Siamese Multi-Scale Aggregation Network for UAV Tracking

Meiyu Yao[1], Na Wu[1*], Shuo Hu[1], Hui Yu[2]

***Abstract*—The Siamese-based trackers have received much attention due to their great performance in the field of target tracking. However, it ignores the relationships and interdependencies between different features, impeding the robustness under various conditions. In addition, most Siamese-based trackers suffer from multiple special challenges, such as Fast Motion, Occlusion in UAV tracking. In this paper, we propose an anchor-free based object tracking algorithm with multi-scale aggregation Siamese Network. The proposed method consists of three parts: the feature extraction network, Encoder and Decoder. A multi-scale receptive field structure is designed in the encoder to deal with the problem of multi-scale change. The design of adaptive anchor in the decoder effectively reduces the relevant hyper-parameters. Experiments on three challenging UAV tracking benchmarks have demonstrated the robustness and effectiveness of the proposed method.**

***Keywords*—Object tracking, UAV tracking, Anchor-free, Siamese-based tracker, intelligent vehicles**

I. INTRODUCTION

Target tracking is an elementary and challenging task for computer vision and intelligent vehicles [1]. It starts with the initial frame target at the position and size of a given video or image sequence, tracking the indicated target frame by frame. In recent years, UAV has gradually attracted more attention by many scholars in the field of target tracking due to its small size, simple operation and flexible action, such as path planning[2], aerophotography[3].

As a type of conventional single target tracking methods, UAV tracking is based on the initial frame state of the target to predict its position in the subsequent frame. Generally, due to the limited resources of UAV resources, UAV target tracking needs more attention in terms of long-time tracking and low computational cost. However, how to improve the accuracy of the UAV trackers with robustness while keep efficiency is a challenging issue.

Generally, UAV tracking systems also face lots of exceptional challenges from movable terraces, such as fast movement, low resolution, severe occlusion, and long-term tracking. At present, the target tracking algorithms are divided into two categories.

*Research supported by National Natural Science Foundation of China (No.62073279).

M. Yao is with the School of Electrical Engineering, Yanshan University, Qinhuangdao, 066004 China (e-mail: 653706476@qq.com).

N. Wu is with the School of Electrical Engineering, Yanshan University, Qinhuangdao, 066004 China (corresponding author, Phone:0086-3358387556, e-mail: wunamay@sina.com).

S. Hu is with the School of Electrical Engineering, Yanshan University, Qinhuangdao, 066004 China (e-mail: hus@ysu.edu.cn)

H. Yu is the School of Creative Technologies, University of Portsmouth, Portsmouth, PO1 2DJ U.K(e-mail: hui.yu@port.ac.uk)

Generally, UAV tracking systems also face lots of exceptional challenges from movable terraces, such as fast movement, low resolution, severe occlusion, and long-term tracking. At present, the target tracking algorithms are divided into two categories. One is Correlation Filter-based algorithms, and the other is Deep Learning-based. In literature, online CF-based trackers can reduce the amount of calculation[4][5]**Error! Reference source not found.Error! Reference source not found.**, so it can be adopted on a large scale. Although highly efficient, the CF-based trackers are difficult to complete the tracking task of the trackers in difficult scenes, such as similar background, occlusion and low resolution in terms of accuracy and robustness, while the DL-based algorithms have can successfully complete the tracking task by exploiting deep feature extraction[6][7][8]. Consequently, the DL-based algorithms are a promising approach to balancing performance and speed[9]. Currently, researchers propose a series of anchor-based algorithms[10][11]. Since the anchors of the above algorithms are predefined, the efficiency of the anchor-based trackers is low, and its tracking performance needs to be further improved. Moreover, the introduction of predefined anchors without prior knowledge restricts the performance of trackers. Subsequently, researchers propose [12][13] the anchor-free based algorithm to strengthen the generality of the tracker. This algorithm tracks the target by predicting the distance between the central point and the four sides of the ground-truth box. Despite the anchor-free based algorithm improves the efficiency of the tracker, the problem of sample imbalance is still not well resolved.

In this paper, we propose an anchor-free based object tracking algorithm with multi-scale aggregation Siamese Network. As shown in Figure 1, the proposed architecture contains a feature extraction network, Encoder and Decoder. The feature extraction network extracts hierarchical semantic information and feeds it into Encoder. The structure of an Encoder can be regarded as a single in and out encoder, and it includes two parts: The Mapping layer and the Remainder blocks. The Decoder structure is composed of two parts, which are the Anchor Proposal Network (APN) and the multi-classification regression network.

In summary, the contributions of the proposed method can be summarized as follows:

1. An anchor-free based target tracking method is proposed, and the proposed algorithm significantly improves the performance of the tracker in challenging scenarios in the UAV tracking filed.

2. To resolve the problems of the multi-scale change in object tracking, an Encoder structure is designed by fusing multi-scale receptive field structure

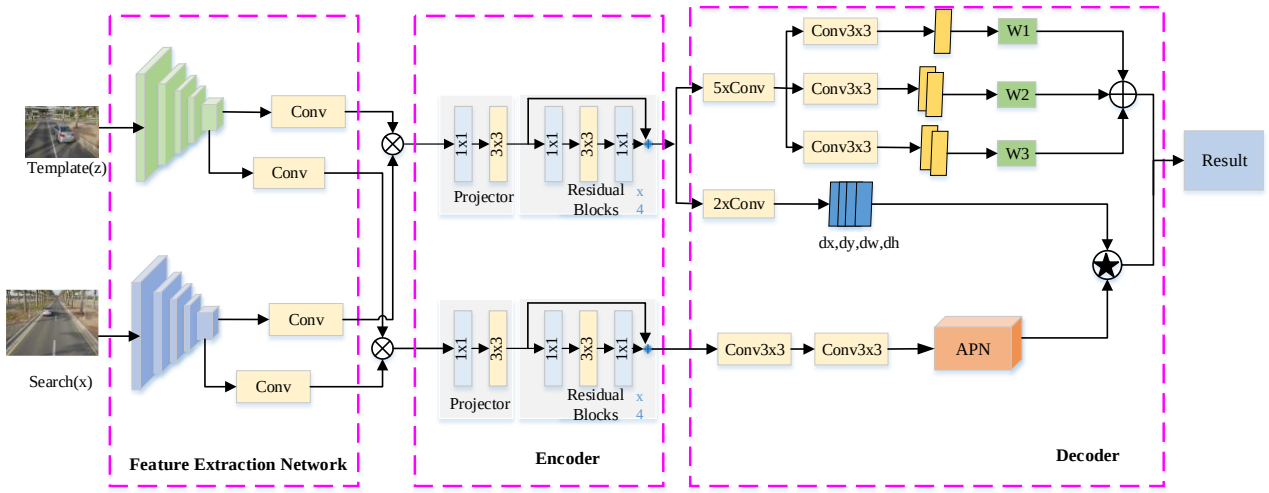


Figure 1 The General Framework

3.To reduce the hyperparameters and the computational complexity, the adaptive anchor points the are introduced in the decoder, which improves the robustness of the tracker.

II. Related Work

Generally, the CF-based trackers can be widely used to the UAV field owing to the high efficiency and scalability[2]. Nevertheless, manual features, such as color features and gray scale features, required by this method need to have strong prior information, which makes it difficult to learn and exploit deeper features.

Currently, Siamese network-based network shows great potential in the field of object tracking. SiamFC[9] proposes an end-to-end tracking method that aims to utilize full Convolutional Neural Networks (CNN) to calculate the similarity between template and search frames. SiamRPN++[10] adopts a deeper feature fusion method to extract semantic information, which further improves the tracking effect. DaSiamRPN[11] proposes a new training method that significantly improves the tracking performance. Although good tracking performance has been achieved, the efficiency of the anchor-based trackers still needs to be further improved, and these models have sample imbalance problems. To further optimize the algorithm, SiamFC++[12] is proposed accordingly. The scalability of the trackers is improved by redesigning the regression branch. However, the effects of the sample imbalance remain. SiamAPN[13] proposes a new approach that employs an a priori structure of the anchor proposed network to increase the proportion of positive samples. However, it increases the complexity and the training time of the model due to using multi-level features for attention fusion.

Recently, the multi-scale feature fusion has received much attention in many areas. In the target detection and segmentation, Libra R-CNN[14] proposes IoU-balanced sampling to solve the imbalance between samples and balance feature pyramid to strengthen multi-level features. To cutting down the high-level and low-level feature fusion path, Shu[15] develops the backbone network structure and strengthens the feature pyramid structure. They also propose a more flexible RoL pooling method to achieve richer semantic information. However, they are effective in the case of multi-scale changes of targets.

In this paper, we propose an Encoder structure, which can guarantee the robustness of the tracker when the target has multiscale changes in the tracking task. In addition, the anchor proposed network (APN)[12] structure in Decoder effectively improves the anti-interference and efficiency of the tracker.

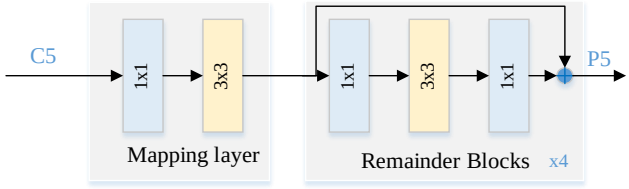


Figure 2 Structural description of the encoder.

III. Proposed Method

In this section, we will describe the proposed algorithm in detail through three blocks. As shown in Figure 1, the algorithm contains three parts: the feature extraction network, Encoder, and Decoder.

A. *The feature extraction network*

As shown in Figure 1, the feature extraction network is composed of a Siamese Network. It includes a reference frame and a search frame. The input to the reference frame is a template image and the input to the search frame is a search image. Generally, the template frame is represented as z and the search frame is represented as x. The AlexNet contains five convolution blocks, and x and z are output φ_5 (x) and φ_5 (z). The features extracted by the convolution of the fifth layer contain enough semantic information.

B. *Encoder*

The encoder structure is shown in Figure 2. The proposed algorithm uses the Encoder structure to replace the Feature

Pyramid Network (FPN)[16]. The FPN can be regarded as a Multiple-in-Multiple-out encoder. In contrast, the proposed encoder is a Single-in-Single-out structure.

The proposed encoder contains two parts: The Mapping layer and the Remainder blocks. The Mapping layer firstly adopts the 1×1 convolutional layer to cut down the number of the channel, and then uses the 3×3 convolutional layer to enhance the semantic information of the context[16]. There are four consecutive remainder blocks in the Encoder, which have different expansion rates. The Remainder blocks is composed of three series convolutional layers. The first is 1×1 convolution for channel reduction. The second expands the receiving domain with a 3×3 expanding convolution. Finally, the second 1×1 convolution is adopted to return to the number of channels. And we sequentially stack four remainder blocks with different expansion rates in order to further expand the receptive filed and extract the scale feature information with multiple receiving domains.

Using expansion convolution to expand the feature's receptive is a very effective method in object detection and object tracking[17]. For example, TridentNet[18] uses this strategy to enhance expression of multi-scale features. It utilizes a multi-branching structure and weight sharing mechanism to overcome the scale change problem of object detection. In order to enhance the feature expression ability and extracts more target feature information. DetNet[19] also adopts dilated residual blocks. In contrast, our method aims to make full use of contextual information and generate features with multiple receptive fields. The proposed encoder enables our method to fully extract the different multi-scale feature information, thus further improving the tracking performance.

C. Decoder

The Decoder consists of an Anchor Proposal Network (APN) and a multi-classification & regression network.

1) Anchor Proposal Network

Inspired by the baseline tracker SiamAPN++[20], we introduce the new APN structure in this paper. In the baseline algorithm, cross-ANN is designed in the APN structure to maintain the cross interdependent similarity of the two features. Different from the baseline tracker[20], the features extracted by convolution in the fifth layer of the backbone network are sent to the encoder, and then sent to the APN structure.

In this paper, the APN structure reduces the number of anchors and takes full advantage of these anchors. As shown in Figure 1, φ_5 (x) and φ_5 (z) are convolved with the kernel function of 3×3 and then passed into the encoder. APN defines the features extracted from the encoder as the suggested anchor point. Using the convolutional operator, we can get more comprehensive and stable anchors points. Unlike the Anchor-based approach, we select automatically positive and negative samples with adaptive anchor strategy. The proposed method needs only less hyper-parameters in different scenarios.

The APN is used to generate corresponding anchor points for the similarity feature map. It can be represented to the search patch for every position on the proposed anchor map $E(p, q,:)$. Such as, setting the location (p, q) of the proposed anchor corresponds to the position (kp, kq). The (kp, kq) is the center of the receive field of (p, q).

The upper-left and lower-right corners of the ground truth bounding box are denoted by (g_{x1}, g_{y1}) and (g_{x2}, g_{y2}) respectively, and the regression labels of the proposed anchors $w_{(p,q)}$ can be computed by:

$$w^{0}_{(p,q)} = k_p - g_{x1},\ w^{1}_{(p,q)} = k_q - g_{x1} \tag{1}$$

$$w^{2}_{(p,q)} = g_{x2} - k_p,\ w^{3}_{(p,q)} = g_{y2} - k_q \tag{2}$$

Obviously, the center point far away from the target easily leads to the drift of the tracking frame, and thus, affecting the robustness of the model. In order to obtain more accurate data, the quality weight ω is introduced into the loss function of the regression sub branch. the center point far away from the target is prone to producing inaccurate bounding boxes, degrading the performance of the tracking system. The loss of APN was included in the regression loss.

2) Multi-classification & regression network

Inspired by the baseline tracker SiamAPN++[20], we also adopted the similar classification branches. As shown in Fig.1, the three classification branches export three corresponding feature maps, which are $M^{cls1}_{w\times h\times 2}$, $M^{cls2}_{w\times h\times 2}$ and $M^{cls3}_{w\times h\times 2}$. Combining these classification branches, the total loss function of the classification branches is computed result as follows[20]:

$$L_{cls} = \lambda_{cls1}L_{cls1} + \lambda_{cls2}L_{cls2} + \lambda_{cls3}L_{cls3} \tag{3}$$

The L_{cls1} and L_{cls2} are the cross-entropy loss functions, $Lcls3$ indicates the binary cross-entropy loss function, and λ_{cls1} λ_{cls2} λ_{cls3} are the weight coefficient of these three classification branches respectively.

In order to improve the performance of the tracker, the regression branch was redesigned. The $M^{loc}_{w\times h\times 4}$ indicates a regression feature map of the regression branch outputs. Where, we represent the label of regression as a four-dimensional vector ($\tilde{v}^{0}_{(p,q)}$, $\tilde{v}^{1}_{(p,q)}$, $\tilde{v}^{2}_{(p,q)}$, $\tilde{v}^{3}_{(p,q)}$). Let g_x, g_y represent the central point of the ground truth box and g_w, g_h represent the scale information of the ground truth box. The range between them is calculated as[20]:

$$\tilde{v}^{0}_{(p,q)} = \frac{g_x - t_x}{t_x},\ \tilde{v}^{1}_{(p,q)} = \frac{g_y - t_y}{t_y} \tag{4}$$

$$\tilde{v}^{2}_{(p,q)} = \ln\frac{g_w}{t_w},\ \tilde{v}^{3}_{(p,q)} = \ln\frac{g_h}{t_h} \tag{5}$$

Let $t' = (t_x, t_y, t_w, t_h)$ indicates the central point and dimensions of the anchor point, it can be expressed as[20]:

$$t' = \psi(E(p,q,:),t) \tag{6}$$

Where ψ represents the transformation operation, that is, the proposed anchor is obtained by counting the offset E (p, q, :) and primeval central t. In order to achieve accurate regression, we use two loss functions that are the smooth L1loss and IoUloss for regression, so the loss function of calculation of regression is presented as follows:

$$L_{loc} = \lambda_{loc} L1(M^{loc}(p,q,:), \tilde{v}_{(p,q)}) - (1 - S_{ious}) * (\beta - S_{ious}) * \log(S_{ious}) \tag{7}$$

The β is the hyper-parameter, mainly representing the propensity towards positive and negative samples. And S_{ious} represents the IoU fraction, which is calculated between the proposed anchors and real frame.

IV. Experiment

A. Implementation details

In this section, the proposed model is comprehensively tested on three mostly used benchmarks, i.e., UAV123[21], UAV20L[21] and UAV123@10fps[21]. The UAV123[21] includes 8 other well-known target tracking algorithms, i.e., SiamRPN[22], SiameAPN++[20], ECO[23], ECO-HC[24], UPDT[25], GCT[26], CCOT[27] and DaSiamRPN[11]. Meanwhile the UAV20L[21] and UAV123@10fps[21] include 7 other well-known target tracking algorithms, i.e, ECO[23], SiamAPN++[20], SiamFC++[12], SiamAPN[13], DaSiamRPN[11], SiamRPN++[10] and SiamFC[9]. The model in this paper adopts offline training. During the training, the learning rate is initially set to 5×10^{-4}. Meanwhile, we set the weight decay to 1×10^{-4}.

The platform and parameters used in this paper are shown in table 1.

Table 1 Experimental platform parameters

Operating system	Processor	Implementation software	Framework implementation
Linux	1080XTi 64GB	Python	Pytorch

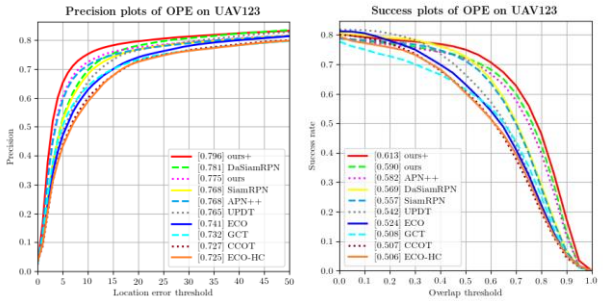
B. Evaluation metrics

The UAV dataset defines two evaluation indicators, namely, the average precision plot and the average success plot. Particularly, the average success rate is calculated by the boundary box coincidence rate. And all trackers are sorted due to the area under the curve (AUC) of the average success rate map. A widely applied evaluation indicator of the average precision map is the center location error (CLE). It refers to the percentage of video frames whose distance between the predicted boundary box center position and the target real box center position is less than threshold.

C. Results on UAV benchmarks

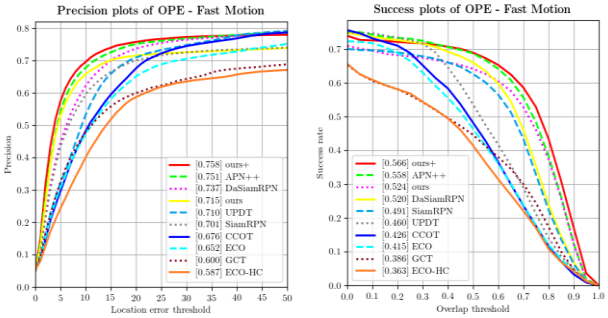
Compared with other trackers, the proposed tracker of this paper has achieved excellent results on three classical datasets. The algorithm named "ours" adopts AlexNet as the backbone network and the algorithm named "ours+" adapts GoogLeNet as the backbone network.

UAV123: UAV123 contains 123 aerial sequences at a low altitude, with a total of more than 110,000 frames, which can be used to test the comprehensive adaptability of the tracker from the perspective of the UAV. The overall average performance index is shown in Fig.3. From Fig.3, we can know that the proposed method "ours+" outperforms other SOTA trackers in terms of precision (0.796) and success (0.613). And the proposed method "ours" ranks third in terms of precision (0.775), while ranks second in terms of success (0.590). UAV123 includes 12 challenging scenarios, including Aspect Ratio Change (ARC), Background Clutters (BC), Camera Motion (CM), Fast Motion (FM), Full Occlusion (FO), Illumination Variation (IV), Low Resolution (LR), Out-of-View (OV), Partial Occlusion (PO), Scale Variation (SV), Similar Object (SO), and Viewpoint Change (VC). Three of these performance metrics are shown in Fig.4-6. In most scenarios, the trackers of "ours" and "ours+" outperform the other well-known trackers.



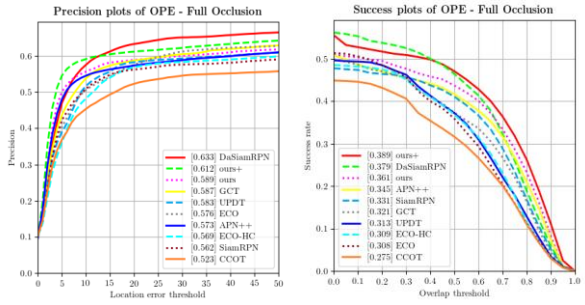
(a)Precision (b)Success plots

Figure 3 Results on UAV123



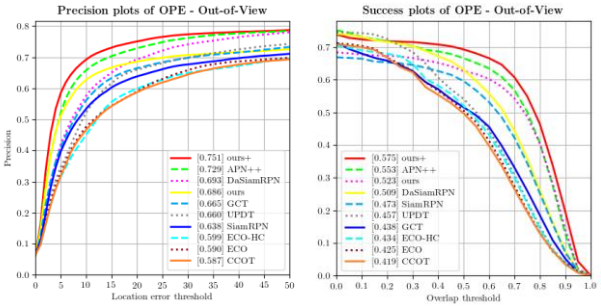
(a)Precision (b)Success plots

Figure 4 Fast Motion attribute graph



(a)Precision (b)Success plots

Figure 5 Full Occlusion attribute graph



(a)Precision (b)Success plots

Figure 6 Out-of-View attribute graph

Table 2 RESULTS ON UAV123@10FPS

Trackers	Precision(%)	Success(%)
ours+	76.6	59.1
ours	77.4	59.4
SiamAPN++	76.4	58.0
SiamAPN	75.2	56.6
SiamRPN++	73.5	55.1
SiamFC++	74.5	57.6
DaSiamRPN	69.2	48.3
SiamFC	68.0	47.3
ECO	67.5	43.0

Table 3 RESULTS ON UAV20L

Trackers	Precision(%)	Success(%)
ours+	75.0	58.5
ours	68.0	52.7
SiamAPN++	73.6	56.0
SiamAPN	72.1	53.9
SiamRPN++	69.6	52.8
SiamFC++	69.5	53.3
DaSiamRPN	66.5	46.5
SiamFC	59.5	40.2
ECO	58.3	36.3

UAV123@10fps: UAV123@10fps contains 123 sequences, selected from the video sequence of 30FPS. When the interval between frames is more than 30 FPS, the challenge difficulty of most scenes increases, such as Scale Change and Low Resolution. Consequently, in order to evaluate the trackers robustness more comprehensively, we also conducted experimental research on the UAV123@10fps. The comparison results are shown in Table. 2. “Ours” ranks first in terms of precision (0.774) and success (0.594), meanwhile “ours+” ranks second in terms of precision (0.766) and success (0.591).

UAV20L: UAV20L contains 20 long-time sequences and a variety of challenging scenes, such as, Fast Motion, Camera Motion, Occlusion, and Scale Change. These sequences average 2934 frames each. Consequently, we conduct experiments on the UAV20L to test the stability of the proposed tracker for long-term tracking. As shown in the Table. 3. In terms of precision, “ours+” ranks the highest with a score of 0.750, and in terms of success rate, it also ranked first with a score of 0.585, which is far superior to other tracking algorithms.

We conduct analysis of the experimental results. Fig.7 demonstrates the bounding box prediction outcome of our proposed methods on the UAV123, and the proposed models can be intuitively obtained outperform SiamAPN++[20] and DaSiamRPN[11], in the challenging scenes of Fast Motion, Scale Variation and so on.

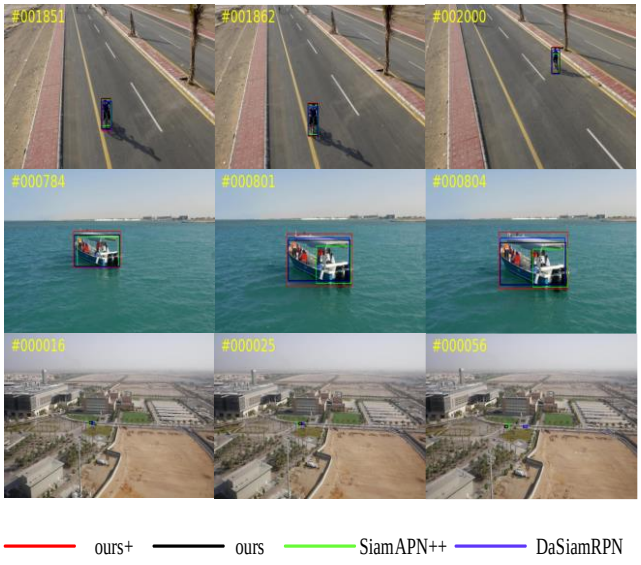


Figure 7 Comparison between the proposed trackers and state-of-the-art screenshots of bike1, boat3, and car14.

V. CONCLUSION

In this work, in order to further improve the robustness and efficiency of UAV tracker in the complex scenes, a new UAV tracking algorithm based on the Siamese network is proposed. A multi-scale receptive field structure is designed in the encoder to deal with the matters of multi-scale change. The design of the adaptive anchor in the decoder effectively reduces the relevant hyperparameters. The tracker designed in this paper has generalization and efficiency in different scenarios. Experimental results on three benchmark datasets indicate that the proposed algorithm has effective utility and generalization.

REFERENCES

[1] J. Wang et al., "Parallel Vision for Long-Tail Regularization: Initial Results from IVFC Autonomous Driving Testing," in IEEE Transactions on Intelligent Vehicles, vol. 7 (2), 2022, pp. 286-299.

[2] Xiumin Zhu, Lingling Wang, et al., "Path planning of multi-UAVs based on deep Q-network for energy-efficient data collection in UAVs-assisted IoT", Vehicular Communications, Vol. 36, 2022.

[3] Efstratios Kakaletsis, Ioannis Mademlis, et al. "Multiview vision-based human crowd localization for UAV fleet flight safety", Signal Processing: Image Communication, Vol.99, 2021.

[4] Xin Yang, Yong Song, Zishuo Zhang, et al. "Robust correlation filter tracking based on response map analysis network", Signal Processing: Image Communication, Vol.108, 2022.
[5] Hao Z, Liu G, Gao J, et al. "Robust Visual Tracking Using Structural Patch Response Map Fusion Based on Complementary Correlation Filter and Color Histogram". Sensors, Vol.19(19), 2019.
[6] S. Hu, L. Sun and H. Yu, "Accurate Visual Tracking with Attention Feature Fusion," 2021 26th International Conference on Automation and Computing (ICAC), 2021, pp. 1-6, doi: 10.23919/ICAC50006.2021.9594244.
[7] H. Zhang, G. Luo, Y. Tian, K. Wang, H. He and F. -Y. Wang, "A Virtual-Real Interaction Approach to Object Instance Segmentation in Traffic Scenes," in IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 2, pp. 863-875, Feb. 2021, doi: 10.1109/TITS.2019.2961145.
[8] Zhaoyang Niu, Guoqiang Zhong, Hui Yu, A review on the attention mechanism of deep learning, Neurocomputing, Vol.452, 2021, pp. 48-62.
[9] L. Bertinetto, J. Valmadre, and et al, "Fully-Convolutional Siamese Networks for Object Tracking," in Proceedings of ECCV, 2016, pp. 850–865.
[10] B. Li, W. Wu, and et al., "SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks," in Proceedings of CVPR, 2019, pp. 4277–4286.
[11] Z. Zhu, Q. Wang, and et al., "Distractor-Aware Siamese Networks for Visual Object Tracking," in Proceedings of ECCV, 2018, pp. 101–117.
[12] Y. Xu, Z. Wang, Z. Li, Y. Yuan, and G. Yu, "SiamFC++: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines," in Proceedings of AAAI, 2020, pp. 12549–12556.
[13] C. Fu, Z. Cao, Y. Li, J. Ye, and C. Feng, "Siamese Anchor Proposal Network for High-Speed Aerial Tracking," in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2021, pp. 1–7.
[14] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In Proceedings of CVPR, 2019, pp. 821–830.
[15] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In Proceedings of CVPR, 2018, pp. 8759–8768.
[16] Tsung-Yi Lin, Piotr Doll´ar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2117–2125, 2017. 1, 2, 4, 7.
[17] Lifang Zhou, Yu He, Weisheng Li, et al. "IoU-guided Siamese region proposal network for real-time visual tracking", Neurocomputing, Vol.462, 2021, pp. 544-554.
[18] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In Proceedings of the IEEE international conference on computer vision, 2019, pp 6054–6063.
[19] Zeming Li, Chao Peng, Gang Yu, et al. Detnet: Design backbone for object detection. In Proceedings of ECCV, 2018.4, pp 334–350.
[20] Cao Z, Fu C, Ye J, et al. SiamAPN++: Siamese attentional aggregation network for real-time uav tracking[C]//2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2021: 3086-3092.
[21] M. Mueller, N. Smith, and B. Ghanem, "A Benchmark and Simulator for UAV Tracking," in Proceedings of ECCV, 2016, pp. 445–461.
[22] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High Performance Visual Tracking with Siamese Region Proposal Network," in Proceedings of CVPR, 2018, pp. 8971–8980
[23] Danelljan, M., Robinson, A., et al.." Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking," in Proceedings of ECCV, 2016, pp. 472–488.
[24] Danelljan, M., Bhat, G., et al. "ECO: Efficient Convolution Operators for Tracking," in Proceedings of CVPR, 2017, pp. 6638–6646.
[25] Bhat, G., Johnander, J., Danelljan, M., et al. "Unveiling the Power of Deep Tracking," in Proceedings of ECCV, 2018, pp. 483–498.
[26] Gao, J., Zhang, T., Xu, C. "Graph Convolutional Tracking," in Proceedings of CVPR, 2019, pp. 4649–4659.
[27] Danelljan M, Robinson A, Shahbaz Khan F, et al. Beyond correlation filters: Learning continuous convolution operators for visual tracking[C]//European conference on computer vision. Springer, Cham, 2016: 472-488.