseline

The selected competitors include: ECO [14], C-COT [15], CFnet [16], MEEM [17], SGT [18], where ECO, C-COT and MEEM are well known RGB based tracking methods. CFNet is the well known deep learning based RGB tracker. SGT is the state-of-the-art RGB-T trackers.

### C. Overall performance

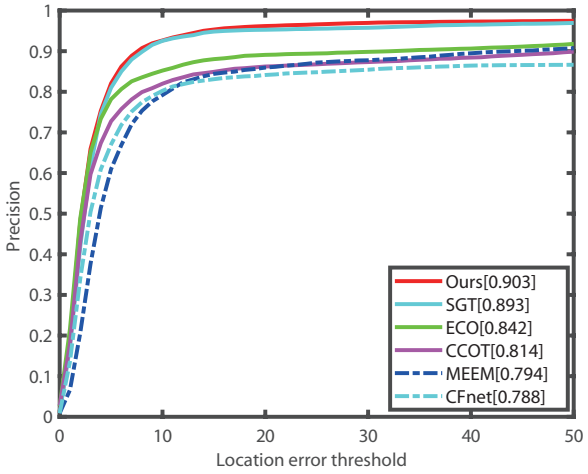
In this paper, we choose GTOT dataset [19] to carry out the experiments. From Fig. 2 we could clearly see that our method wins the first place in both precision and success plots. ECO is a classic tracking methods that are widely used for testing the tracking performance, thus the precision score of our method in precision plot is higher than ECO over 2%. Similarly, the AUC score of our method in success plot is also higher than ECO over 3%. Those two results can give a strong support for verifying the proposed network.

### D. Attribute based performance

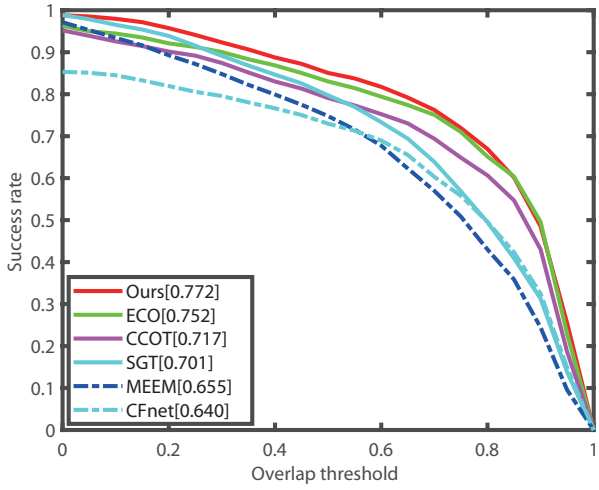
GTOT dataset contains 50 grayscale-thermal video pairs with 7 kinds of challenges: OCC (Occlusion), LSV (Large Scale Variation), FM (Fast Motion), LI (Low Illumination), TC (Thermal Crossover), SO (Small Object) and DEF (Deformation). In this test, we show attribute based performance on 7 challenges (see Table 1). From this test, we could clearly see that our methods wins top 2 in all challenges. Especially in OCC, LSV, LI and DEF scenarios, the average overlap scores are higher than other 6 methods, which can validate our advantage that the GCN based bilinear pooling module can explore the partial feature interaction between the RGB and thermal targets.

### E. Subjective performance

In this part, we randomly select two examples in GTOT dataset to show the subjective performance (see Fig. 3). Since the target in video sequence is very small, we enlarge the local area in the selected frames, which can show the tracking performance more clearly. It should be noted that the video frame in the selected video sequence are randomly.




(a)




(b)

Fig. 2. The overall tracking performance on GTOT dataset: (a) the precision plot, (b) the success plot.

## V. Conclusion

In this paper, 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 GCN based bilinear pooling module that can adopt co-attention mechanism to explore the partial feature interaction between RGB and thermal targets. Extensive experiments on GTOT datasets indicated that the proposed network can give a


Fig. 3. The subjective results on two video pairs: (a) Pool, (b) FastMotorNig.

TABLE I
MEAN VALUE OF OVERLAP SCORE OVER DIFFERENT VIDEO SUBSETS IN GTOT DATASET. THE BEST RESULT IS DENOTED AS RED

| Attr. \ Meth. | Ours | ECO | C-COT | SGT | MEEM | CFNet |
|---|---|---|---|---|---|---|
| OCC | 60.3 | 55.2 | 52.2 | 48.6 | 45.9 | 42.3 |
| LSV | 56.9 | 52.1 | 50.9 | 46.8 | 44.8 | 31.2 |
| FM | 58.3 | 55.8 | 53.3 | 47.9 | 46.1 | 42.2 |
| LI | 54.6 | 57.3 | 51.3 | 48.1 | 45.3 | 42.5 |
| TC | 62.4 | 60.9 | 55.4 | 50.2 | 47.2 | 42.7 |
| SO | 59.6 | 57.1 | 53.1 | 50.3 | 45.6 | 42.0 |
| DEF | 55.5 | 54.9 | 50.3 | 44.9 | 41.2 | 37.3 |
| Average | 58.2 | 56.2 | 52.6 | 48.1 | 45.4 | 40.2 |

superior performance as compared to the state-of-the-art RGB and RGB-T trackers.

## VI. ACKNOWLEDGMENTS

This work was jointly supported in part by the National Natural Science Foundation of China under grant (62171232,61801242,62001110,61876093), the Natural Science Foundation of Jiangsu Province under grant BK20200353 and BK20181393, the Guangdong Basic and Applied Basic Research Foundation under grant 2020A1515110145, and the Shenzhen Science and Technology Program under grant RCBS20200714114858072. SRTP program 202110286065.

## REFERENCES

[1] C. Li, C. Zhu, J. Zhang, B. Luo, and J. Tang, "Learning local-global multi-graph descriptors for rgb-t object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 2913–2926, 2019.

[2] 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 Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9887–9897, 2019.

[3] Y. Zhu, C. Li, B. Luo, J. Tang, and X. Wang, "Dense feature aggregation and pruning for rgbt tracking," in *Proc. of the ACM International Conference on Multimedia*, 2019.

[4] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. of European Conference on Computer Vision*, 2016.

[5] X. Li, C. Ma, B. Wu, Z. He, and M.-H. Yang, "Target-aware deep tracking," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1369–1378.

[6] M. Gao, L. Jin, Y. Jiang, and B. Guo, "Manifold siamese network: A novel visual tracking convnet for autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 4, pp. 1612–1623, 2020.

[7] 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. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4282–4291.

[8] 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. of the Empirical Methods in Natural Language Processing*, 2016.

[9] 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. of the International Conference on Learning Representation*, 2017.

[10] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proc. of the IEEE international conference on computer vision*, 2015, pp. 1449–1457.

[11] 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 Transactions on Image Processing*, vol. 28, no. 12, pp. 6116–6125, 2019.

[12] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. of the Conference and Workshop on Neural Information Processing Systems (NIPS)*, 2016.

[13] Y. Qi, S. Zhang, L. Qin, Q. Huang, H. Yao, J. Lim, and M.-H. Yang, "Hedging deep features for visual tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 5, pp. 1116–1130, 2019.

[14] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "Eco: Efficient convolution operators for tracking," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[15] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. of European Conference on Computer Vision*, 2016.

[16] 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. of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[17] J. Zhang, S. Ma, and S. Sclaroff, "Meem: Robust tracking via multiple experts using entropy minimization," in *Proc. of European Conference on Computer Vision*, 2014.

[18] C. Li, N. Zhao, Y. Lu, C. Zhu, and J. Tang, "Weighted sparse representation regularized graph learning for rgb-t object tracking," in *Proc. of the ACM international conference on Multimedia*, 2017.

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