SiamADN: Siamese Attentional Dense Network for UAV Object Tracking

WANG Zhi[1,2], *WANG Ershen*[3*], *HUANG Yufeng*[3], *YANG Siqi*[3], *XU Song*[3]

1. Zhejiang Key Laboratory of General Aviation Operation Technology, General Aviation Research Insititute of Zhejiang Jiande, Jiande 311612, P.R.China;

2. Department of General Aviation, Civil Aviation Management Institute of China, Beijing 100102, P.R.China;

3. School of Electronic and Information Engineering, Shenyang Aerospace University, Shenyang 110136, P.R.China

(Received 15 June 2021; revised 10 July 2021; accepted 15 August 2021)

Abstract: Single object tracking based on deep learning has achieved the advanced performance in many applications of computer vision. However, the existing trackers have certain limitations owing to deformation, occlusion, movement and some other conditions. We propose a siamese attentional dense network called SiamADN in an end-to-end offline manner, especially aiming at unmanned aerial vehicle (UAV) tracking. First, it applies a dense network to reduce vanishing-gradient, which strengthens the features transfer. Second, the channel attention mechanism is involved into the Densenet structure, in order to focus on the possible key regions. The advance corner detection network is introduced to improve the following tracking process. Extensive experiments are carried out on four mainly tracking benchmarks as OTB-2015, UAV123, LaSOT and VOT. The accuracy rate on UAV123 is 78.9%, and the running speed is 32 frame per second (FPS), which demonstrates its efficiency in the practical real application.

Key words: unmanned aerial vehicle (UAV); object tracking; dense network; corner detection; siamese network

CLC number: TN925 **Document code:** A **Article ID:** 1005-1120(2021)04-0587-10

0 Introduction

Visual object tracking (VOT) has received considerable attention due to its wide applications in video surveillance, human-computer interaction, unmanned driving and other fields[1]. Its main function is to estimate the position of certain target in a video sequence, as the target position in the initial frame is given. Although the VOT algorithms have been studied for decades, and great progress has been made in recent years[2], it is still a difficult problem under the unconstrained conditions of scale and position variation, background clutters, and serious occlusions.

Current tracking methods can be roughly divided into two categories: Discriminative correlation filter (DCF) based methods[3-5] and Deep Learning based approaches[1, 6-10]. The traditional DCF methods achieve higher calculation in the Fourier domain for tracking process[3]. The improved DCF-based approaches demonstrate to construct robust models[3-5], imposing learning restrictions or mitigate the filter degradation[3]. Recently, deep learning-based way has been widely applied in the tracking process. CFNet uses correlation filters in the feature extraction layer to improve the tracking accuracy[6]. DSiam studies the use of feature transformation to modify the Siamese branch and improve the accuracy by suppressing the background[11]. RASNet contains a variety of attention mechanisms to make the tracking model adapt to the target[12]. In order to obtain more accurate bounding box, SiamRPN introduces region proposal network (RPN)[13] in SiamFC to avoid the complicated calculation of obtaining multiple scales[14]. Inspired by SiamRPN, DaSiamRPN improves the recognition ability by adding hard negative data in the training process[15].

*Corresponding author, E-mail address: wanges_2016@126.com.

How to cite this article: WANG Zhi, WANG Ershen, HUANG Yufeng, et al. SiamADN: Siamese attentional dense network for UAV object tracking[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2021, 38(4): 587-596.

http://dx.doi.org/10.16356/j.1005-1120.2021.04.005

Advanced Siamese networks, such as SiamRPN++[7], SiamMask[16] and SiamDW[17] use modern deep networks to optimize the architecture. SPM tracker combines coarse matching and fine matching to improve robustness and recognition ability[18]. The anchor designes in these trackers avoid time-consuming multi-scale feature extraction. As the networks are deepen, the problem of gradient disappearance may become more obvious in the deep learning process. Densenet can help to reduce vanishing-gradient, strengthen features transfer, and reduce related parameter numbers to a certain extent.

In this paper, we design an effective Siamese attentional dense keypoints network (SiamADN) based on anchor-free Siamese network to solve the state estimation issue, and achieve good performance. First, the dense network is introduced to advance the feature extraction process, which effectively improves the performance of the follow-up tracking process. Then we use the global attention, and further improve the accuracy of the tracking frame. Inspired by the related anchor-free detector methods, our framework relies on anchor points and uses corner detection to accurately predict the bounding box. The key contributions of our paper are summarized as follows.

To our knowledge, the attentional dense network is proposed for tracking feature exaction. The Densenet is applied to directly receive and transport features through all the preceding layers, which can alleviate vanishing gradients and enhance feature propagation. Then the channel attention mechanism is performed to obtain global information and compress integrated information.

An advanced corner detection network is introduced to progress the tracking part, since it gets an efficient way to find the missing target and increase the tracking speed. To speed up the prediction step, an anchor-free corner strategy is used to extract information for predicting the bounding box.

Comprehensive experiments are performed on four challenging datasets (OTB-2015, LaSOT, VOT and UAV123). The tracking results demonstrate that our designed method outperforms the existing state-of-the-art approaches.

1 Related Work

We mainly review the Siamese related tracking methods because they dominate the tracking performance in recent years. Tracking researchers are committed to design faster and more accurate tracking algorithms, such as feature extraction, template update, classifier design, and bounding box regression. Early feature extraction mainly uses color texture or other manual features. Thanks to the development of deep learning, convolution feature CNN is currently widely used.

Current studies show that the online training and offline tracking methods based on deep neural networks are feasible. Siamese structure is involved to achieve better balance between accuracy and efficiency. As a pioneering job, SiamFC constructs a fully convolutional Siamese network to train the tracker[8]. Encouraged by its success, many researchers follow this work and propose some updated models. CFNet introduces a relevant filter layer in the SiamFC framework and performs online tracking to improve accuracy[6]. DSiam proposes to modify the Siamese branches through two online transformations and learns a dynamic Siamese network to improve accuracy under acceptable speed loss[11]. SAsiam constructs a dual Siamese network with semantic and appearance branch[19]. These two branches are trained separately to maintain the heterogeneity of features, but merged to improve tracking accuracy during the testing. In order to deal with scale change problems, these Siamese networks need to perform multi-scale search, which is time-consuming. Attention mechanism can help to discover the important components, seeking to increase the tracking capability. MemTrack is to use the attention scheme to search the object for matching in the following tracking part[20]. SiamDA uses a dual attention module to select the features, that is effectively integrated into the Siamese network for tracking[21]. Further, a siamese network with multi-attention map is proposed to deal with the visual object tracking[22].

Inspired by the RPN[14], the SiamRPN tracker performs region proposal extraction after the output of the Siamese network[13]. SiamRPN uses the joint training of a classification branch and a regression branch for region suggestion, avoiding the time-consuming step of extracting multi-scale feature maps due to the invariance of target scale, and obtaining very effective results. However, it is difficult to deal with distractors that look similar to objects. These RPN-based trackers use anchor points for regional recommendations. In addition, the anchor box can use depth feature mapping to avoid repeated calculations and greatly accelerate the tracking speed. Although SiamRPN++ uses a very deep neural network, it can still work at a fairly fast real-time speed. The accuracy and speed of most advanced anchor-free trackers (such as ECO[23]) still lag behind these anchor-based trackers on challenging benchmarks such as GOT-10K[24]. However, the tracking performance is very sensitive to the relative hyper-parameters of anchor, requireing careful adjustment and empirical skills to achieve the desired performance. In addition, the size and aspect ratio of the anchor box are fixed. Even with heuristically adjusted parameters, these trackers are still difficult to handle objects with large shape deformation and posture changes. In this paper, our method can greatly alleviate these problems. In addition, we also prove that a tracker with a much simpler structure can achieve better performance than most advanced trackers.

2 Method

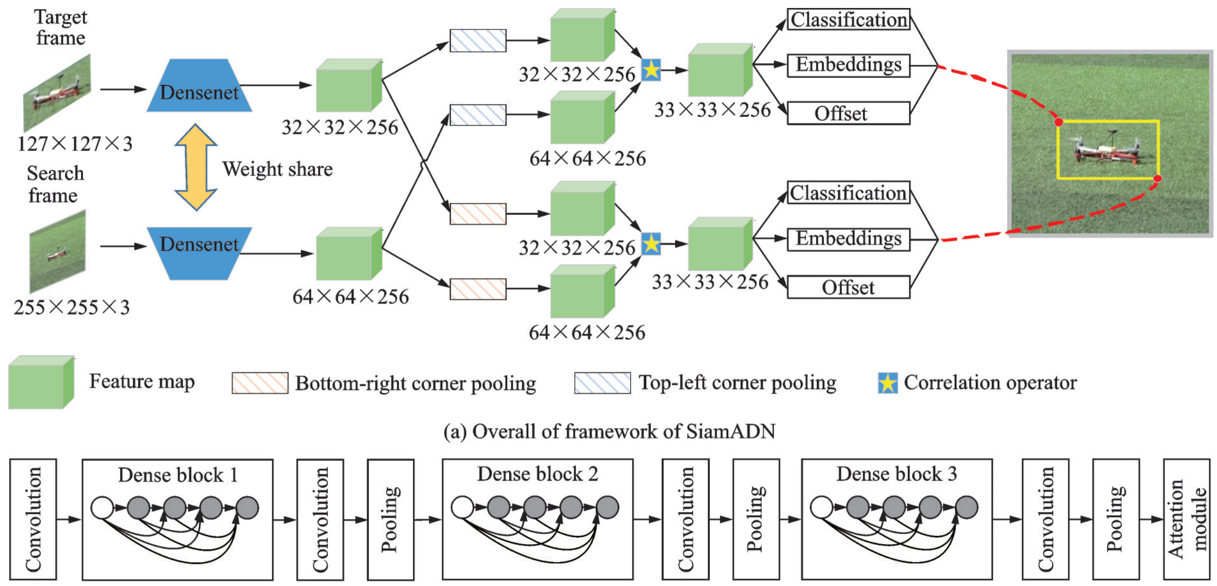
In this section, we first briefly introduce the framework of siamese networks and the structure of object tracking using siamese networks. Then we predict the bounding box by introducing the key point prediction module. The corner detection sub-networks predict the classification, regression and embedding information of the corner point for tracking.

2.1 Siamese network

Siamese network[25] is a neural network framework. Its specific implementation is not limited to a certain network. In simple terms, it can be implemented with convolutional neural networks (CNNs) or recurrent neural networks (RNNs). The Siamese network can be specifically explained as a "conjoined neural network", which has two sub-neural networks like twins. The two sub-networks have the same structure, and share same weight settings.

Specifically, the siamese network generally consists of two equal branches: The template branch and the search branch. The two branches share weights to ensure that each has similar transformations.

The overall network is illustrated in Fig.1(a). The left side of Fig.1(a) is Siamese subnetwork for feature extraction. We adopt Densenet as backbone



(a) Overall of framework of SiamADN

(b) Detailed structure of Densenet

Fig.1 Overall framework and detail diagram of SiamADN

network in Fig.1(b). In the middle green box is the corner detection module, which has two branches one for top-left and another for bottom-right. In each branch, we predict the classification, embedding and offset of the key points.

2. 2 Dense network based siamese tracking

Based on the Siamese network[25], the tracking algorithm is applied to formulate target tracking as a cross-correlation problem and learns tracking similarity from a deep model[8]. The Siamese network is divided into two branches: One branch is used to learn the feature representation of tracking object, and the other is used to learn the search area. We introduce a dense network with attention mechanism as the backbone of the tracking network. It contains multiple dense blocks that can be directly connected from any layer to all subsequent layers, in order to improve the information flow and gradients in the entire network, which makes them easy to train.

Inspired by the visual attention mechanism, we add a feature channel attention module to the dense network, which is called the global attention dense module in Fig.2. The attention mechanism is considered to automatically obtain feature information, with the purpose of automatically improving useful features and suppressing unimportant features. We use an improved global attention module in each dense block. The spatial dimension is compressed, and each feature map is globally pooled and averaged to a true value. To some extent, the real number has a global receptive field.

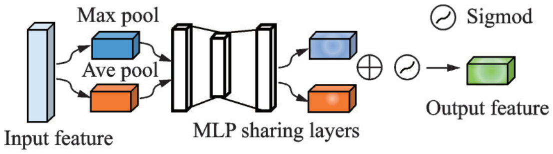


Fig.2 Structure of channel attention module

The Densenet used in our experiment has three dense blocks, and each dense block has an equal number of layers. A convolution with 16 output channels is performed on the input image. For a convolutional layer with a kernel size of 3×3, each side of the input is filled with one pixel to keep the feature map size of each layer fixed. We use 1×1 convolution, and then use 2×2 average pooling as a transition layer between two consecutive dense blocks. At the end of dense block, the network performs global average pooling, and then attaches a SoftMax classifier.

2. 3 Corner detection

We detect an object as a pair of key points, namely the top-left corner and the bottom-right corner of the bounding box. The convolutional network predicts two heatmaps to represent the position of the corners of different object categories, as one for the upper left corner and the other for the lower right corner. The network also predicts the embedding vector (embedding) of each detected corner point, so that the embedding distance between the two corner points and the same object is close. In order to generate a tighter bounding box, the network also predicts the offset to slightly adjust the position of the corner points. Using the predicted heatmap, embedding, and offset, we apply a simple post-processing algorithm to obtain the final bounding box.

We combine different information to get the top-left corner pooling and the bottom-right corner pooling. The structure of the pooling module in the top-left corner is shown in Fig.3. We add the top pooling to the left pooling, and the left pooling to the top pooling, and then the two parts are combined to get the top-left pooling. The bottom-right pooling is also processed following the above process. We add the right pooling to the bottom pooling, and the bottom pooling to the right pooling, and then combine the two to get the bottom right pooling. In this way, the corner points can get richer

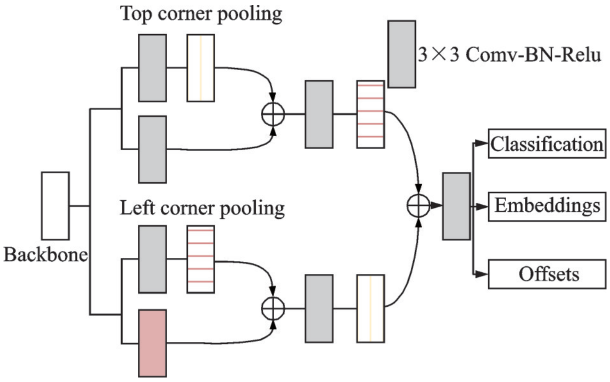


Fig.3 General structure of Corner detection

object information.

We calculate the top-left branch and the bottom-right branch, and then predict that the heatmap representing the position of the corner point. The corner point contains the target embedding vector, as the classification part and the offset of the corner point pair to shorten the distance between the targets. We use a simple post-processing method to locate the final target bounding box.

2.4 Tracking process

The overall network structure is mainly divided into two parts: Feature extraction and corner detection. Through the siamese network, the template frame and search frame are extracted separately. And the siamese network can be used to make the two branches share weighs to ensure that the conversion of each branch is similar. At the same time, in order to reduce calculation, we increase the convolution level to reduce the image feature resolution. In the feature extraction part, the dense network with channel attention mechanism we designed can be used to better spread the feature information and alleviate gradient disappearance. Then the corner detection is performed on the feature information, and the classification, regression and embedding information can be obtained.

The procedure of the proposed object track algorithm is summarized as follows[26].

Step 1 Train the network to generate model.

Step 2 Take a block near the target in the first frame, then resize it to a size of 127 pixel×127 pixel.

Step 3 Send the trained model and save weights for regression, classification and embedding, and do not update later.

Step 4 In the search frame, take the previous frame position as the center and intercept a relatively large area, then resize it to 255 pixel×255 pixel.

Step 5 Use the weight for classification, regression and embedding retained in the first frame to deconvolve the corresponding features extracted from the search frame; get the classification, regression vector and object information of the upper left corner and the lower right corner, so as to locate the tracking object position.

The overall flow chart of the method is shown in Fig.4.

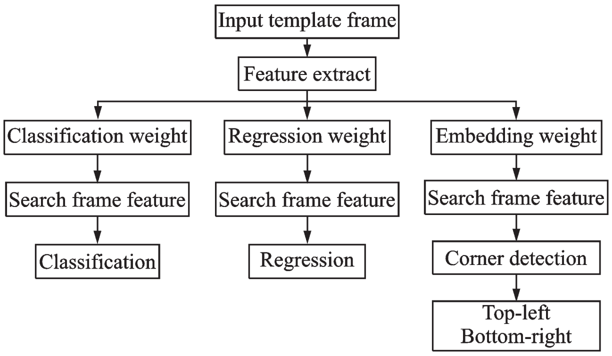


Fig.4 Overall flowchart of the proposed method

2.5 Loss function

We use the "pull" loss L_{pul} to group the corner points belonging to the object, and the "push" loss L_{pus} to separate the corner points between the foreground and the background. We also use focus loss L_{att} to predict the attention map. We optimize the complete training loss function by combining the above loss functions to train our network end-to-end

$$L=\alpha L_{\mathrm{att}}+\gamma L_{\mathrm{cls}}+\gamma L_{\mathrm{emb}}+\delta L_{\mathrm{off}} \tag{1}$$

where $L_{\mathrm{emb}}=\theta L_{\mathrm{pul}}+\omega L_{\mathrm{pus}}$ denotes the loss of embedding; α, β, γ and δ denote the weights for balance the full training loss.

3 Experiments

3.1 Training dataset and evaluation training

Our method is implemented in Python3.9 using PyTorch and run on NVIDIA Tesla V100 GPU. The backbone network of our architecture[12] is not pretrained on any dataset. We train the network on the training sets of COCO[27], ImageNet DET[28] and YouTube-Bounding Boxes Dataset[29] and to learn how to measure the similarity between general objects for the general concept of visual tracking. In training and testing, we set the input size of the template patch to 127 pixel×127 pixel, and the size of the search patch to 255 pixel×255 pixel. The code will be published on GitHub.

We focus on the single object tracking on OTB2015[30], LaSOT[31] and UAV123[32]. We use homemade UAV datasets to evaluate the tracking effect of UAVs in the long-term tracking.

3.2 Comparison

3.2.1 OTB-2015 Dataset

The standardized OTB benchmark provides a fair testbed on robustness. The siamese based tracker formulates the tracking as one-shot detection task without any online update, thus resulting in inferior performance on this no-reset setting benchmark. However, we identify the limited representation from the shallow network as the primary obstacle, which prevents siamese based trackers from surpassing top-performing methods, such as C-COT variants. We compare our tracker on the OTB2015 with seven state-of-the-art trackers including SiamRPN[13], CREST[33], SINT[34], CFNet[6], SiamFC[8], Staple[35], and MemTrack[20]. Fig.5 shows the comparison success and precision plots of OPE on OTB-2015 Dataset. It indicates that our tracker produces outstanding results in both plots.

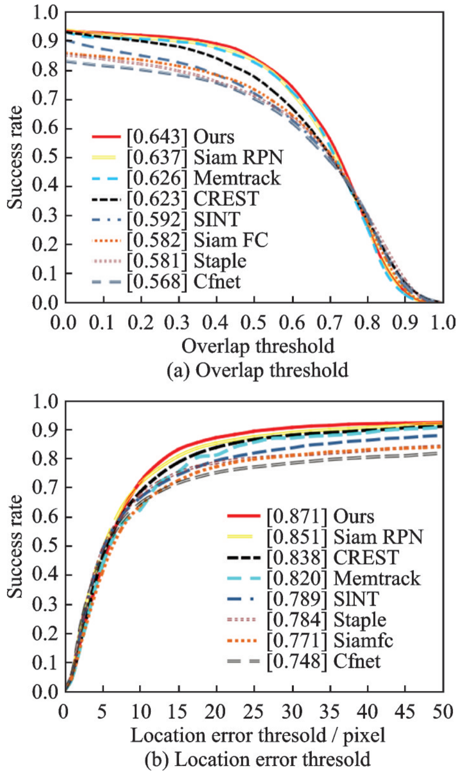


Fig.5 Evaluation results of trackers on OTB-2015 Dataset

3.2.2 UAV123 Dataset

UAV123 dataset includes 123 sequences with an average sequence length of 915 frames. Memtrack is a tracking algorithm that uses the attention mechanism. Our method also uses channel attention to enhance tracking performance. In addition to comparison with the Memtrack tracker, ECO, ECO-HC, KCFDP, SiamRPN, SiamFC, Staple, and SAMF are all added on comparison. Fig.6 illustrates the precision and success plots of the compared trackers. Specifically, our tracker achieves a success score of 0.562, outperforming SiamRPN (0.557) and ECO (0.525) with a large margin. Our tracker can get the best results in both success plot and precision plot, significantly outperforming the baseline SiamRPN and other approaches.

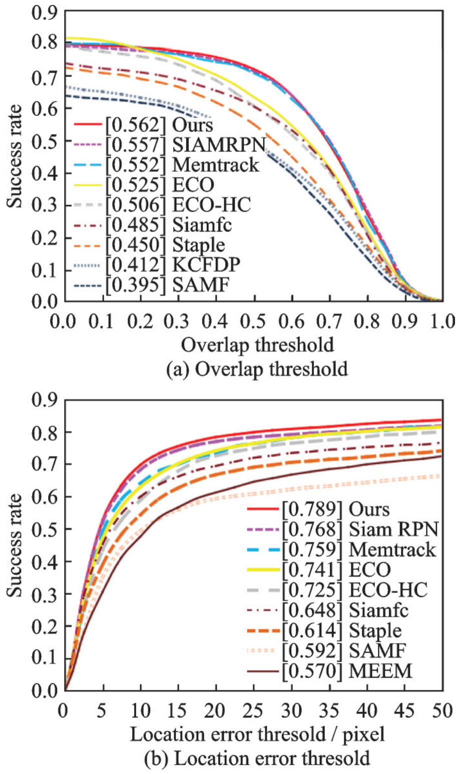
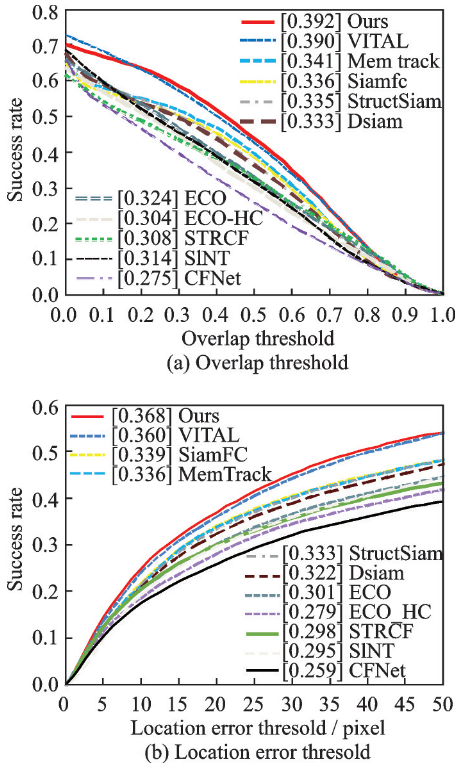


Fig.6 Comparisons among recent update trackers on UAV123

3.2.3 LaSOT Testing Set

In order to further verify the proposed framework on a larger and more challenging dataset[36], we conducted experiments on LaSOT Testing Set. The dataset provides large-scale, high-quality dense annotations, with a total of 1 400 videos and 280 videos in the testing set. Fig.7 reports the overall performance of our tracker. Without bells and whistles, our model is sufficient to achieve a state-of-the-



(a) Overlap threshold

(b) Location error thresold

Fig.7 Comparisons on the LaSOT with state-of-the-art tracking methods by using success and precision plots

art accuracy (AUC) score of 39.2%. The leading results on such a large dataset show that our method has good generalization ability for object tracking.

3.2.4 VOT

There are about 60 sequences in VOT2016, and ten different sequences in VOT2018. VOT datasets have three commonly used evaluation indicators: Accuracy, robustness, and expected average overlap (EAO). The accuracy is used to evaluate the average overlap between the predicted bounding box and the ground truth box during successful tracking. The robustness is used to evaluate the failure rate. EAO combines accuracy and robustness. We evaluate our tracker on VOT2018 and the comparation result is show in Table 1. Through the table, we can see that our designed trackers get the highest value among the three classical evaluation criteria.

Table 1 Comparison with the state-of-the-art in terms of EAO, robustness, and accuracy on the VOT2018 benchmark

Tracker	Accuracy	Robustness	EAO
SiamRPN	0.586	0.276	0.383
MFT	0.505	0.140	0.385
DaSiamRPN	0.569	0.337	0.326
DeepSTRCF	0.523	0.215	0.345
Ours	0.593	0.255	0.386

3.2.5 Homemade UAV datasets

We use our self-made UAV datasets to test the method we designed. The datasets acquire a large number of high-definition images including drone targets through real-life photography, web crawlers and data augmentation. The test results are shown in Fig.8. Our tracker can distinguish bright, dim and

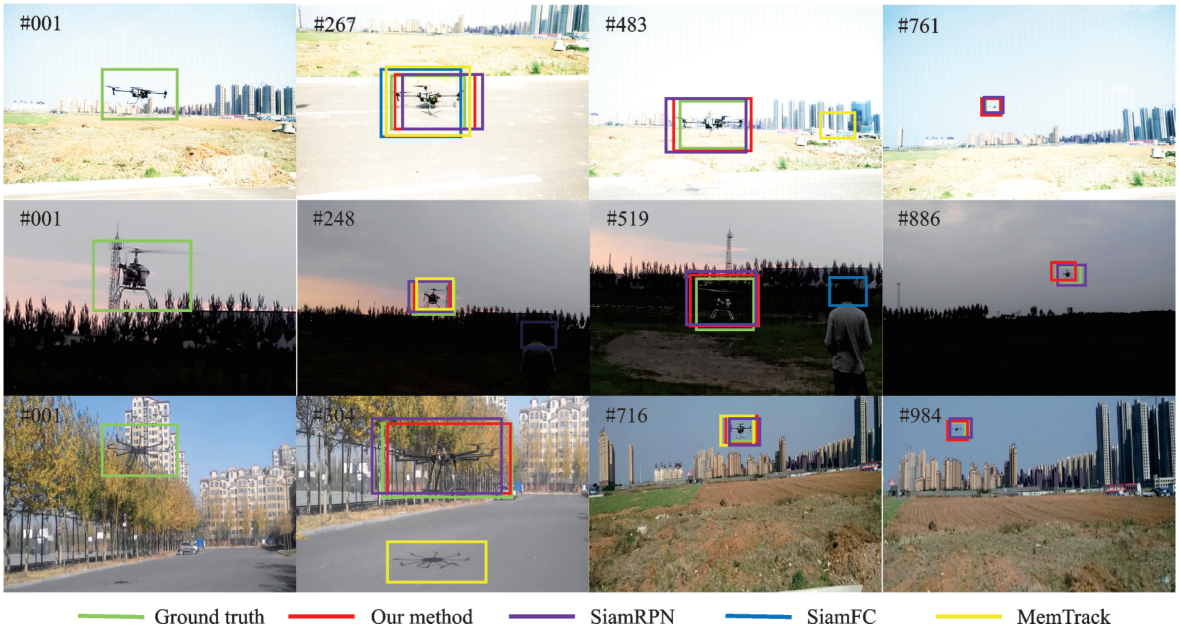


Fig.8 Example tracking results of our approach with other siamese network based trackers on the drone datasets

complex backgrounds from the object. The Memtrack and SiamFC may lost the target or confuse at certain frame. With the help of attention, dense and corner detections, our model can help to get better tracking result.

3.3 Ablation study

We study the impact of backbone architecture and the attention mechanism on VOT2017. We compare the Densenet-121 and Densenet-169 as backbone. The modified Densenet-169 adds attention mechanism (AM). And it can be seen that after adding the AM, tracker produces better results[37]. As shown in Table 2, under the same network, EAO can increase by 3.2% after adding AM, and speed only decays 4 frame/s.

Table 2 Comparison with the different backbone networks in terms of EAO and speed on the VOT2017 benchmark

Backbone	EAO	Speed/(frame•s^{-1})
Densenet-121	0.248	37
Densenet-169	0.252	36
Densenet-121+AM	0.256	33
Densenet-169+AM	0.261	32

4 Conclusions

In summary, we propose an effective siamese attentional dense network, SiamADN, for visual target tracking. By using channel attention module, the network integrates global and local information and contextual information. We do not perform multi-scale search, in order to further simplify the search process method. SiamADN uses the state of a pair of corner points with a cascaded corner point pooling, and designs anchors without any prior knowledge. Extensive experiments on four different tracking benchmarks can be verified, and our method achieves the most advanced performance. The tracker we designed can track UAVs very well.

References

[1] GUO D, WANG J, CUI Y, et al. SiamCAR: Siamese fully convolutional classification and regression for visual tracking[EB/OL]. (2019-12-13) [2021-04-01] https://arxiv.org/abs/1911.07241.

[2] WU Y, LIM J, YANG M H. Online object tracking: A benchmark[C]//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE, 2013: 2411-2418.

[3] DANELLJAN M, HGER G, KHAN F S, et al. Accurate scale estimation for robust visual tracking [C]// Proceedings of the British Machine Vision Conference (BMVC). Nottingham, UK: BMVA, 2014: 1-11.

[4] HENRIQUES J F, CASEIRO R, MARTINS P, et al. High-speed tracking with kernelized correlation filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583-596.

[5] LI Y, ZHU J. A scale adaptive kernel correlation filter tracker with feature integration[C]//Proceedings of Computer Vision-ECCV 2014 Workshops. [S. l.] : Springer, 2014: 254-265.

[6] XU H, GAO Y, YU F, et al. End-to-end learning of driving models from large-scale video datasets[EB/OL]. (2017-07-23) [2021-04-01]. https://arxiv.org/abs/1911.07241.

[7] LI B, WU W, WANG Q, et al. SiamRPN++ : Evolution of Siamese visual tracking with very deep networks[EB/OL]. (2018-12-31) [2021-04-01]. https://arxiv.org/abs/1812.11703.

[8] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional siamese networks for object tracking[C]//Proceedings of Computer Vision-ECCV 2016 Workshops. [S. l.] : Sprigner, 2016: 850-865.

[9] VALMADRE J, BERTINETTO L, HENRIQUES J F, et al. End-to-end representation learning for correlation filter based tracking[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017: 5000-5008.

[10] NAM H, HAN B. Learning multi-domain convolutional neural networks for visual tracking[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016: 4293-4302.

[11] GUO Q, FENG W, ZHOU C, et al. Learning dynamic siamese network for visual object tracking[C]// Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 1781-1789.

[12] WANG Q, TENG Z, XING J, et al. Learning attentions: Residual attentional siamese network for high performance online visual tracking[C]//Proceedings

of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 4854-4863.

[13] REN S, HE K, GIRSHICK R B, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Trans Pattern Anal Mach Intell, 2017, 39(6): 1137-1149.

[14] LI B, YAN J, WU W, et al. High performance visual tracking with siamese region proposal network[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 8971-8980.

[15] ZHU Z, WANG Q, LI B, et al. Distractor-aware siamese networks for visual object tracking[C]//Proceedings of Computer Vision. [S.l.]: Springer, 2018: 103-119.

[16] WANG Q, ZHANG L, BERTINETTO L, et al. Fast online object tracking and segmentation: A unifying approach[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019: 1328-1338.

[17] ZHANG Z, PENG H. Deeper and wider siamese networks for real-time visual tracking[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019: 4591-4600.

[18] WANG G, LUO C, XIONG Z, et al. SPM-tracker: Series-parallel matching for real-time visual object tracking[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019: 3643-3652.

[19] HE A, LUO C, TIAN X, et al. A two fold siamese network for real-time object tracking[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 4834-4843.

[20] YANG T, CHAN A B. Learning dynamic memory networks for object tracking[C]//Proceedings of Computer Vision—ECCV. [S. l.] : Springer, 2018: 153-169.

[21] YANG Kang, SONG Huihui, ZHANG Kaihua. Real-time visual tracking based on dual attention Siamese network[J]. Journal of Computer Applications, 2019, 39(6): 1652-1656.(in Chinese)

[22] QI Tianhui, ZHANG Hui, LI Jiafeng, et al. Siamese network with multi-attention map for visual object tracking[J]. Journal of Signal Processing, 2020, 36(9): 1557-1566.(in Chinese)

[23] DANELLJAN M, BHAT G, KHAN F S, et al. ECO: Efficient convolution operators for tracking[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017: 6931-6939.

[24] HUANG L, ZHAO X, HUANG K. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild[J]. IEEE Trans Pattern Anal Mach Intell, 2021, 43(5): 1562-1577.

[25] BROMLEY J, BENTZ J W, BOTTOU L, et al. Signature verification using a “Siamese” time delay neural network[J]. Int J Pattern Recognit Artif Intell, 1993, 7(4): 669-688.

[26] LI H, DENG Y, XU X, et al. Eagle-vision-based object detection method for UAV formation in hazy weather[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2020, 37(4): 517-527.

[27] LIN T, MAIRE M, BELONGIE S J, et al. Microsoft COCO: Common objects in context[C]//Proceedings of Computer Vision.[S.l.]: Springer, 2014: 740-755.

[28] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. Int J Comput Vis, 2015, 115(3): 211-252.

[29] REAL E, SHLENS J, MAZZOCCHI S, et al. YouTube-BoundingBoxes: A large high-precision human-annotated data set for object detection in video[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017: 7464-7473.

[30] WU Y, LIM J, YANG M. Object tracking benchmark [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834-1848.

[31] KRISTAN M, LEONARDIS A, MATAS J, et al. The sixth visual object tracking vot2018 challenge results[C]//Proceedings of Computer Vision—ECCV 2018 Workshops. [S.l.]: Springer, 2018: 3-53.

[32] MUELLER M, SMITH N, GHANEM B. A benchmark and simulator for UAV tracking[C]//Proceedings of Computer Vision—ECCV 2016. [S. l.] : Springer, 2016: 445-461.

[33] SONG Y, MA C, GONG L, et al. CREST: Convolutional residual learning for visual tracking[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 2574-2583.

[34] TAO R, GAVVES E, SMEULDERS A W M. Siamese instance search for tracking[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016: 1420-1429.

[35] BERTINETTO L, VALMADRE J, GOLODETZ S, et al. Staple: Complementary learners for real-time tracking[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016: 1401-1409.

[36] WANG E, WANG D, HUANG Y, et al. Siamese attentional cascade keypoints network for visual object tracking[J]. IEEE Access, 2021(9): 7243-7254.

[37] LYU Z, PAN F, XU X. A novel deep neural network compression model for airport object detection[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2020, 37(4): 562-573.

Acknowledgements This study was supported by the Zhejiang Key Laboratory of General Aviation Operation Technology (No. JDGA2020-7), the National Natural Science Foundation of China (No. 62173237), the Natural Science Foundation of Liaoning Province (No. 2019-MS-251), the Talent Project of Revitalization Liaoning Province (No. XLYC1907022), the Key R & D Projects of Liaoning Province (No.2020JH2/10100045), and the High-Level Innovation Talent Project of Shenyang (No.RC190030).

Authors Dr. **WANG Zhi** received the Ph. D. degree in mechanical engineering from Beijing University of Technology, China, in 2020. He is currently a lecturer in Civil Aviation Management Institute of China. His current research interests mainly concern unmanned aerial vehicle operation and route optimization.

Prof. **WANG Ershen** received the Ph.D. degree from Dalian Maritime University in 2009. He is a professor at College of Electronic and Information Engineering in Shenyang Aerospace University. His research currently focuses on global navigation satellite system (GNSS) positioning algorithms, and UAV surveillance technology.

Author contributions Dr. **WANG Zhi** designed the tracking model, analysed the data and wrote the original paper. Prof. **WANG Ershen** contributed to the model and interpretated the tracking results. Dr. **HUANG Yufeng** contributed to the tracking model and discussed the background. Mr. **YANG Siqi** contributed to the background and discussed the result. Mr.**XU Song** contributed to data analysis. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor:ZHANG Bei)

SiamADN:用于无人机目标跟踪的孪生注意密集网络

王 志[1,2],王尔申[3],黄煜峰[3],杨斯淇[3],徐 嵩[3]

(1.浙江建德通用航空研究院浙江通用航空运行技术研究重点实验室,建德311612,中国;
2.中国民航管理干部学院通用航空系,北京 100102,中国;
3.沈阳航空航天大学电子信息工程学院,沈阳110136,中国)

摘要:基于深度学习的单目标跟踪技术在诸多计算机视觉应用中取得了卓越的效果。然而,现有跟踪器在图像目标变形、遮挡、移动等场景下的应用具有局限性。本文提出一种融合注意力机制和密集网络的孪生网络离线跟踪方法(SiamADN),重点针对无人机等小目标的跟踪任务。首先,算法应用密集网络来减少消失梯度,从而加强特征转移;其次,在密集网络结构中引入通道注意力机制,增强感兴趣的关注度;引入高级角点检测网络以改进持续跟踪过程。在OTB-2015、UAV123,LaSOT 和VOT等跟踪数据集上进行了广泛的实验。实验结果表明,在UAV123上的准确率为78.9%,运行速度为每秒32帧,具有较好的实际应用效果。

关键词:无人机;目标跟踪;密集网络;角点检测;孪生网络