International Journal of Cognitive Informatics and Natural Intelligence
Volume 15 • Issue 4

Object Tracking Based on Global Context Attention

Yucheng Wang, School of Computer Science, Wuhan University, China
Xi Chen, School of Computer Science, Wuhan University, China
Zhongjie Mao, School of Computer Science, Wuhan University, China
Jia Yan, Department of Electrical Engineering, School of Electronic Information, Wuhan University, China

ABSTRACT

Previous research has shown that tracking algorithms cannot capture long-distance information and lead to the loss of the object. When the object was deformed, the illumination changed, and the background was disturbed by similar objects. To remedy this, this article proposes an object-tracking method by introducing the global context attention module into the multi-domain network (MDNet) tracker. This method can learn the robust feature representation of the object through the global context attention module to better distinguish the background from the object in the presence of interference factors. Extensive experiments on OTB2013, OTB2015, and UAV20L datasets show that the proposed method is significantly improved compared with MDNet and has competitive performance compared with more mainstream tracking algorithms. At the same time, the method proposed in this article achieves better results when the video sequence contains object deformation, illumination change, and background interference with similar objects.

KEYWORDS

Channel Attention, Computer Vision, Convolutional Network, Global Context Attention, Implicit Template, Multi-Domain Learning, Object Tracking, Spatial Attention

INTRODUCTION

Object tracking is a fundamental task in the field of computer vision. The tracker needs to accurately predict the object's position and size change in the subsequent video frames according to the object given in the first frame of the video sequence. Although much progress has been made in research on object tracking (Marvasti-Zadeh et al., 2021; Abbass et al., 2021), it is still a challenging task because the object is often disturbed by external factors, such as size change, illumination change, and occlusion (K. Zhang et al., 2014).

In recent years, object-tracking algorithms have mainly been divided into two categories according to the template type. The first type is based on an explicit template tracking algorithm, and the second type is based on an implicit template tracking algorithm. The tracking algorithms based on Siamese neural networks, such as the Fully-Convolutional Siamese Network (SiamFC; Bertinetto et al., 2016) and the Deeper and Wider Siamese Network (Zhang & Peng, 2019), are the representatives of the first kind of method. They take the object branch as the template and find the most similar region to the object in the search region through the cross-correlation method. The representative of the second kind of method is the Multi-Domain Network tracking algorithm MDNet (Nam & Han, 2016).

DOI: 10.4018/IJCINI.287595

The algorithm divides the convolutional neural network into a shared feature-extraction layer and a domain-specific layer. The shared layer conducts pre-training on the dataset to learn the common features of the object. In contrast, the domain-specific layer uses online training during the tracking process to generate a domain-specific visual tracking classifier. Because the implicit template-based method takes the surface model as the last full connection layer for online learning, the accuracy is usually better than that of the explicit template-based method based on the Siamese neural network.

However, MDNet's shared feature-extraction layer only has a few convolutional layers, so it cannot capture the remote dependency information. When there is deformation of the tracked object, illumination change, or interference of similar objects in the background, the shared feature-extraction layers cannot extract more discriminative and robust features of the tracked object from the global information, leading to a great loss of tracking algorithm accuracy. Therefore, the tracker proposed in this article introduces the Global Context module (Y. Cao et al., 2019) in MDNet and combines the complete Global Context attention module with the split Global Context attention module to better enhance the recognition of features. The Global Context attention module simplifies the autocorrelation operation in non-local modules by learning a shared attention map for all query positions in the feature map and adopts a method similar to Squeeze-and-Excitation Network (SENet; Hu et al., 2018) to obtain the attention between channels. This method not only captures space and channels attention at the same time, but also greatly reduces the number of additional parameters and computational complexity.

In summary, the main contributions of this work are the following:

1. This article introduces the Global Context module in the position where the size of the feature map and the number of channels are more balanced to make full use of both space and channel information to enhance the feature representation of the tracked object.
2. By splitting the Global Context module, this article integrates the global attention modeling part into the more prominent position of the feature graph to capture the remote information, and integrates the channel-attention modeling part into the region with more channels to select the features with differentiability and robustness.
3. The method proposed in this article has been extensively tested on three datasets, OTB2013 (Wu et al., 2013), OTB2015 (Wu et al., 2015), and UAV20L (Mueller et al., 2016). The results show that the proposed method is not only greatly improved compared with the benchmark algorithm MDNet, but also achieves better results when the video sequence contains object deformation, illumination change, and background interference with similar objects.

The rest of the article is organized as follows. The second section reviews the related work of visual tracking and attention mechanism. Section 3 describes the Global Context module and the algorithm proposed in this article. In the fourth section, the implementation details of the proposed algorithm and the experimental results on three benchmark datasets are given. Finally, the fifth section is the conclusion.

RELATED WORK

Visual Tracking

Object tracking has been an active research field in computer vision for many years. Because convolutional neural networks can extract the feature representation of the object well, their use dramatically increases the success rate and accuracy of the object-tracking algorithm (Li et al., 2018). The current typical object-tracking algorithm consists of a region-selection model, a feature-extraction model, and an observation model. The region-selection model mainly selects the search region of the tracked object in the current frame according to the object of the previous frame. The feature-

extraction model is mainly responsible for extracting features of the selected search area. Finally, the observation model is responsible for locating the tracked object in the current frame based on the confidence judgment of the extracted features. The features extracted from the feature-extraction model directly affect the accuracy of the object-tracking algorithm. Therefore, the work of this article mainly extracts a more robust representation of the object features to distinguish the tracked objects against the chaotic background.

The Attention Mechanism

The attention mechanism can mimic the internal process of the biological observation of objects. It combines global information to assign different weights to different locations to improve the accuracy of observing specific areas. However, using stacked convolutional layers to capture remote information requires enormous computational costs, and the layers are difficult to optimize. Moreover, the information between remote pixels is difficult to transmit under this structure. The current attention mechanisms that can capture remote information are mainly divided into self-attention and channel-attention mechanisms. Self-attention mechanisms have recently been successfully applied to various tasks, such as machine translation and visual recognition. X. Wang et al. (2018) proposed a non-local neural network, which generates an attention map for each query position of the feature graph similar to self-attention. Although the structure of this module is brief, its computational complexity is large, so it is inappropriate to apply it directly to visual tracking. The representative work of the channel-attention mechanism includes the SENet and the Point-Wise Spatial Attention Network (Zhao et al., 2018). These methods rescale different channels to capture the information between the channels to learn the weight information that corresponds to different channels.

The Tracking Algorithm Based on the Attention Mechanism

To solve the interference of similar targets, background confusion, and other factors, and to enhance the feature representation of the tracked target, an intuitive method is to increase the depth of the convolutional neural network to capture information in a larger range. However, in visual tracking, the algorithm speed is required to be very high, so it is not appropriate to simply increase the network depth to enhance the object representation. Since the attention mechanism (Hu et al., 2018; X. Wang et al., 2018) can directly learn the weight information of different positions and channels in the feature map, it has been widely used in image classification, segmentation, and other computer vision tasks (Wang et al., 2017; W. Cao et al., 2019) to obtain more robust object features. For visual tracking, Wang and Yeung (2013) obtained a confidence graph from RNN to distinguish the object from the chaotic background. Choi et al. (2019) proposed an attention mechanism that adaptively selects a subset of the associative filter template. In terms of trackers based on the Siamese network, Q. Wang et al. (2018) used a residual attention network to capture more robust object features. Shen et al. (2019) develop a hierarchical attention Siamese network for visual tracking, while Zhang et al. (2020) utilize a series of structures to emphasize important semantic information. Tan et al. (2021) develops a target-aware non-local block to leverage the long-range dependency, and a location-aware non-local block to associate multiple response maps. In terms of multi-attention in visual tracking, Chen et al. (2019) use an attention mechanism with long short-term memory units to capture multi-level visual attention in a historical context.

In contrast to the above methods, the approach proposed here aims to find a simple and efficient way to introduce the global attention mechanism into the shallow feature-extraction network. In this way, a more robust feature representation of the tracking target can be obtained. Specifically, this article integrates the Global Context module into two positions in MDNet. On one hand, this article embeds the complete Global Context module into the second convolutional layer of MDNet to simultaneously enhance the feature representation of both space and channel. On the other hand, by splitting the Global Context module, the global attention modeling part is added to the first convolutional layer

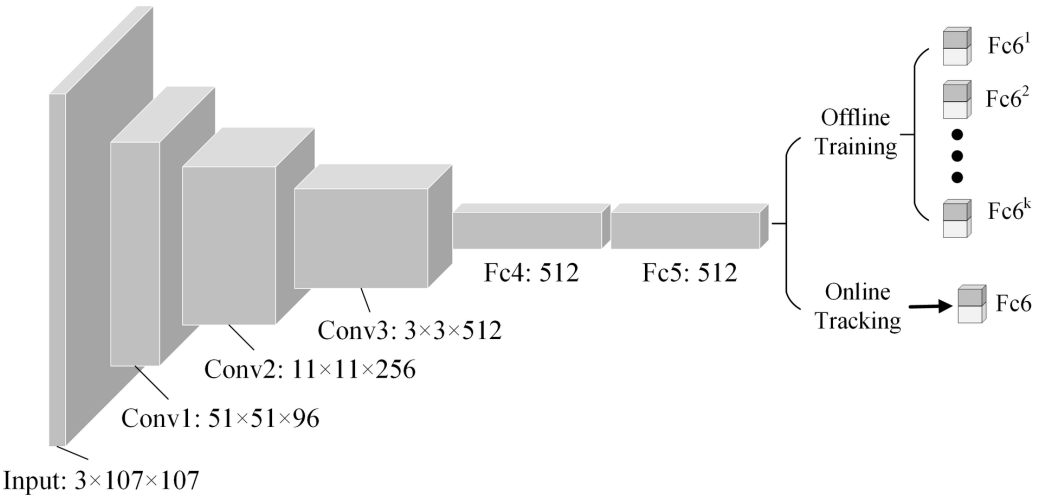
of MDNet to extract remote information. The channel-attention modeling part is added to the third convolutional layer to select features with discriminability and robustness.

METHOD

Baseline Approach: MDNet

MDNet treats each video frame sequence as a single domain and proposes a multi-domain convolutional network based on Visual Geometry Group Medium (VGG-M; Chatfield et al., 2014) architecture. As shown in Figure 1 adapted from the paper by Nam&Han (2016), Conv1-Conv3 and Fc4-Fc5 are domain-independent shared layers, and the initial parameters are obtained through offline training. Fc6 is a specific domain layer, and parameters are randomly initialized in each iteration training or tracking a specific domain. The values 1, 2, . . ., k represents k video sequences of the iterative training model during offline training. MDNet models domain-independent information in the shared layer to obtain a common feature representation, and trains a binary classifier in the Fc6 layer to distinguish the object from the background. Through the above two methods, MDNet finally achieves the purpose of tracking the moving object.

Figure 1. Architecture of the MDNet



MDNet cuts the samples taken from the target search domain into a uniform size of 107×107, and inputs them into the network in Figure 1. The output is the confidence of the target through the Fc6 layer of the network, and classifies negative samples that are difficult to distinguish by the model as complex samples. To reduce the interference of the imbalance of positive and negative samples on the online training of the model, MDNet adopts a hard minibatch mining strategy. This strategy takes the first 96 negative samples with the highest p-values and combines this group of complex samples with 32 positive samples obtained using Gaussian sampling around the target. Finally, the model parameters are iteratively trained and updated through the loss function, represented by Equation 1, and the stochastic gradient descent algorithm:

$$L\left(p,y\right)=-\left(y\cdot ln\left(p\right)+\left(1-y\right)\cdot ln\left(1-p\right)\right) \tag{1}$$

where $y \in \left(0,1\right)$ is the sample category label, and $p \in \{x \mid 0 \leq x \leq 1\}$ is the model-estimated probability of the sample belonging to the target.

Global Context Attention Module

Long-Distance Information Capture Method

The non-local neural network (X. Wang et al., 2018) uses a self-attention mechanism to model the pixel-pair relationship. However, it learns an attention map that is not restricted to each position, causing a waste of computing resources. The non-local neural network aims to gather information from other locations to enhance the characteristics of the current location; x and z are defined as the input and output of the network structure, and the non-local neural network can be expressed as Equation 2:

$$z_i = x_i + W_z \sum_{j=1}^{Np} \frac{f\left(x_i, x_j\right)}{C\left(x\right)} \left(W_v, x_j\right) \tag{2}$$

where C(x) is the normalization factor, Wz and Wv represent the weights of different 1×1 convolutional layers, i is the index of the position, and j is all possible positions enumerated.

The non-local neural network aggregates each query position in a Global Context, providing a pioneering method for capturing non-local features. This method aims to extract a global understanding of the visual scene and is widely used in recognition, object detection, and segmentation.

SENet mainly includes three processes. The first is global average pooling in the network for context modeling to enhance location features; the second is channel weight calculation, including 1×1 convolution, Rectified Linear Unit (ReLU), and Sigmoid, during which feature conversion is used at the same time to obtain the dependence between channels; and the third is the weighted channel features.

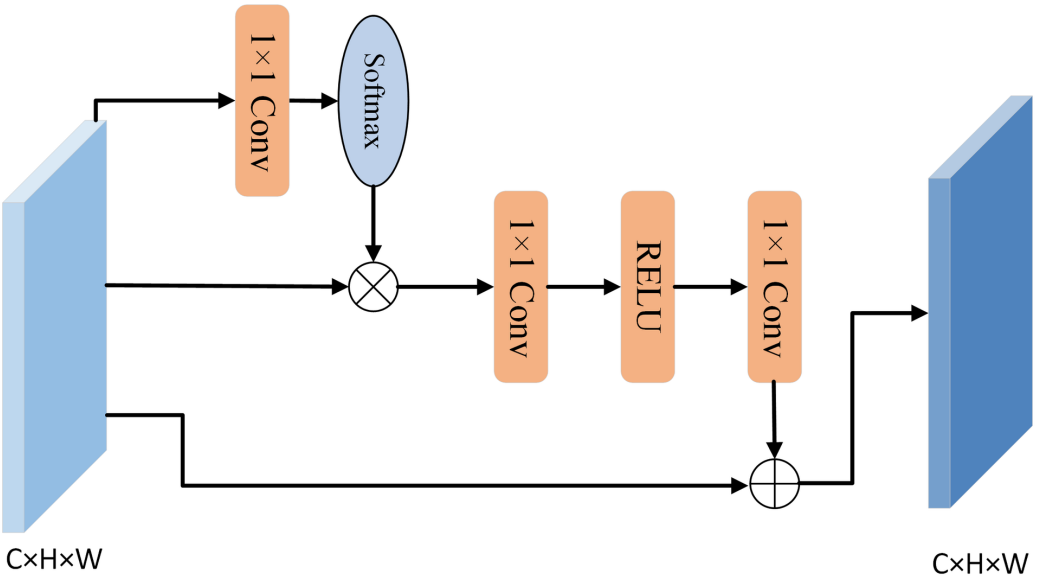
Global Context Network Structure

The Global Context module combines SENet and a simplified non-local neural network. This module has a relatively small amount of calculation, can integrate global information well, and has achieved good results in object detection. Attention modeling in the Global Context module can be divided into two processes. The first is the global attention concentration mechanism for context modeling. This part uses 1×1 convolution and the Softmax function to obtain self-attention weights and then performs attention focused on obtaining global background features. Second, the feature conversion part is used to obtain channel dependence. The detailed structure of the Global Context module is shown in Figure 2, which can be expressed as Equation 3:

$$z_i = x_i + W_{v2} RELU\left(LN\left(W_{v1} \sum_{j=1}^{N_p} \frac{e^{W_k x_j}}{\sum_{m=1}^{N_p} e^{W_k x_m}} x_j\right)\right) \tag{3}$$

where $\frac{e^{W_k x_j}}{\sum_{m=1}^{N_p} e^{W_k x_m}}$ represents the weight of the global attention pool, $W_{v2} RELU\left(LN\left(W_{v1}\left(\sim\right)\right)\right)$ represents channel-attention modeling; x is the input feature map; z is the output feature map; i is the index of the query position; j is the index of enumerating other possible positions; W_{v2}, W_{v1}, and W_k represent the weights of different 1×1 convolutional layers; RELU represents the ReLU activation function; and LN represents the layer normalization.

Figure 2. Architecture of the Global Context module



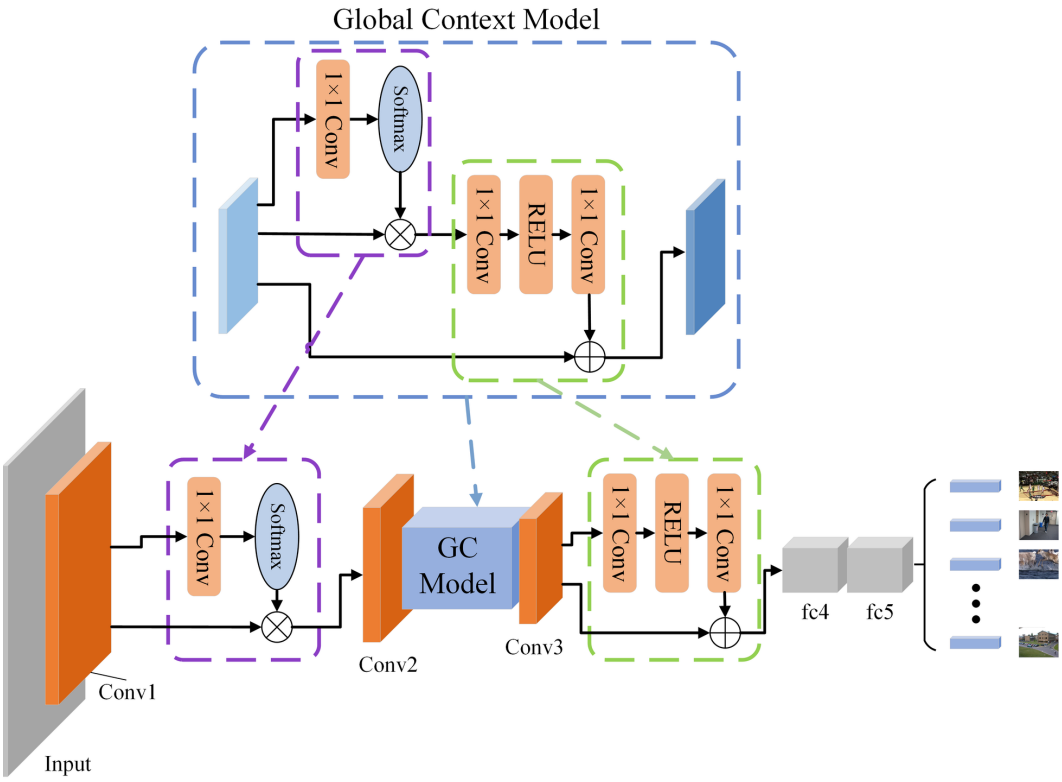
The Global Context module is a lightweight model that can obtain remote non-local features and be flexibly inserted into the network architecture of various visual problems. This article puts the Global Context module into the shared feature-extraction layer of MDNet to improve the generalization performance of network training.

Proposed Approach

The overall framework of the algorithm in this article is shown in Figure 3. This approach improves the network structure of MDNet by introducing a Global Context module embedded in different parts. As can be seen from the figure, the algorithm is divided into a shared feature-extraction layer, an attention modeling layer, and a domain-specific layer. The shared feature-extraction part includes three pre-trained VGG-M model layers. The first two are composed of the convolutional layer, the RELU activation layer, and the maximum pooling layer, while the third only includes the convolutional layer and the RELU activation layer.

In the attention modeling part, this article embeds the Global Context module into the shared feature-extraction part in the following two ways so that it can capture the object's more discriminative and robust features. First, since the input image has a dimension of 11×11×256 after being processed in the second layer of the shared feature-extraction part, the size of the feature map and the number of channels are relatively balanced at this time. Likewise, the Global Context module can capture richer information in both space and channel dimensions. Therefore, this article embeds the complete Global Context module into the second layer of the shared feature-extraction part to learn different weight information in the spatial and channel dimensions. Second, since the dimension of the input picture in the first layer of the shared feature-extraction part is 51×51×96, the dimension after the third layer processing is 3×3×512. Therefore, this article embeds the global attention modeling part of the Global Context module behind the first layer to distinguish the weight information of each location so that the region of the tracked target is clearer. The global attention modeling part is shown in Equation 4:

Figure 3. Architecture of the proposed network



$$z_i = x_i + W_v\left(\sum_{j=1}^{N_p}\frac{e^{W_k x_j}}{\sum_{m=1}^{N_p} e^{W_k x_m}}\right)x_j \tag{4}$$

where x is the input feature map, z is the feature map incorporating spatial attention, i is the index of the query position, j is the index of enumerating other possible positions, and W_v and W_k represent the weights of different 1×1 convolutional layers, where $\frac{e^{W_k x_j}}{\sum_{m=1}^{N_p} e^{W_k x_m}}$ represents the Softmax function.

Since the feature map obtained after the third layer is only 3×3 in size, its information in the spatial dimension is already concentrated enough. However, it still has a wealth of knowledge in the channel dimension that can be extracted. This article embeds the channel-attention modeling part in the Global Context module after the third convolutional layer to enhance the importance of the channel matching the current environment. When the tracked object is disturbed, the channel-attention module can make the characteristics of the object more robust. The attention modeling part of this channel is shown in Equation 5:

$$z_i = x_i + W_{v1}\left(RELU\left(LN\left(W_{v2}\left(x_i\right)\right)\right)\right) \tag{5}$$

where x is the input feature map, z is the feature map incorporating channel attention, W_{v1} and W_{v2} represent the weights of different 1×1 convolutional layers, RELU represents the ReLU activation function, and LN represents the layer normalization.

EXPERIMENTS

Implementation Details

The algorithm was implemented using PyTorch, and the running environment was Ubuntu 18.04, the central processing unit was an Intel Core i9, and the graphics processing unit was Nvidia 2080Ti. The ImageNet-VID dataset was chosen to train the offline model. In this study, the authors selected 700 video sequences from which the probability of these sequences containing the target was approximately 0.5. Twelve frames from each of these videos were selected, generating 100 positive patches and 400 negative patches in each frame. If the Intersection over Union (IOU) of the patch and ground truth was greater than 0.7, it was marked as a positive patch; if it was less than 0.5, it was marked as a negative patch.

This study uses the same online update and training method as MDNet. We set the time interval between the long-term update model and the short-term update model to 100 and 20, respectively. When updating online, the learning rates of the shared feature-extraction layer, the attention layer-modeling layer, and the domain-specific layer were set to 0.0005, 0.006, and 0.002, respectively. At the same time, the model generated 256 candidate regions around the tracking target in the previous frame according to the Gaussian distribution. Similar to offline training, we selected 400 positive samples and 100 negative samples according to the IOU values of the image patch and ground truth. For the bounding box regression model, we used 2,000 samples for training.

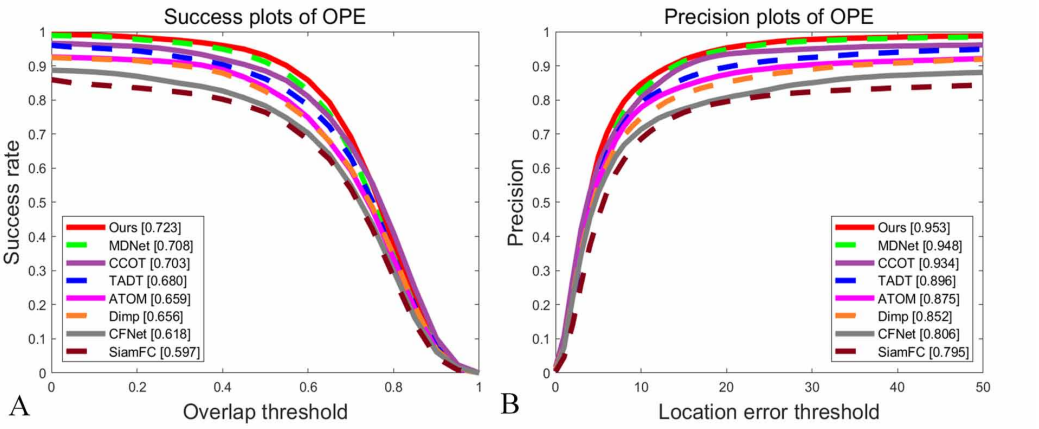
Experimental Dataset and Evaluation Criteria

The algorithm was tested on the OTB2013, OTB2015, and UAV20L datasets. The OTB2013 dataset is a short-term tracking dataset. It has 51 video sequences, and 11 attributes are marked, including BC (background clutters), DEF (target deformation), IV (illumination variation), IPR (in-Plane rotation), LR (low resolution), MB (motion blur), and SV (scale change). The OTB2015 dataset supplements 49 video sequences based on the OTB2013 dataset, with 100 video sequences, and annotates the same 11 attributes as the OTB2013 dataset. The UAV20L dataset contains long-term tracking sequences of 20 Unmanned Aerial Vehicle (UAV) scenes, and the average sequence length reaches 2934 frames. Compared with the OTB2013 and OTB2015 datasets, the UAV20L dataset has a longer sequence time span, and the target is prone to drastic changes during the tracking process, such as the target deforming or leaving the field of view. The above three datasets can test the accuracy and robustness of this algorithm in various situations in short-term tracking and long-term tracking.

The OTB2013, OTB2015, and UAV20L datasets all use two evaluation indicators: the precision plot and the success plot. The success plot represents the percentage of the total number of frames whose overlap ratio between the target frame predicted by the algorithm and the manually labeled target frame is greater than 0.5. The accuracy plot represents the percentage of the total number of frames whose distance between the center position of the predicted target and the center position of the manually marked target is within 20 pixels. At the same time, the test method used in this experiment is a one-time run evaluation strategy (OPE), which means that all test videos are run only once.

This article compared the algorithm proposed in this article with the tracking algorithm target-aware deep tracking (TADT; Li et al., 2019), accurate tracking by overlap maximization (ATOM; Danelljan et al., 2019), the discriminative model prediction tracker (DIMP; Bhat et al., 2019), the correlation filter network (CFNet; Valmadre et al., 2017), SiamFC (Bertinetto et al., 2016), and the continuous convolution operator tracker (C-COT; Danelljan et al., 2016) on the OTB2013 and

Figure 4. Performance comparison using OTB2013. The left panel (A) is the success plot and the right panel (B) is the precision plot. The lines in the figure and the legend correspond from top to bottom.



OTB2015 datasets. These six algorithms use the depth features extracted by the convolutional neural network. Therefore, this paper compares the proposed algorithm with other tracking algorithms, such as ATOM, DIMP, and C-COT, which can effectively resist these interference factors. These comparisons can efficiently evaluate the effectiveness of the global attention mechanism introduced in this paper. At the same time, we compare the proposed algorithm with the tracking algorithms, which are easily disturbed by similar targets, such as CFNet and SiamFC. These comparisons can effectively evaluate the importance of the global attention mechanism introduced in this paper in the tracking framework.

Using the UAV20L dataset, the authors compare the algorithm proposed here with the tracking algorithm spatially regularized correlation filters tracker (SRDCF; Danelljan et al., 2015), the multi-store tracker (MUSTER; Hong et al., 2015), the scale adaptive with multiple features tracker (SAMF; Li & Zhu, 2014), the multiple experts using entropy minimization tracker (MEEM; J. Zhang et al., 2014), the discriminative scale space tracker (DSST; Danelljan et al., 2014), the structed output tracker (Struck; Hare et al., 2015), and SiamFC (Bertinetto et al., 2016). Since the length of the video sequence in the UAV20L dataset is longer than in the OTB2015 dataset, we chose for comparison the correlation filter-based algorithm SRDCF, MUSTER, and other algorithms that are more suitable for long-sequence tracking.

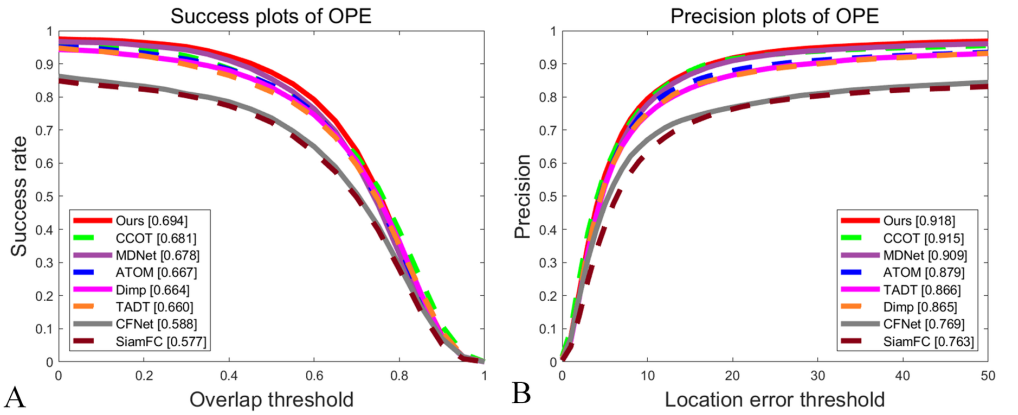
Results for the Full Dataset

Figures 4–6 show the test results of this algorithm using the OTB2013, OTB2015, and UAV20L datasets. As can be seen in the figure, compared with the benchmark algorithm MDNet, the success rate and accuracy of the algorithm in this article have been improved to a certain extent. At the same time, it has also achieved competitive results with the current mainstream trackers.

With the OTB2013 data set, the algorithm's success rate reached 72.3%, and the accuracy rate reached 95.3%, which are 2.1% and 0.5% higher, respectively, than the baseline algorithm MDNet. Using the OTB2015 dataset, the algorithm's success rate reached 69.4%, and the accuracy rate reached 91.8%, which are 2.4% and 1% higher, respectively, than the baseline algorithm MDNet. For the UAV20L dataset, the algorithm's success rate reached 51.9%, and the accuracy rate reached 68.3%, which are 8.4% and 8.2% higher, respectively, than the baseline algorithm MDNet.

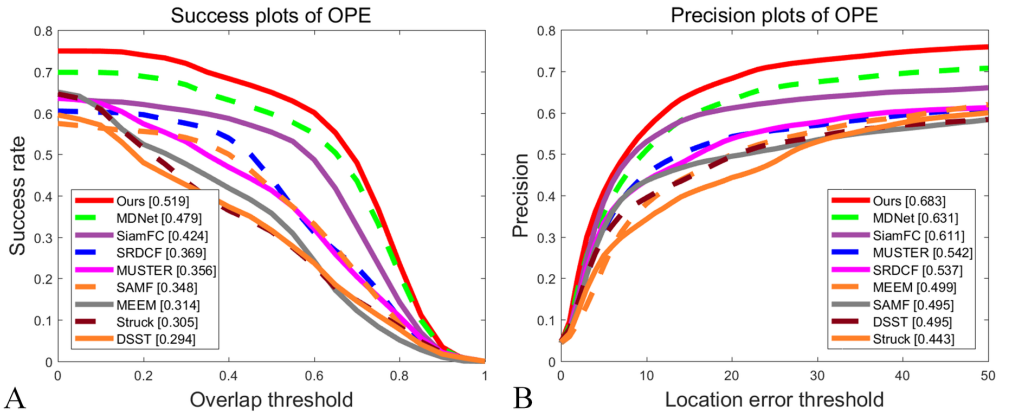
As can be seen from Figures 4 and 5, we compare the proposed algorithm with ATOM, which introduces a target classification module to increase the model's ability to distinguish similar targets, and TADT, which uses a target-aware approach similar to MDNet. The algorithm in this paper achieves

Figure 5. Performance comparison using OTB2015. The left panel (A) is the success plot and the right panel (B) is the precision plot. The lines in the figure and the legend correspond from top to bottom.



better performance because it extracts more robust features of the target through the efficient global attention module. Because the length of the video sequence in the UAV20L dataset is longer than that in the OTB2015 dataset, the tracking target is more likely to be disturbed by factors such as similar targets and background confusion. Figure 6 shows that the filter templates of the algorithms based on correlation filtering, such as DSST, SAMF, and MEEM, are vulnerable to contamination. However, the proposed algorithm has a great advantage over these alternatives because it can focus on the target's adaptively more robust features. Compared with tracking algorithms that can effectively resist the interference of similar targets and target deformation, such as the SRDCF algorithm that introduces filter regularization, and the MUSTER algorithm that introduces long- and short-term memory, the algorithm offered in this paper extracts more discriminative features of targets through the global attention module. Therefore, the algorithm introduced here has certain advantages in terms of success rate and accuracy compared with the above algorithms when tracking long sequences.

Figure 6. Performance comparison using UAV20L. The left panel (A) is the success plot and the right panel (B) is the precision plot. The lines in the figure and the legend correspond from top to bottom.

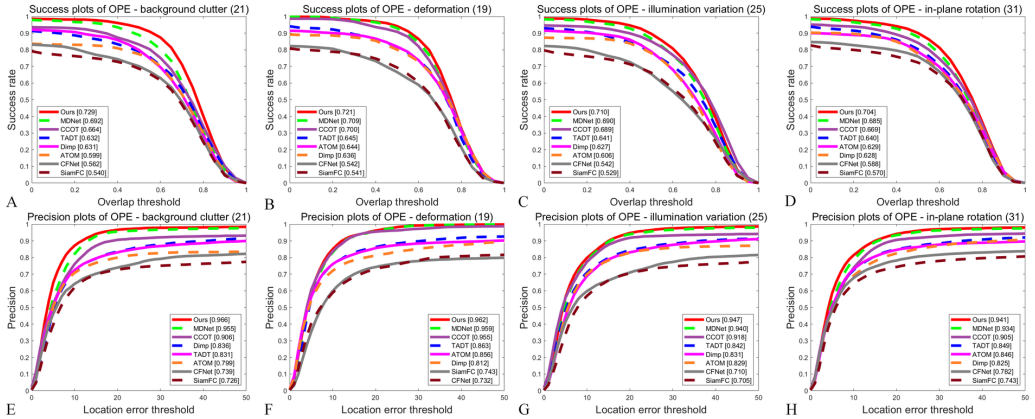


Results for the Dataset with Different Attributes

Since our algorithm introduces the Global Context attention module, the authors selected background clutter, deformation, illumination change, and in-plane rotation in the OTB2013 and OTB2015 datasets and tested it under the condition that four interference factors exist separately. The results are shown in Figures 7 and 8.

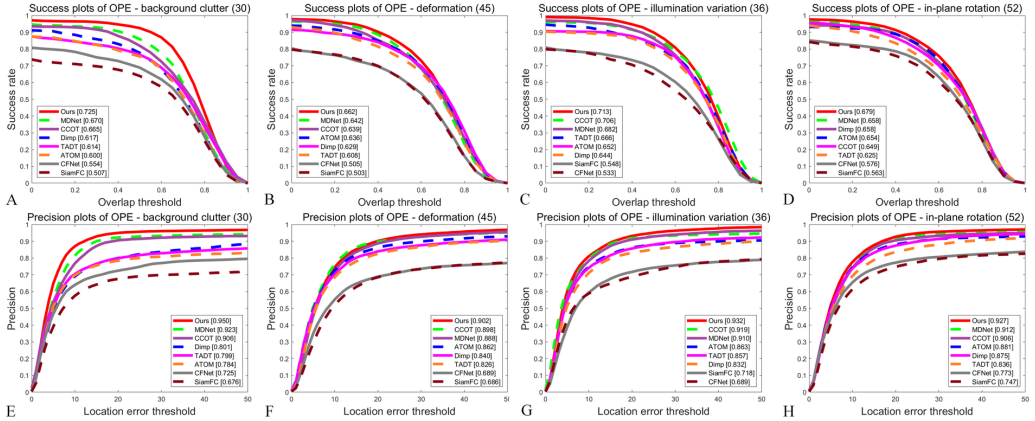
The test results for the four attributes in the OTB2013 dataset are shown in Figure 7. Compared with the benchmark algorithm MDNet, the algorithm in this article improves the success rate of the four attributes by 5.3%, 1.7%, 2.9%, and 2.8%, respectively. At the same time, the accuracy of the algorithm in the four attributes is improved by 1.2%, 0.3%, 0.7%, and 2.8%, respectively.

Figure 7. Performance comparison using OTB2013 with DEF, BC, IV, and IPR. Subgraphs (A), (B), (C), and (D) are the success plots. Subgraphs (E), (F), (G), and (H) are the precision plots. The lines in the figure and the legend correspond from top to bottom.



The test results for the four attributes in the OTB2015 dataset are shown in Figure 8. Compared with the benchmark algorithm MDNet, the algorithm in this article improves the success rate of the

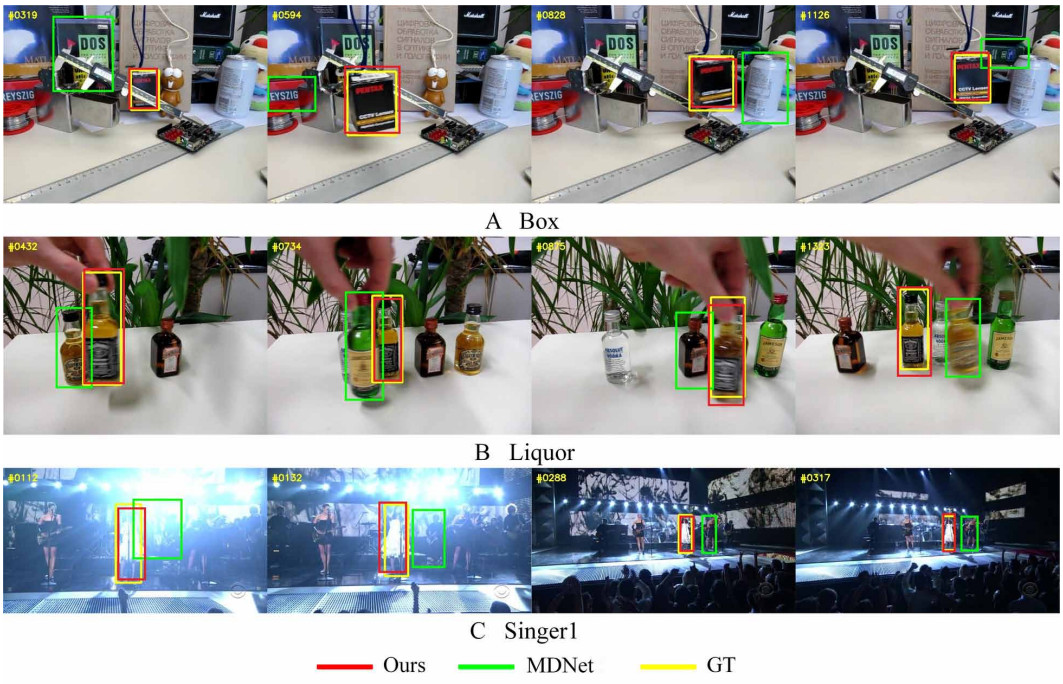
Figure 8. Performance comparison using OTB2015 with DEF, BC, IV, and IPR. Subgraphs (A), (B), (C), and (D) are the success plots. Subgraphs (E), (F), (G), and (H) are the precision plots. The lines in the figure and the legend correspond from top to bottom.



four attributes by 8.2%, 3.1%, 4.5%, and 3.2%, respectively. At the same time, the accuracy of the algorithm in the four attributes is improved by 2.9%, 1.6%, 2.4% and 1.6%, respectively.

It can be seen from the Figures 7 and 8 that, when the four interference factors exist alone, the novel algorithm not only indicates considerable improvement over the baseline algorithm MDNet, but also achieves the best results among all the compared algorithms. It benefits from the global attention module learning the more robust features of the target, and can pay attention to the target adaptively. Therefore, the global attention module can guide the attention of the tracking framework to the area where the target is most likely to appear, reducing the influence of similar targets and background confusion. These improvements fully illustrate the effectiveness of the Global Context attention module introduced in this article.

Figure 9. Tracking results of the tracker proposed in this article compared to MDNet on challenging sequences (Box, Liquor, Singer1) in the OTB2015 dataset.



Qualitative Evaluation of the Algorithm

To qualitatively evaluate the tracking effect of the proposed tracker, we selected three video sequences to further analyze the performance of the algorithm proposed in this article. These video sequences contain attributes such as similar target interference and target occlusion. The tracking video screenshots are shown in Figure 9. The video names, from top to bottom, are Box, Liquor, and Singer1. Since MDNet's shared feature-extraction part has a few convolutional network layers and cannot capture long-distance information, it is easy to cause target loss when objects are similar to the tracked target in the background. Suppose the tracked target is partially or completely occluded. In that case, the MDNet algorithm will not be able to accurately extract the distinguishable features of the target, which will result in loss of the target.

As shown in Figure 9A, in the Box video sequence, because there is a large number of square targets similar to the tracked target in the background, the MDNet algorithm loses the target at frame 319 and cannot recapture the target in subsequent frames. In Figure 9B, because there are many similar glass bottles, the MDNet algorithm mistakenly identifies similar targets in the background as being tracked after the target has been occluded several times. As shown in Figure 9C, in the Singer1 video sequence, there are people similar to the tracked target in the background and dramatic changes in the illumination around the tracked target. The MDNet algorithm cannot extract robust features because the target is affected by changes in illumination, so the target is lost when it is interfered with by similar targets. Since a Global Context module is introduced into MDNet, the improved MDNet algorithm can learn more robust features of the tracked target and capture remote information. The method proposed in this study can resist the interference caused by partial or complete occlusion of the target and accurately locate the tracked target when there are similar targets in the background. It can be seen from Figure 9 that the method in this article has achieved excellent results.

CONCLUSION AND FUTURE WORK

This article has made two improvements to MDNet. The first is to use the complete Global Context module to simultaneously capture the space and channel dimension information contained in the feature map with a relatively balanced size and number of channels in MDNet; the second is to split the Global Context attention module and use the global attention modeling part of MDNet to capture spatial attention on the larger feature maps in MDNet. The channel-attention modeling part captures channel attention on the feature maps with more channels in MDNet. The full experiments on the three datasets of OTB2013, OTB2015, and UAV20L show that the algorithm introduced here displays greater improvement in tracking success rate and accuracy compared to the benchmark algorithm MDNet, and compared with other mainstream tracking algorithms. The novel algorithm also achieved the highest performance.

Although the proposed tracker can achieve favorable tracking results in the three benchmark datasets, our method can still be further improved. First, although the Global Context module we introduced in MDNet reduces the extra computation amount as much as possible when capturing space and channel attention, our algorithm still cannot meet the requirement of real-time performance, and its running speed needs to be further improved. In fact, one of the factors limiting the speed of the algorithm in this paper is the complex update strategy; as a result, we plan to incorporate the attention mechanism into the update strategy in future work to obtain a better and faster update strategy. Second, since the algorithm in this paper adopts a training strategy similar to MDNet, it only focuses on the classification between tracking targets and backgrounds while ignoring the differences between cross-domain targets. This problem can be ameliorated by introducing cross-domain target loss into future algorithm training. Third, the three convolutional layers in the shared feature-extraction layer are enhanced respectively by the attention mechanism in this paper. However, due to the different information richness contained in the features extracted by each convolutional layer, such an enhancement method may cause certain information loss. In future work, the author plans to integrate the features learned by the attention mechanism in different convolutional layers to extract more robust and differentiated features of the tracking target. In addition, the main purpose of the shared feature-extraction layer is to model the domain-independent information to obtain the common feature representation of the tracking target. Therefore, the authors plan to use more datasets to fully train the algorithm's shared feature-extraction layer in future work to improve the shared feature extraction layer and obtain better performance.

REFERENCES

Abbass, M. Y., Kwon, K. C., Kim, N., Abdelwahab, S. A., El-Samie, F. E. A., & Khalaf, A. A. (2021). A survey on online learning for visual tracking. *The Visual Computer*, *37*(5), 993–1014. doi:10.1007/s00371-020-01848-y

Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., & Torr, P. H. (2016). Fully-convolutional Siamese networks for object tracking. In G. Hua, & H. Jégou (Eds.), *Proceedings of the European Conference on Computer Vision* (pp. 850-865). Springer. doi:10.1007/978-3-319-48881-3_56

Bhat, G., Danelljan, M., Gool, L. V., & Timofte, R. (2019). Learning discriminative model prediction for tracking. In M. L. Kyoung, & F. David (Eds.), *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6181–6190). IEEE. doi:10.1109/ICCV.2019.00628

Cao, W., Li, Y., & He, Z. (2019). Weighted optical flow prediction and attention model for object tracking. *IEEE Access: Practical Innovations, Open Solutions*, *7*(99), 144885–144894. doi:10.1109/ACCESS.2019.2944649

Cao, Y., Xu, J., Lin, S., Wei, F., & Hu, H. (2019). GCNet: Non-local networks meet squeeze-excitation networks and beyond. In M. L. Kyoung, & F. David (Eds.), *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* (pp. 1–10). IEEE. doi:10.1109/ICCVW.2019.00246

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. In M. Valstar, & A. French (Eds.), *Proceedings of the British Machine Vision Conference* (pp. 1–11). BMVA Press. doi:10.5244/C.28.6

Chen, B., Li, P., Sun, C., Wang, D., Yang, G., & Lu, H. (2019). Multi attention module for visual tracking. *Pattern Recognition*, *87*, 80–93. doi:10.1016/j.patcog.2018.10.005

Choi, J., Chang, J. H., Yun, S., Fischer, T., Demiris, Y., & Young Choi, J. (2017). Attentional correlation filter network for adaptive visual tracking. In J. Dan & S. Harry (Eds.), *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4807–4816). IEEE. doi:10.1109/CVPR.2017.513

Danelljan, M., Bhat, G., Khan, F. S., & Felsberg, M. (2019). Atom: Accurate tracking by overlap maximization. In D. Larry, & T. Philip (Eds.), *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4660–4669). IEEE. doi:10.1109/CVPR.2019.00479

Danelljan, M., Hager, G., Khan, F., & Felsberg, M. (2014). Accurate scale estimation for robust visual tracking. In M. Valstar, & A. French (Eds.), *Proceedings of the British Machine Vision Conference* (pp. 1–11). BMVA Press. doi:10.5244/C.28.65

Danelljan, M., Hager, G., Shahbaz Khan, F., & Felsberg, M. (2015). Learning spatially regularized correlation filters for visual tracking. In R. Bajcsy, & G. Hager (Eds.), *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4310–4318). IEEE. doi:10.1109/ICCV.2015.490

Danelljan, M., Robinson, A., Khan, F. S., & Felsberg, M. (2016). Beyond correlation filters: Learning continuous convolution operators for visual tracking. In G. Hua, & H. Jégou (Eds.), *Proceedings of the European Conference on Computer Vision* (pp. 472–488). Springer. doi:10.1007/978-3-319-46454-1_29

Hare, S., Golodetz, S., Saffari, A., Vineet, V., Cheng, M. M., Hicks, S. L., & Torr, P. H. (2015). Struck: Structured output tracking with kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(10), 2096–2109. doi:10.1109/TPAMI.2015.2509974 PMID:26700968

Hong, Z., Chen, Z., Wang, C., Mei, X., Prokhorov, D., & Tao, D. (2015). Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. In H. Bischof, & D. Forsyth (Eds.), *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 749–758). IEEE. doi:10.1109/CVPR.2015.7298675

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In M. Brown, & B. Morse (Eds.), *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7132–7141). IEEE. doi:10.1109/TPAMI.2019.2913372

Li, P., Wang, D., Wang, L., & Lu, H. (2018). Deep visual tracking: Review and experimental comparison. *Pattern Recognition*, *76*, 323–338. doi:10.1016/j.patcog.2017.11.007

Li, X., Ma, C., Wu, B., He, Z., & Yang, M. H. (2019). Target-aware deep tracking. In D. Larry, & T. Philip (Eds.), *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1369–1378). IEEE. doi:10.1109/CVPR.2019.00146

Li, Y., & Zhu, J. (2014). A scale adaptive kernel correlation filter tracker with feature integration. In D. Fleet, & T. Pajdla (Eds.), *Proceedings of the European Conference on Computer Vision* (pp. 254–265). Springer. doi:10.1007/978-3-319-16181-5_18

Marvasti-Zadeh, S. M., Cheng, L., Ghanei-Yakhdan, H., & Kasaei, S. (2021). Deep learning for visual tracking: A comprehensive survey. *IEEE Transactions on Intelligent Transportation Systems*, *1*(1), 1–26. doi:10.1109/TITS.2020.3046478

Mueller, M., Smith, N., & Ghanem, B. (2016). A benchmark and simulator for UAV tracking. In G. Hua, & H. Jégou (Eds.), *Proceedings of the European Conference on Computer Vision* (pp. 445–461). Springer. doi:10.1007/978-3-319-46448-0_27

Nam, H., & Han, B. (2016). Learning multi-domain convolutional neural networks for visual tracking. In R. Bajcsy, & F. F. Li (Eds.), *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4293–4302). IEEE. doi:10.1109/CVPR.2016.465

Shen, J., Tang, X., Dong, X., & Shao, L. (2019). Visual object tracking by hierarchical attention Siamese network. *IEEE Transactions on Cybernetics*, *50*(7), 3068–3080. doi:10.1109/TCYB.2019.2936503 PMID:31536029

Tan, H., Zhang, X., Zhang, Z., Lan, L., Zhang, W., & Luo, Z. (2021). Nocal-Siam: Refining visual features and response with advanced non-local blocks for real-time Siamese tracking. *IEEE Transactions on Image Processing*, *30*(7), 2656–2668. doi:10.1109/TIP.2021.3049970 PMID:33439844

Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A., & Torr, P. H. (2017). End-to-end representation learning for correlation filter based tracking. In R. Chellappa, & Z. Zhang (Eds.), *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2805–2813). doi:10.1109/CVPR.2017.531

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., & Tang, X. et al. (2017). Residual attention network for image classification. In R. Chellappa, & Z. Zhang (Eds.), *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3156–3164). IEEE. doi:10.1109/CVPR.2017.683

Wang, N., & Yeung, D. Y. (2013). Learning a deep compact image representation for visual tracking. In C.J.C. Burges, & L. Bottou (Eds.), *Proceedings of the 26th International Conference on Neural Information Processing Systems* (pp. 1–9). MIT Press. doi: doi:10.5555/2999611.2999702

Wang, Q., Teng, Z., Xing, J., Gao, J., Hu, W., & Maybank, S. (2018). Learning attentions: Residual attentional Siamese network for high performance online visual tracking. In M. Brown, & B. Morse (Eds.), *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4854–4863). IEEE. doi:10.1109/CVPR.2018.00510

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In M. Brown, & B. Morse (Eds.), *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7794–7803). IEEE. doi:10.1109/CVPR.2018.00813

Wu, Y., Lim, J., & Yang, M. H. (2013). Online object tracking: A benchmark. In G. Medioni, & R. Zabih (Eds.), *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2411–2418). IEEE. doi:10.1109/CVPR.2013.312

Wu, Y., Lim, J., & Yang, M. H. (2015). Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(9), 1834–1848. doi:10.1109/TPAMI.2014.2388226 PMID:26353130

Zhang, D., Zheng, Z., He, X., Su, L., & Chen, L. (2020). Learning fine-grained similarity matching networks for visual tracking. In C. Gurrin, & B. Jónsson (Eds.), *Proceedings of the 2020 International Conference on Multimedia Retrieval* (pp. 296–300). doi:10.1145/3372278.3390729

Zhang, J., Ma, S., & Sclaroff, S. (2014). MEEM: Robust tracking via multiple experts using entropy minimization. In D. Fleet, & T. Pajdla (Eds.), *Proceedings of the European Conference on Computer Vision* (pp. 188–203). Springer. doi:10.1007/978-3-319-10599-4_13

Zhang, K., Zhang, L., & Yang, M. H. (2014). Fast compressive tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(10), 2002–2015. doi:10.1109/TPAMI.2014.2315808 PMID:26352631

Zhang, Z., & Peng, H. (2019). Deeper and wider Siamese networks for real-time visual tracking. In D. Larry, & T. Philip (Eds.), *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4591–4600). IEEE. doi:10.1109/CVPR.2019.00472

Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C. C., Lin, D., & Jia, J. (2018). Psanet: Point-wise spatial attention network for scene parsing. In V. Ferrari, & M. Hebert (Eds.), *Proceedings of the European Conference on Computer Vision* (pp. 267–283). Springer. doi:10.1007/978-3-030-01240-3_17

Yucheng Wang's research fields include computer vision, object detection and object tracking. Yucheng Wang is currently studying for a bachelor's degree in computer science and technology in the Department of Computer Science and Technology, School of Computer Science and Technology, Wuhan University, Hubei, China. He plans to pursue a master's degree to continue his research in the field of object tracking, object detection and computer vision. Yucheng Wang's next research direction plans to continue to focus on the attention mechanism, Anchor-free tracking algorithm and Transformer tracking algorithm. He has published some original articles in some international conferences and international journals.

Xi Chen's research field include artificial intelligence, image processing, computer vision and target tracking. Professor Xi Chen obtained his bachelor's, master's and doctorate degrees from the School of Computer Science, Wuhan University, Hubei, China. He is currently a professor and doctoral supervisor. He has more than 8 years of teaching experience for engineering undergraduates and postgraduates. At the same time, he has supervised many doctoral students. He has published many research papers in famous international journals. He presided over 21 national defense and aerospace model projects, published more than 20 SCI/EI papers, and obtained 4 national invention patents and 1 utility model patent.

Zhongjie Mao has obtained his bachelor's and master's degrees from the Department of Computer Science and Technology, School of Computer Science, Wuhan University, Wuhan, China, and is currently studying for a PhD in the Department of Computer Science and Technology, School of Computer Science, Wuhan University. Zhongjie Mao's research field include artificial intelligence, image processing, computer vision and target tracking. He has many years of engineering project experience and academic research experience in the field of computer vision and target tracking, and has a relatively deep understanding of the development of the target tracking field. He has published many articles in high-level international academic conferences and journals.

Jia Yan has obtained his bachelor's, master's and doctorate degrees from the School of Electronic Information, Wuhan University, Hubei, China, and is currently a professor of the School of Electronic Information of Wuhan University. Jia Yan's research fields include Image and video processing, Blind image quality assessment and Visual tracking. Jia Yan has many years of teaching experience in teaching engineering undergraduates and graduate students. He has many years of research experience in the field of computer vision and target tracking, and has a deep understanding of the field. He has published many articles in high-level international academic conferences and journals.