Siamese Network based on MLP and Multi-head Cross Attention for Visual Object Tracking*

Piaoyang Li[1], Shiyong Lan*[1], Shipeng Sun[1], Wenwu Wang[2],
Yongyang Gao[1], Yongyu Yang[1], and Guangyu Yu[1]

[1] College of Computer Science, Sichuan University, Chengdu, China
[2] University of Surrey, Guildford, GU2 7XH, United Kingdom

Abstract. Visual object tracking is an important prerequisite in many applications. However, the performance of the tracking system is often affected by the quality of the visual object's feature representation and whether it can identify the best match of the target template in the search area. To alleviate these challenges, we propose a new method based on Multi-Layer Perceptron (MLP) and multi-head cross attention. First, a new MLP-based module is designed to enhance the input features, by refining the internal association between the spatial and channel dimensions of these features. Second, an improved head network is constructed for predicting the location of the target, in which the multi-head cross attention mechanism is used to find the optimal matching between the template and the search area. Experiments on four datasets show that the proposed method offers competitive tracking performance as compared with several recent baseline methods. The codes will be available at https://github.com/SYLan2019/MLP-MHCA.

Keywords: Visual Object Tracking · Siamese Network · Attention.

1 Introduction

Visual object tracking is an active field of research in computer vision. Traditional tracking algorithms are based on either generative or discriminative models. In generative algorithms [1], the target features are extracted for constructing an appearance model and then matched with those from the searching area. However, the performance of the generative models degrades in a complex environment, in the presence of illumination changes and occlusions. The discriminative models [9,17,5,31,30,4] convert the tracking problem into a classification and localization problem. To effectively identify the target in the search area, it is crucial to obtain robust and accurate feature representations for the targets.

One of the most classic discriminative models is the Siam network [2], which simply translates the tracking problem into a problem of learning the matching

* This work was funded by 2035 Innovation Pilot Program of Sichuan University, China. * Corresponding author. E-mail:lanshiyong@scu.edu.cn.

between the target template and the search area. Siamese-based trackers usually consist of a backbone network and a head network, in which the backbone network is used to extract features, and the head network is used for target classification and localization. Existing Siamese-based trackers can be divided into convolutional neural networks (CNN)-based Siamese trackers [2,18] and Transformer-based Siamese trackers [6,5,7]. In CNN-based Siamese trackers [2], the target template is used to match with the search area through sliding convolution, and the area with the maximum response value is then obtained as the target position. However, during the correlation operation, CNN-based Siamese trackers tend to give locally optimal solutions, as the correlation operation itself is a local linear matching process[5]. In Transformer-based Siamese trackers [5,7], the correlation operation is replaced with Transformer, which can prevent the algorithm from converging to local minimum with the global information extracted from the image, but the entire transformer architecture leads to a high computational load.

Recently, several studies have been performed on replacing transformers with more efficient methods. For example, Tolstikhi et al [27] advocated that using only MLP can achieve the same performance as using transformers in visual classification tasks, but with significantly improved computational efficiency. Motivated by this work, we introduce MLPs in our tracking task to enhance feature representations, in order to improve the discrimination of the target in the search area for target localization. In addition, we design a simple cross attention module followed by another MLP for predicting the location of the target, rather than using a correlation operation between the template and the search area as used in previous work [2].

Our main contributions can be summarized as follows:

- We propose a new tracking framework based on MLPs and multi-head cross attention. In our proposed framework, a new MLP-based module is designed to enhance the feature representation by associating the channel and the spatial information within the input features. Our ablation experiments show the effectiveness of our modification.
- An improved head network is constructed with simple multi-head cross attention instead of using a conventional correlation filter for predicting the position of the target in a local search area.
- Extensive experiments on the OTB2015, UAV123, NFS, and VOT2020 datasets show that the proposed method outperforms the compared baselines.

2 Proposed Approach

2.1 Architecture

The architecture of the Siamese network that we propose for object tracking can be divided into three parts: backbone for feature extraction, neck network for feature enhancement, and head network for classification and regression, as

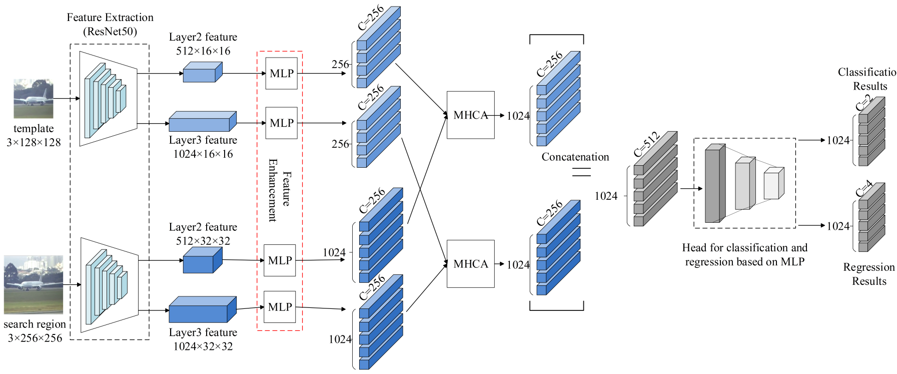


Fig. 1. The framework of the proposed Siamese network based on MLPs and multi head cross attention.

shown in Fig. 1. Unlike TransT [5], our backbone Resnet50 extracts the features of different layers, including semantic information from deep layers and textural information from shallow layers, that are helpful for improving the performance of classification and regression [6]. In addition, we utilize a feature enhancement module based on MLPs to extract features at different scales and improve the ability of the network in feature representation. In the head network, we design an efficient multi-head cross attention to overcome local optimization problems of target localization in the search area.

2.2 Feature Enhancement Module Based on MLP

Convolution pays more attention to the spatial information on the feature map and ignores the information on the channel dimension. Here, we propose a feature enhancement network based on multi-layer perceptron to enhance the features. As shown in Fig. 2, the features extracted by the backbone are first passed through layer normalization, then they are reshaped and input into a multi-layer perception composed of two fully-connected layers and a GeLU [14] activation layer. Then, we reshape them and feed them into the next multilayer perceptron. That is, the spatial feature is extracted first, followed by the channel-wise feature. As shown in Fig. 1, we extract the features of the second and third layers in Resnet50, whose channel dimensions are 512 and 1024, respectively. To ensure that the feature dimensions from different layers are consistent to enable feature fusion, we chose to employ two fully-connected layers to lower the channel dimension. In this way, we can aggregate spatial features to achieve feature enhancement, in addition, we can reduce the computational cost to some extent.

2.3 Head Network Based on Multi-head Cross Attention (MHCA)

Existing Siamese network trackers use correlation operations [31] to find similarity between the template and the search area. However, the usual correlation

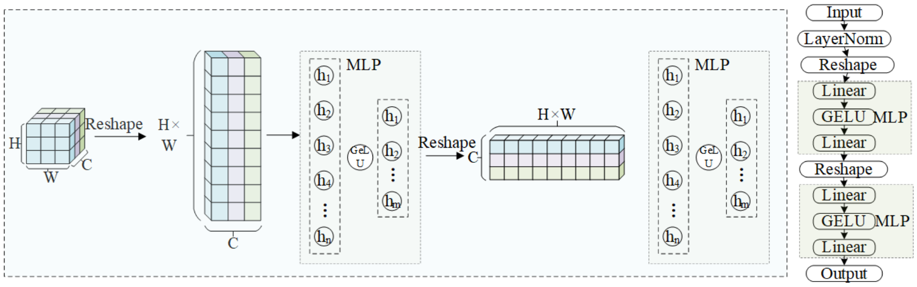


Fig. 2. Feature enhancement network based on MLP. The left figure shows the core idea, while the right shows the general process. The *linear* operation in the right figure corresponds to a full-connected layer in MLP. In addition, channel interaction is achieved in each Mixer operation by applying one-dimensional global average pooling and one-dimensional fully connected layers. The global average pooling captures global statistical information for each channel, while the fully connected layers in Mixer operation learn weights to model the relationships between channels.

operation itself is a local linear matching process[5], which is prone to taking the local optimal matching as the final result. To solve this problem, we introduce a multi-head cross attention module to identify the similarity between them and retains semantic features, as shown in Fig. 3. To determine the relationship between the features from the template and the features from the search region, we use the search region feature of dimension 1024×256 as K and V, and the template feature of dimension 256×256 as Q. Then, Q and K are passed through two different fully-connected layers, and then multiplied with each other. After that, the shape becomes 1024×256, then this result is mapped to the V vector (i.e., by multiplying the matrix with V). Finally, it is passed through the fully-connected layer to adjust the number of channels. In this way, we replace the correlation operations with the multi-head cross attention mechanism. The following is the initial cross-attention formula:

$$SelfAttention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

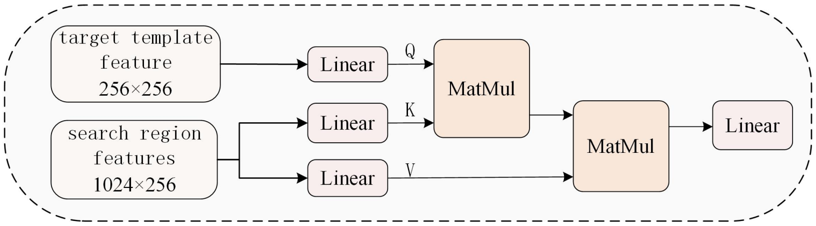


Fig. 3. The cross attention module.

To construct the head networks, we use two cross-attention structures in our approach. The results are spliced through the channels to decrease the number of parameters. The formula is given as:

$$MultiHead(Q, K, V) = Linear(Concat(H_1, H_2)) \tag{2}$$

where $Concat(\cdot)$ means concatenate operation, H_1 and H_2 are the output of the MHCA module. The $Linear(\cdot)$ operation is achieved with a fully-connected layer that controls the number of channels.

2.4 Adaptive Multi-Layer Feature Fusion

We extracted both shallow location information and deep semantic information for the tracking task in order to improve the performance of the model. The features from shallow layers contain detailed textural information which are suitable for localization, while the features from deep layers contain semantic information which are useful for classification [6,18]. As shown in Fig. 1, we extracted the layer2 and layer3 features from Resnet50. As a result, the features of the target template have a shape $512 \times 16 \times 16$ and $1024 \times 16 \times 16$. Both become $256 \times 16 \times 16$ after they are passed through the feature enhancing network. For the feature from the search region, its shape will eventually become $256 \times 32 \times 32$. In a typical feature fusion method, the shallow and deep features are concatenated before applying dimension reduction with convolution layers. Unlike this method, we treat them independently first, then splice them, as indicated in the formula:

$$\begin{aligned} &P^{cls}_{w\times h\times 2} \quad , \quad P^{reg}_{w\times h\times 4} \\ &= MLP([mHead_1(f_{t1}, f_{s1}), mHead_2(f_{t2}, f_{s2})]) \end{aligned} \tag{3}$$

The shallow and deep features of the target template are denoted by f_{t1} and f_{t2}, while the shallow and deep features of the search area are denoted by f_{s1} and f_{s2}, respectively. $mHead$ is a multi-head cross attention based head network proposed in this paper, and the symbol $[\cdot]$ denotes the channel splicing operation. P^{cls} and P^{reg} represent the prediction results of classification and regression, respectively.

3 Experiments

3.1 Experimental Setup

We train our model on four typical datasets including COCO [19], LaSOT [10], GOT-10K [15], and VOT2020 [16]. ImageNet [25] pretrained Resnet-50 [13] is used to initialize the parameters of the backbone, whereas Xavier init [12] is used to initialize the remaining parameters in our model. We employed two RTX3090 GPUs to train our model, with 10^{-5} as the learning rate for the backbone, and 10^{-4} for the others. The default batch size is 36, with each epoch having 1000 iterations and a total of 500 epochs. We use AdamW [20] as the optimizer. The number of heads in multi-head cross attention is 8. The number of channels in the hidden layer was set to 2048.

Table 1. Evaluation results on OTB2015, NFS and UAV123. Red and blue represent the top two track results, respectively. The symbol - is used to denote that the corresponding test results are not included in the official model.

Trackers	Years	OTB2015		UAV123			NFS
		AUC	Prec.	AUC	P	NP	AUC
DaSiamRPN [31]	2018	0.650	0.880	0.568	0.796	-	-
SiamRPN++ [18]	2019	0.696	0.914	0.613	0.807	-	-
DiMP [3]	2019	0.688	0.900	0.597	0.152	0.441	0.620
SiamBAN [6]	2020	0.696	0.910	0.597	0.178	0.452	0.594
STARK [29]	2021	0.696	-	0.692	0.882	0.660	0.652
KeepKtack [22]	2021	0.709	-	0.697	-	-	0.664
TransT [5]	2021	0.696	-	0.691	0.876	0.694	0.657
RTS[24]	2022	-	-	0.676	0.894	0.816	0.654
ToMP [21]	2022	0.701	-	0.690	-	-	0.669
Mixformer [7]	2022	0.700	0.929	0.687	0.895	-	-
Ours	-	0.701	0.909	0.701	0.898	0.703	0.671

Table 2. Evaluation results on VOT2020. Red and blue represent the best two results respectively. The symbol - is used to denote that the corresponding test results are not included in the official model.

	Ours	Mixformer [7]	ToMP [21]	RTS [24]	CSWinTT [26]	TransT [5]	STARK50 [29]	D3S [16]	ATOM [8]	DiMP [3]
		2022	2022	2022	2022	2021	2021	2020	2019	2019
EAO	0.509	0.527	0.297	0.506	0.304	0.293	0.308	0.439	0.271	0.274
Accuracy	0.723	0.746	0.453	0.710	0.480	0.477	0.478	0.699	0.462	0.457
Robustness	0.828	0.833	0.789	0.845	0.787	0.754	0.799	0.769	0.734	0.734
$\varDelta$ EAO to ours	-	**↓0.018**	**↑0.212**	**↑0.003**	**↑0.205**	**↑0.216**	**↑0.201**	**↑0.07**	**↑0.238**	**↑0.235**

3.2 Results and Analysis

We perform test of the model on several datasets, including OTB2015 [28], NFS [11], UAV123 [23] and VOT2020 [16]. Our model was trained on the LaSOT [10], GOT-10K [15] and COCO [19] datasets before the testing. We compare our method with state-of-the-art (SOTA) tracking algorithms qualitatively and quantitatively on OTB2015, NFS and UAV123 datasets. Table 1 shows AUC, Precision, Norm Precision results. It can be observed that our model outperforms all other methods on the UAV123 and NFS datasets while achieving competitive results on the OTB2015 dataset. Compared with TransT [5] and other SOTA methods, the AUC and precision scores in OTB2015 are both increased by 1.3% and 1.7%, respectively. On the NFS dataset, the AUC score is increased by 2.1%. On the UAV123 dataset, the precision score is increased by 2.5%. Furthermore, the NFS and UAV123 datasets involve more background clutter and camera viewpoint change. Our method achieves better performance on these two datasets than SOTA baselines. This demonstrates that the feature enhance module in our model can effectively improve the tracking robustness against the changes in visual attributes. In addition, the tracking speed of our model is about 40 FPS, which can meet the requirement of real-time tracking. These results show that the proposed method achieves competitive performance as compared with the SOTA baselines.

Fig. 4 shows a comparison between our algorithm, SiamBan [6] and TransT [5] for helicopter tracking. The red, green, blue and cyan boxes represent the ground-truth position, and the position estimated by the proposed method, SiamBan and TransT, respectively. In total, eight frames are selected. The helicopter video sequence represents a challenging case with scale change of the target. It can be found that our algorithm can still accurately predict the location and size of the target when the target scale has changed.

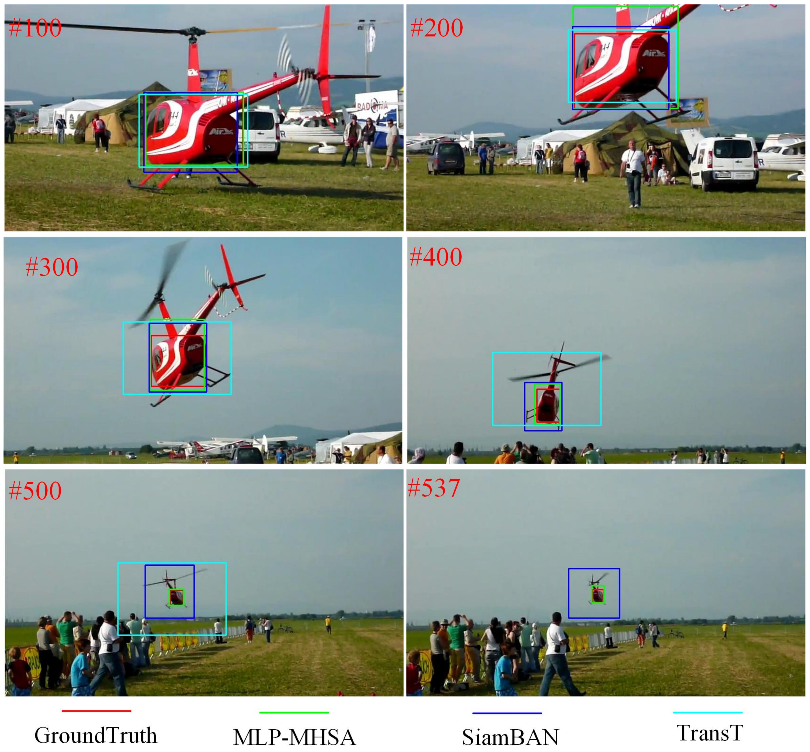


Fig. 4. The results of the methods in tracking the helicopter. The number in the upper left corner of each picture is the frame number in the helicopter video sequence.

We also evaluated our model on the VOT2020 [16] short-term tracking challenge, and compared it with recent trackers. The results are shown in Table 2, where EAO, A, R are classic tracking performance indicators used on this dataset, representing Expected Average Overlap, Accuracy and Robustness respectively. The proposed method offers a higher EAO score, as compared with RTS [24], CSwinTT [26] and ToMP [21], just a little less than Mixformer [7]. In addition, our method ranks second and third in terms of Accuracy and Robustness indicators, respectively. However, the proposed method gives a much higher EAO score than TransT [5], reaching 50.9%. All these experimental results demonstrate that our method can achieve competitive performance in short-term tracking challenge.

In addition, our proposed method can alleviate the computational limitations of the transformer module in object tracking to a certain extent. Table 3 shows that our method is more effective than TransT [5] and Mixformer [7]. Although MLP operations can also be computationally intensive, our approach employs a dimensionality reduction process in the MLP module, which enhances feature representation and reduces computation (as shown in Figure 5, the computational cost is adjusted by controlling the scaling ratio of the output to the input in the FC layer). Therefore, our approach presents a promising alternative to the transformer module in terms of efficiency and effectiveness for object tracking.

Table 3. The comparison of the size of the three methods. "Linear" refers to the fully connected layer

method	TrasnT	Mixformer	MLP-MHCA
Number of parameters(MB)	23.0	35.1	20.0

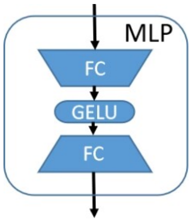


Fig. 5. The internal structure of an MLP.

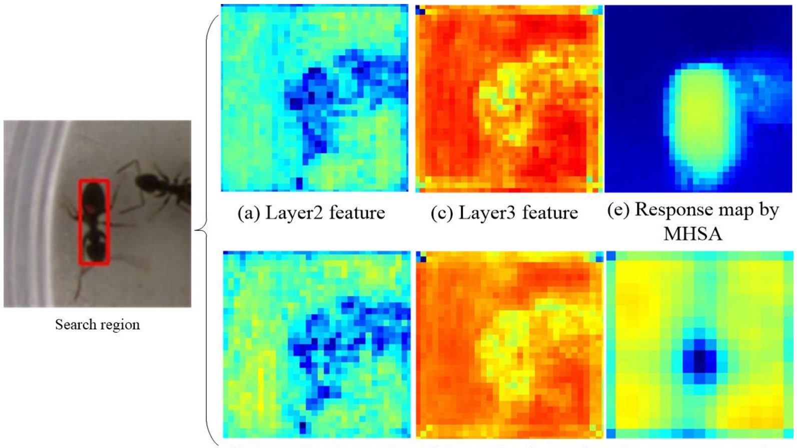
3.3 Ablation Study

We carried out ablation experiments on the VOT2020 [16] dataset to demonstrate the effectiveness of various modules in the proposed model, including multi-head cross-attention, feature enhancement network based on multi-layer perceptron and multi-layer feature adaptive fusion. The ablation experiment results are shown in Table 4. Among them, mHead represents multi-head cross-attention, the correlation operation in Siamese network trackers is represented by Cor, layer2 represents the second layer feature of Resnet50, layer3 represents the third layer feature of Resnet50, and EN-MLP is the feature enhancement module based on multi-layer perceptron.

Experiment 1 represents our proposed method where the correlation operation is replaced with attention. Experiment 2 shows the outcomes of an experiment employing only the second layer of the Resnet50. The result of using only the third layer of Resnet50 is shown in experiment 3. Through the comparison of experiments 1, 2 and 3, it can be seen that the simultaneous selection of the second and third layers improves tracking performance. Experiment 4 shows

Table 4. Ablation studies on the VOT2020 dataset. ✓ means the component is used, while × means that it is not used in the model.

number	**Whether to use**						VOT2020		
	mHead	Cor	Layer2	EN-MLP	Layer3	EN-MLP	A	R	EAO
1	✓	×	✓	✓	✓	✓	0.723	0.828	0.509
2	✓	×	✓	✓	×	✓	0.652	0.787	0.466
3	✓	×	×	✓	✓	✓	0.676	0.809	0.483
4	×	✓	✓	✓	✓	✓	0.563	0.749	0.448
5	✓	×	✓	×	✓	×	0.573	0.722	0.446
6	✓	×	×	×	✓	×	0.545	0.730	0.421



(a) Layer2 feature

(c) Layer3 feature

(e) Response map by MHSA

(b) Layer2 enhanced feature

(d) Layer3 enhanced feature

(f) Response map by correlation operation

Fig. 6. Visualization results of feature map and response map. The two pictures in the first column show the features of Resnet Layer 2 that are not enhanced by MLP and the features of Resnet Layer 2 that are enhanced by MLP. The second column represents the performance results of Resnet Layer 3 features and MLP. The top of the third column represents the response map fused with the MHCA. The figure below represents feature map obtained by related operations.

the results of utilizing the traditional convolution operation rather than the attention mechanism. We can see that all the performance indexes decreased as compared with those in experiment 1, especially the Robustness and EAO, which showed that the multi-head cross-attention was helpful to improve the accuracy and robustness of the tracker. Experiment 5 did not use the MLP based feature enhancement module as in experiment 1, thus the performance scores obtained are also lower than those in experiment 1. In experiment 6, only the feature of the third layer of Resnet50 is used, and as a result, the accuracy and robustness scores are greatly reduced. This shows that the use of multi-layer features can improve the performance of the tracker.

Fig. 6 visualises the feature map and response map in the ablation experiment. The red box in the search area surrounds the object *ant*. The features from both layer 2 and layer 3 show the contour of the object *ant* and the nearby *ant*. With the help of the MLP based feature enhancement module, the contour of the object becomes clearer with less clutter in surrounding area, which helps mitigate the impact of interference on the tracker. The response map of multi-head cross-attention has a dimension of 32×32. The region with a high response value still lies around the object in the response map of multi-head cross-attention (see Fig. 6 (e)). However, as shown in Fig. 6 (f), the high response value of correlation operation has deviated from the target location. Therefore, the experimental results demonstrate that the multi-head cross attention is more beneficial to improve the tracking accuracy and robustness over the correlation operation.

To sum up, the proposed multi-head cross-attention, feature enhancement module and multi-layer feature adaptive fusion indeed improve the performance of the tracker.

4 Conclusions

We have presented a Siamese network based on multi-layer perceptron and multi-head cross attention for visual tracking. We studied a new paradigm of MLP as feature enhancement, and the use of multi-head cross attention to replace the correlation operation in the Siamese network. This enables the extraction of shallow location information and deep semantic information simultaneously while utilizing multi-layer perceptron to enhance the features. The experiments on the OTB2015, VOT2020, NFS and UAV123 datasets show the effectiveness of our proposed method, as compared with several SOTA baseline methods. In the future, we will further study visual object tracking by taking the temporal information into account.

References

1. Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T.: Atutorial on particle filters for online nonlinear/non-gaussianbayesian tracking. IEEE TSP **50**(2), 174–188 (2002)
2. Bertinetto, L., Valmadre, J., Henriques, J., Vedaldi, A., Torr, P.: Fully-convolutional siamese networks for object tracking. In: European Conference on Computer Vision. pp. 850–865 (2016)
3. Bhat, G., Danelljan, M., Gool, L., Timofte, R.: Learning discriminative model prediction for tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6182–6191 (2019)
4. Chen, B., Tsotsos, J.K.: Fast visual object tracking with rotated bounding boxes. arXiv preprint arXiv:1907.03892 (2019)
5. Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H.: Transformer tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8126–8135 (2021)
6. Chen, Z., Zhong, B., Li, G., Zhang, S., Ji, R.: Siamese box adaptive network for visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6668–6677 (2020)

7. Cui, Y., Jiang, C., Wang, L., Wu, G.: Mixformer: End-to-end tracking with iterative mixed attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13608–13618 (2022)
8. Danelljan, M., Bhat, G., Khan, F., Felsberg, M.: Atom: Accurate tracking by overlap maximization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4660–4669 (2019)
9. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Eco: Efficient convolution operators for tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6638–6646 (2017)
10. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: Lasot: A high-quality benchmark for large-scale single object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5374–5383 (2019)
11. Galoogahi, H.K., Fagg, A., Huang, C., Ramanan, D., Lucey, S.: Need for speed: A benchmark for higher frame rate object tracking. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1125–1134 (2017)
12. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. pp. 249–256 (2010)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
14. Hendrycks, D., Gimpel, K.: Bridging nonlinearities and stochastic regularizers with gaussian error linear units. arXiv.1606.08415 (2016)
15. Huang, L., Zhao, X., Huang, K.: Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(5), 1562–1577 (2019)
16. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., et al.: The eighth visual object tracking vot2020 challenge results. In: European Conference on Computer Vision. pp. 547–601 (2020)
17. Lan, S., Li, J., Sun, S., Lai, X., Wang, W.: Robust visual object tracking with spatiotemporal regularisation and discriminative occlusion deformation. In: IEEE International Conference on Image Processing. pp. 1879–1883 (2021)
18. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: Siamrpn++: Evolution of siamese visual tracking with very deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4282–4291 (2019)
19. Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision. pp. 740–755 (2014)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
21. Mayer, C., Danelljan, M., Bhat, G., Paul, M., Paudel, D., Yu, F., Van Gool, L.: Transforming model prediction for tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8731–8740 (2022)
22. Mayer, C., Danelljan, M., Paudel, D., Gool, L.V.: Learning target candidate association to keep track of what not to track. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13444–13454 (2021)
23. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for uav tracking. In: European Conference on Computer Vision. pp. 445–461 (2016)

24. Paul, M., Danelljan, M., Mayer, C., Van Gool, L.: Robust visual tracking by segmentation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII. pp. 571–588. Springer (2022)
25. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S.and Ma, S., Huang, Z., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
26. Song, Z., Yu, J., Chen, Y., Yang, W.: Transformer tracking with cyclic shifting window attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8791–8800 (2022)
27. Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., et al.: Mlp-mixer: An all-mlp architecture for vision. Advances in Neural Information Processing Systems **34** (2021)
28. Wu, Y., Lim, J., Yang, M.: Online object tracking: A benchmark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2411–2418 (2013)
29. Yan, B., Peng, H., Fu, J., Wang, D., Lu, H.: Learning spatio-temporal transformer for visual tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10448–10457 (2021)
30. Yu, Y., Xiong, Y., Huang, W., Scott, M.: Deformable siamese attention networks for visual object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6728–6737 (2020)
31. Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., Hu, W.: Distractor-aware siamese networks for visual object tracking. In: Proceedings of the European Conference on Computer Vision. pp. 101–117 (2018)