J Inf Process Syst, Vol.19, No.1, pp.80~88, February 2023
https://doi.org/10.3745/JIPS.01.0092

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Dynamic Tracking Aggregation with Transformers for RGB-T Tracking

Xiaohu Liu[1,*] and Zhiyong Lei[2]

Abstract

RGB-thermal (RGB-T) tracking using unmanned aerial vehicles (UAVs) involves challenges with regards to the similarity of objects, occlusion, fast motion, and motion blur, among other issues. In this study, we propose dynamic tracking aggregation (DTA) as a unified framework to perform object detection and data association. The proposed approach obtains fused features based a transformer model and an L1-norm strategy. To link the current frame with recent information, a dynamically updated embedding called dynamic tracking identification (DTID) is used to model the iterative tracking process. For object association, we designed a long short-term tracking aggregation module for dynamic feature propagation to match spatial and temporal embeddings. DTA achieved a highly competitive performance in an experimental evaluation on public benchmark datasets.

Keywords

Cross-modal Fusion, Dynamic Tracking Aggregation, RGB-T Tracking, Transformers

1. Introduction

Recently, multi-object tracking via unmanned aerial vehicles (UAVs) [1,2] has emerged as a topic of active research owing to the widespread popularity of UAVs [3]. Although multi-object tracking methods have been considerably improved, many challenges remain in regards to their application onboard UAVs. Meanwhile, with the popularization of sensors systems with different modalities, applications of visible-thermal (RGB-T) imaging have attracted attentions owing to the advantages of complementary information [4,5]. By combining these complementary features, RGB-T tracking can improve UAV tracking performance. Various algorithms have been developed, for example, Li et al. [6] considered the shared-modality and modality-specific representations and proposed a multi-adaptor learning network. Zhang et al. [7] utilized valid attribute information and proposed an RGB-T tracker to meet the requirement of real-time operation.

However, their method was trained on small-scale datasets such as RGBT210, RGBT234, or synthetic data generated from visible imaging [5,8]; however, its generalization ability was limited and the training process was limited by the available data such that processing similar objects, occlusion, fast motion, and motion blur situations is typically difficult. In this study, we propose an approach called dynamic tracking aggregation with transformers (DTA) that unifies multi-object detection and instance association. Firstly,

Manuscript received August 4, 2022; first revision October 4, 2022; accepted October 10, 2022.
* **Corresponding Author:** Xiaohu Liu (liuxiaohu@st.xatu.edu.cn)
[1] School of Mechatronic Engineering, Xi'an Technological University, Xi'an, China (liuxiaohu@st.xatu.edu.cn)
[2] School of Electronic and Information Engineering, Xi'an Technological University, Xi'an, China (leizy888@163.com)

a Swin transformer [9] based encoder is used to obtain feature representation from two different modalities.

A log-spaced contiguous position bias is used for position embedding to deal with variations in resolution. Secondly, L1-norm based fusion is applied to model the interaction and dependency between the two modalities. Moreover, we also propose a long short-term tracking aggregation (LSTA) block for dynamic feature propagation, which consists of long- and short-term attention to match spatial and temporal embeddings. We also provide the results of experiments conducted to validate the efficacy of the proposed DAT method.

2. Related Work

2.1 RGB-T Tracking

The correspondence and discriminability of multi-modal information [10] can be exploited by RGB-T tracking. Methods to fuse the features of the two modalities can broadly be divided into three levels, including pixel-, feature-, and decision-level fusion. For pixel-level fusion, multiple layers with shared weights are applied to obtain the heterogeneously complementary information. For example, the method considered by Peng et al. [11] is highly dependent on image alignment. In contrast, feature-level fusion uses the features from different modalities as input. Fusion feature methods can be trained using concatenation strategies or attention techniques, and can be optimized with massive unaligned data to achieve significant improvements in performance. Each information modality is independently modeled by decision fusion, and the final candidate is obtained by score fusion. JMMAC [12] utilized a multi-modal network to fuse the modality-level and pixel-level representations.

2.2 MOT with Transformers

Transformers models have shown great advantages in classification, detection, and tracking. TrackFormer [13] takes object and autoregressive track queries as inputs to the decoder of the transformer block to simultaneously detect objects and associate instances of different frames. TransTrack [14] obtained an aggregation embedding of each object by recurrently passing track features extracted from a transformed encoder. TransMOT [15] used convolutional neural networks (CNNs) to extract features, and an affinity matrix was trained using transformers. Nevertheless, these methods still involve some challenges for use onboard UAVs, such as issues with similar objects, occlusion, fast motion, and motion blur, especially in long-term tracking.

3. Dynamic Tracking Aggregation with Transformers

3.1 Overview

Given a video sequence $I_l = \{I^0, I^1, \cdots, I^T\}$, where $l = ir$ for infrared images, and $l = vis$ for visible images, the MOT needs to localize the K objects, and simultaneously maintains the object trajectories $T = \{T_0, T_1, \cdots, T_K\}$ over different frames. We proposed an end-to-end tracking algorithm called DTA to unify object detection and association stages. Our tracker is composed of four main components,

including the encoder module to extract features of different modalities, the fusion layer to aggregation features of different modalities, the LSTA block for dynamic feature matching and propagation, and an association solver for object association, as shown in Fig. 1.

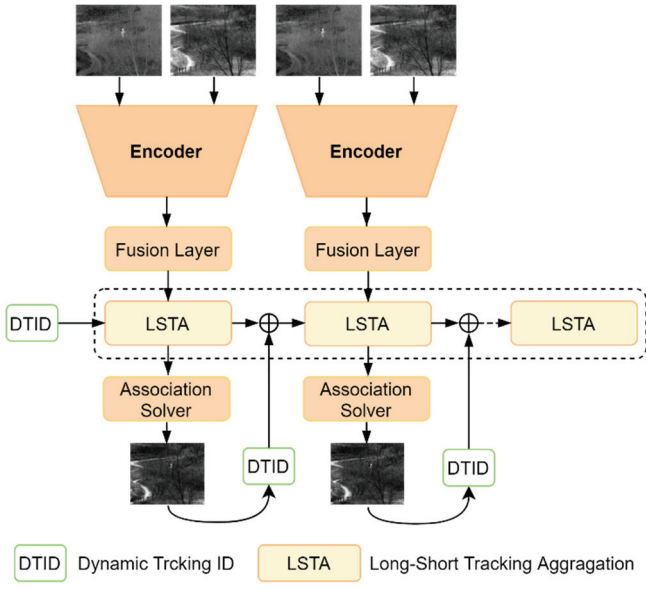


Fig. 1. Overview of the proposed dynamic tracking aggregation.

3.2 Transformer-based Encoder for Feature Extraction

We adopt a Swin transformer [9] model is as the backbone encoder, in comparison to traditional CNN models, and a more compact feature representation can be obtained with richer semantic information which promotes the succeeding networks for localizing the target objects.

We denote the input image as $I_l \in R^{H\times W\times 3}$, and the embedding feature from the input as $Z_l \in R^{\frac{H}{s}\times\frac{W}{s}\times C}$, where s is the stride of the backbone network. Because the VTUAV dataset has a higher resolution of 1,920×1,080, we use the log-spaced contiguous position bias for position embedding to apply the transferred model with high performance [16] as given below in Eq. (1).

$$\widehat{\Delta x} = sign(x) \cdot \log\left(1 + |\Delta x|\right) \tag{1}$$

where $\widehat{\Delta x}$ is the log-spaced coordinates of the relative position bias and Δx is the linear-scaled coordinates of the same. The relative position bias was obtained using a layer MLP with ReLU activation.

3.3 Fusion Layer for Modality-Feature Aggregation

Complementary information is present in RGB-T images; therefore, a robust feature representation can be obtained by propagating information between the two. To model the interaction and dependency between the two modalities, we designed the block-based $L1-norm$ fusion strategy to score the activity level of the row and column dimensions [17].

Firstly, the row vector weights were calculated by $L1-norm$, and the softmax function is adopted to obtain the activity level of feature maps, referred to as $\phi_l^{row}(i)$ by Eq. (2).

$$\phi_l^{row}(i) = \frac{\exp(|Z_l(i)|)}{\sum_l \exp(|Z_l(i)|)}, i \in (-r, r) \tag{2}$$

where r determines the block size. Then, the fused features of the row vector dimension, termed as $\Phi_l^{row}(i,j)$, are as given below in Eq. (3).

$$\Phi_l^{row}(i,j) = \sum_l Z_l(i,j) \times \phi_l^{row}(i) \tag{3}$$

Similar to the above operations, the fused features of column vector dimension, termed as $\Phi_l^{col}(i,j)$, are as given below in Eq. (4).

$$\Phi_l^{col}(i,j) = \sum_l Z_l(i,j) \times \phi_l^{col}(i) \tag{4}$$

Finally, we utilize the element-wise addition to obtain the global fused representation as given Eq. (5).

$$\Phi(i,j) = \Phi_l^{row}(i,j) + \Phi_l^{col}(i,j) \tag{5}$$

3.4 Long Short-Term Tacking Aggregation for Dynamic Feature Propagation

In contrast to previous methods [9,13], which propagate the object features between adjacent frames, we developed a LSTA block for dynamic feature propagation, as shown in Fig. 2, which enables information for linking objects to be efficiently retrieved for a given time span.

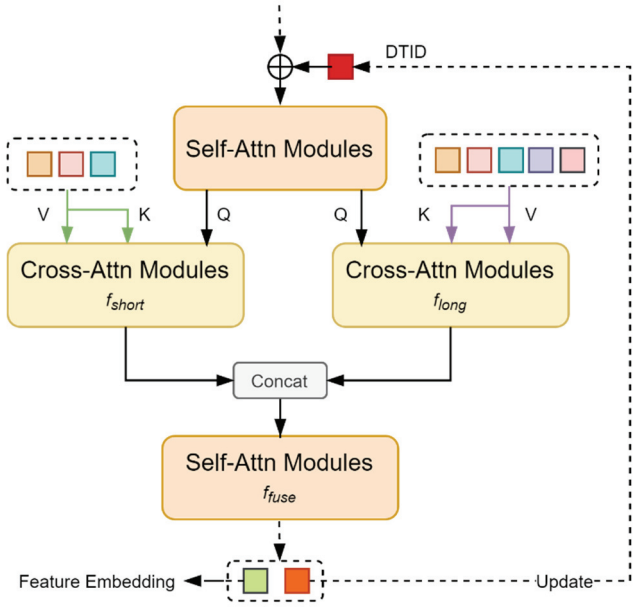


Fig. 2. Long short-term tracking aggregation comprising different attention modules.

Previous methods [13,14] have utilize only static attention modules to aggregate single-object information, but the multi-object association cannot be fully modeled. We constructed an LSTA block for dynamic feature matching and propagated it with multiple attention layers based on transformer blocks.

Specifically, we encode the past frame features and extract the track embedding with stacked attention modules. (i) A self-attention module is utilized to associate the object detected in the current frame. (ii) We apply a short-term block f_{short} to assemble the embeddings of neighboring frames and simultaneously smooth out noise. (iii) We also use a long-term block f_{long} to extract relevant features within the temporal window with parameter δ. Finally, (iv) a fusion block f_{fuse} is used to aggregate the features of short-term and long-term embeddings.

Following [18], better performance can be achieved by updating the queries dynamically using late features. A dynamically updated embedding, called dynamic tracking identification (DTID), is used to model the iterative tracking process, where the previous DTID is used to update the current solution iteratively. The f_{short} takes as previous T_s features as input, while f_{long} applies longer historical features with length T_l ($T_s \ll T_l$). Cross-attention modules with multiple heads are used in f_{short} and f_{long}, where the key and value take the past features as inputs.

3.5 Association Solver

To obtain the target bounding boxes and object class, an auxiliary linear decoder was utilized, and set prediction [19] was used in the association solver with an end-to-end optimization objective. In particular, a bipartite matching mechanism was applied to formulate the loss function. Denote $y = \{y_i\}_{u=1}^{M}$ the ground truth, and $\hat{y} = \{\hat{y}_i\}_{u=1}^{N}$ as the predictions (generally M<N), and the total loss can be defined as:

$$L(y, \hat{y}) = \sum_{i=1}^{N} \left[\lambda_{cls} L_{cls}^{i} + \lambda_{box} 1_{\{y \neq \emptyset\}} L_{box}^{i} + \lambda_{iou} L_{iou}^{i}\right] \tag{6}$$

where L_{box}^{i} is the bounding box regression loss, and L_{class}^{i} is the classification loss and bounding box regression loss.

4. Implementation Details

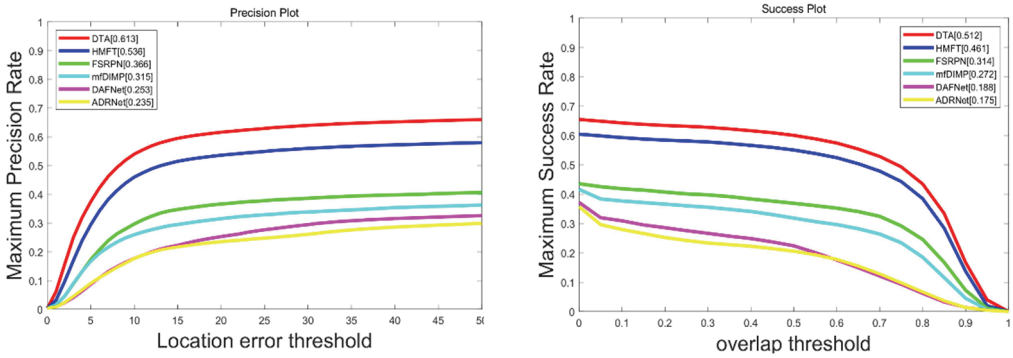
The proposed algorithm was implemented using PyTorch, and all experiments were performed with four Tesla A100 GPUs. In the training process, random flip and crop data augmentations were utilized. We used Swin-B [9] as the encoder with the last stage removed and flattened the features from the encoder to sequences before the LSTA module; and the dimension of the features was set to 256, and the attention head was set to 8.

Following [20], the training consisted of two phases, including (i) pre-training the Swin-B [9] encoder for object detection with multiple image augmentations applied, and (ii) the main training process on the RGBT-234, RGBT-210, and VTUAV benchmarks [5]. For the subsequent transformer module, we reduced the number of layers to four. The short-term s was set to 1, and the long-term δ was set to 7/16 for training and testing. Following [20,21], the coefficients of Hungarian loss with λ_{cls}, λ_{reg}, λ_{iou} were selected as 2, 5 and 2, respectively.

To stabilize the training, an exponential moving average (EMA) [22] was used. In addition, spatial attention was applied first, followed by temporal attention to prompt the training process, and the AdamW [23] optimizer was utilized. In the pre-training stage, the initial learning rate was set to 4×10^{-4}, and in the main training stage, it was set to 2×10^{-4} with a batch size of 16, at the 100-th epoch, it was decreased by 10.

5. Experimental Results

We applied the proposed approach to the VTUAV benchmark and compared its performance with that of five RGB-T trackers—DAFNet [24], ADRNet [5], FSRPN [25], mfDiMP [26], and HMFT [5]. As shown in Fig. 3. DTA achieved the best performance with 51.2% MSR (maximum success rate) and 61.3% MPR (maximum precision rate) on the long-term test subset.



(a) (b)

Fig. 3. Evaluation results on long-term test subset on VTUAV: (a) MPR and (b) MSR.

Compared to other CNN-based methods, DTA exhibited much better tracking performance with a LSTA design with dynamic tracking identification. This is consistent with the results in other tasks [17] that attention-based features can extract global information compared to convolutions. Traditional CNNs take each frame independently, and the features of each frame have a limited effect on future tracking. However, the LSTA module can apply long-term features to compensate for this problem.

Further, from the evaluation on different attributes shown in Fig. 4, it may be observed that DTA is robust to many challenging situations important for UAV tracking, including camera movement, fast motion, partial occlusion, and target blur. For partial occlusion and target blur challenges, considering the temporal variance, we need long-term information to integrate the diverse features, where the context information plays an important role. For camera movement and fast motion challenges, owing to the high similarity of the adjacent frames, the latest feature can be used for object association to remove noise; this renders the short-term features are more important. The proposed method enables a dynamically trade-off in these two situations to obtain complementary information.

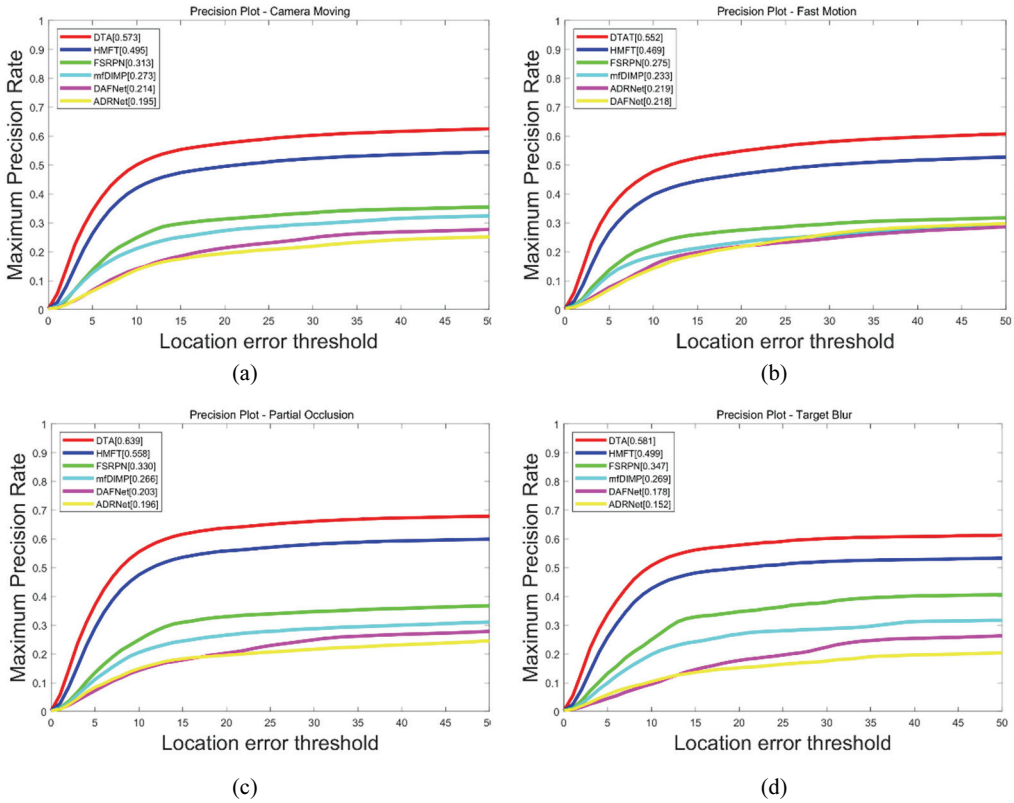


Fig. 4. Evaluation results on different attributes: (a) camera moving, (b) fast moving, (c) partial occlusion, and (d) target blur.

6. Conclusion

In this study, we have proposed a novel and efficient approach for RGB-T tracking by dynamic tracking aggregation, and showed that it achieved competitive performance on benchmarks datasets. The transformer-based encoder and L1-norm strategy can effectively fuse the features of the two modalities. In addition, combined with dynamic tracking identification embedding, we have proposed a long short-term tracking aggregation designed for dynamic feature propagation to match spatial-temporal embeddings. With this structure, DAT can track objects without post-processing. Self-supervised learning methods can also be exploited in future work to construct models of larger scale and improve the efficiency of the training process.

Acknowledgement

The research was funded by the Department of Education of Shaanxi Province, China (No. 21JK0819).

References

[1] S. M. Azimi, M. Kraus, R. Bahmanyar, and P. Reinartz, "Multiple pedestrians and vehicles tracking in aerial imagery using a convolutional neural network," *Remote Sensing*, vol. 13, no. 10, article no. 1953, 2021. https://doi.org/10.3390/rs13101953.

[2] L. Wen, D. Du, P. Zhu, Q. Hu, Q. Wang, L. Bo, and S. Lyu, "Detection, tracking, and counting meets drones in crowds: a benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Virtual Event, 2021, pp. 7812-7821.

[3] I. Delibasoglu, "UAV images dataset for moving object detection from moving cameras," 2021 [Online]. Available: https://arxiv.org/abs/2103.11460.

[4] P. Zhang, D. Wang, and H. Lu, "Multi-modal visual tracking: review and experimental comparison," 2020 [Online]. Available: https://arxiv.org/abs/2012.04176.

[5] P. Zhang, J. Zhao, D. Wang, H. Lu, and X. Ruan, "Visible-thermal UAV tracking: a large-scale benchmark and new baseline," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, 2022, pp. 8876-8885.

[6] C. Li, A. Lu, A. Zheng, Z. Tu, and J. Tang, "Multi-adapter RGBT tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, Seoul, South Korea, 2019, pp. 2262-2270.

[7] P. Zhang, D. Wang, H. Lu, and X. Yang, "Learning adaptive attribute-driven representation for real-time RGB-T tracking," *International Journal of Computer Vision*, vol. 129, pp. 2714-2729, 2021.

[8] T. Zhang, X. Liu, Q. Zhang, and J. Han, "SiamCDA: complementarity-and distractor-aware RGB-T tracking based on Siamese network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1403-1417, 2021.

[9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, 2021, pp. 10012-10022.

[10] H. Zhang, L. Zhang, L. Zhuo, and J. Zhang, "Object tracking in RGB-T videos using modal-aware attention network and competitive learning," *Sensors*, vol. 20, no. 2, article no. 393, 2020. https://doi.org/10.3390/s20020393.

[11] J. Peng, H. Zhao, Z. Hu, Z. Yi, and B. Wang, "Siamese infrared and visible light fusion network for RGB-T tracking," 2021 [Online]. Available: https://arxiv.org/abs/2103.07302.

[12] P. Zhang, J. Zhao, C. Bo, D. Wang, H. Lu, and X. Yang, "Jointly modeling motion and appearance cues for robust RGB-T tracking," *IEEE Transactions on Image Processing*, vol. 30, pp. 3335-3347, 2021.

[13] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "TrackFormer: multi-object tracking with transformers," 2022 [Online]. Available: https://arxiv.org/abs/2101.02702.

[14] P. Sun, J. Cao, Y. Jiang, R. Zhang, E. Xie, Z. Yuan, C. Wang, and P. Luo, "TransTrack: multiple object tracking with transformer," 2021 [Online]. Available: https://arxiv.org/abs/2012.15460.

[15] P. Chu, J. Wang, Q. You, H. Ling, and Z. Liu, "TransMOT: spatial-temporal graph transformer for multiple object tracking," 2021 [Online]. Available: https://arxiv.org/abs/2104.00194.

[16] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, et al., "Swin transformer v2: scaling up capacity and resolution," 2022 [Online]. Available: https://arxiv.org/abs/2111.09883.

[17] Z. Wang, Y. Chen, W. Shao, H. Li, and L. Zhang, "SwinFuse: a residual swin transformer fusion network for infrared and visible images," 2022 [Online]. Available: https://arxiv.org/abs/2204.11436.

[18] J. Cai, M. Xu, W. Li, Y. Xiong, W. Xia, Z. Tu, and S. Soatto, "MeMOT: multi-object tracking with memory," 2022 [Online]. Available: https://arxiv.org/abs/2203.16761.

[19] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: deformable transformers for end-to-end object detection," 2020 [Online]. Available: https://arxiv.org/abs/2010.04159.

[20] Z. Yang, Y. Wei, and Y. Yang, “Associating objects with transformers for video object segmentation,” 2021 [Online]. Available: https://arxiv.org/abs/2106.02638.

[21] G. Bertasius and L. Torresani, “Classifying, segmenting, and tracking object instances in video with mask propagation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, 2020, pp. 9736-9745.

[22] B. T. Polyak and A. B. Juditsky, “Acceleration of stochastic approximation by averaging,” *SIAM Journal on Control and Optimization*, vol. 30, no. 4, pp. 838-855, 1992.

[23] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2019 [Online]. Available: https://arxiv.org/abs/1711.05101.

[24] Y. Gao, C. Li, Y. Zhu, J. Tang, T. He, and F. Wang, “Deep adaptive fusion network for high performance RGBT tracking,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, Seoul, South Korea, 2019, pp. 91-99.

[25] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, R. Pflugfelder, J. K. Kamarainen, et al., “The seventh visual object tracking VOT2019 challenge results,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, Seoul, South Korea, 2019, pp. 2206-2241.

[26] L. Zhang, M. Danelljan, A. Gonzalez-Garcia, J. van de Weijer, and F. Shahbaz Khan, “Multi-modal fusion for end-to-end RGB-T tracking,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, Seoul, South Korea, 2019, pp. 2252-2261.

Xiaohu Liu **https://orcid.org/0000-0003-0896-1810**

He received a master's degree from Xi’an Technological University in 2013. Since 2017, he is with the school of Mechanical and Electrical Engineering from Xi’an Technological University. His current research interests include intelligent information processing and deep learning.

Zhiyong Lei **https://orcid.org/0000-0003-1741-9943**

He is currently a professor in the school of Electronic Information Engineering, Xi’an Technological University. His research interests are computer measurement and control technology, intelligent sensor networks and information fusion, computer vision and image processing.