IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 12, 2019, accepted June 2, 2019, date of publication June 6, 2019, date of current version June 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2921315

Remote Sensing Target Tracking in UAV Aerial Video Based on Saliency Enhanced MDnet

FUKUN BI[1], MINGYANG LEI[1], YANPING WANG[1], AND DAN HUANG[2]
[1]School of Information Science and Technology, North China University of Technology, Beijing 100144, China
[2]China Research and Development Academy of Machinery Equipment, Beijing 100089, China

Corresponding author: Yanping Wang (newauthor@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61601006, in part by the Beijing Natural Science Foundation under Grant 4192021, and in part by the Equipment Pre-Research Foundation under Grant 61404130312.

ABSTRACT Remote sensing target tracking in the aerial video from unmanned aerial vehicles (UAV) plays a key role in public security. As the UAV aerial video has rapid changes in scale and perspective, few pixels in the target region, and multiple similar disruptors, and the main tracking methods in this research field generally have relatively low tracking performance and timeliness, we propose a remote sensing target tracking method for the UAV aerial video based on a saliency enhanced multi-domain convolutional neural network (SEMD). First, in the pre-training stage, we combine the least squares generative adversarial networks (LSGANs) with a multi-orientation Gaussian Pyramid to augment typical easily confused negative samples for enhancing the capacity to distinguish between targets and the background. Then, a saliency module was integrated into our tracking network architecture to boost the saliency of the feature map, which can improve the representation power of a rapid dynamic change target. Finally, in the stage for generating tracking samples, we implemented a local weight allocation model to screen for hard negative samples. This approach can not only improve the stability in tracking but also boost efficiency. The comprehensive evaluations of public and homemade hard datasets demonstrate that the proposed method can achieve high accuracy and efficiency results compared with state-of-the-art methods.

INDEX TERMS Visual tracking, multi-domain learning, saliency enhanced, sample augmentation.

I. INTRODUCTION

With the popularization of high-resolution imaging technology and the progress of artificial intelligence, target tracking in remote sensing video received much attention. One of the important parts is remote sensing target tracking in UAV aerial video due to its significance in military reconnaissance, land monitoring and criminal tracking. Although many object tracking algorithms [1], [2] have been proposed, there are still challenges due to the swaying of UAV platforms, which collect videos with high frequency scale and orientation changes, few pixels in target region, and the targets are often easily confused with the background. Furthermore, deep learning algorithms generally cannot meet the real-time requirements for UAV processing platforms. Therefore, designing an effective and robust tracking method for UAV aerial video remains challenging.

In recent years, many researchers have made efforts to facilitate target tracking research. Hare *et al.* [3] presented a structured output support vector machine for object tracking. Kalal *et al.* [4] proposed a visual tracking algorithm named TLD by improving an online learning mechanism. However, they all need to consume huge calculations, which cannot achieve real-time tracking in the UAV processing platform. Recent techniques based on correlation filters boost efficiency significantly. Bolme *et al.* [5] propose the minimum output sum of squared errors (MOSSE) filter, which works by finding the maximum cross correlation response between the model and candidate patch. Henriques *et al.* [6] improved adaptive performance for diverse scenarios using multichannel HOG features. Danelljan *et al.* [7] handled the scale change of target objects by learning adaptive multiscale correlation filters. However, these methods would generate multiple suspected responses when considering similar objects around the target, which usually occurs from the perspective of UAV platforms.

Recently, it was observed that deep networks have been successful in visual tracking. Wang *et al.* [8] developed a top layer and lower layer to obtain semantic features and

The associate editor coordinating the review of this manuscript and approving it for publication was Zhaoxiang Zhang.

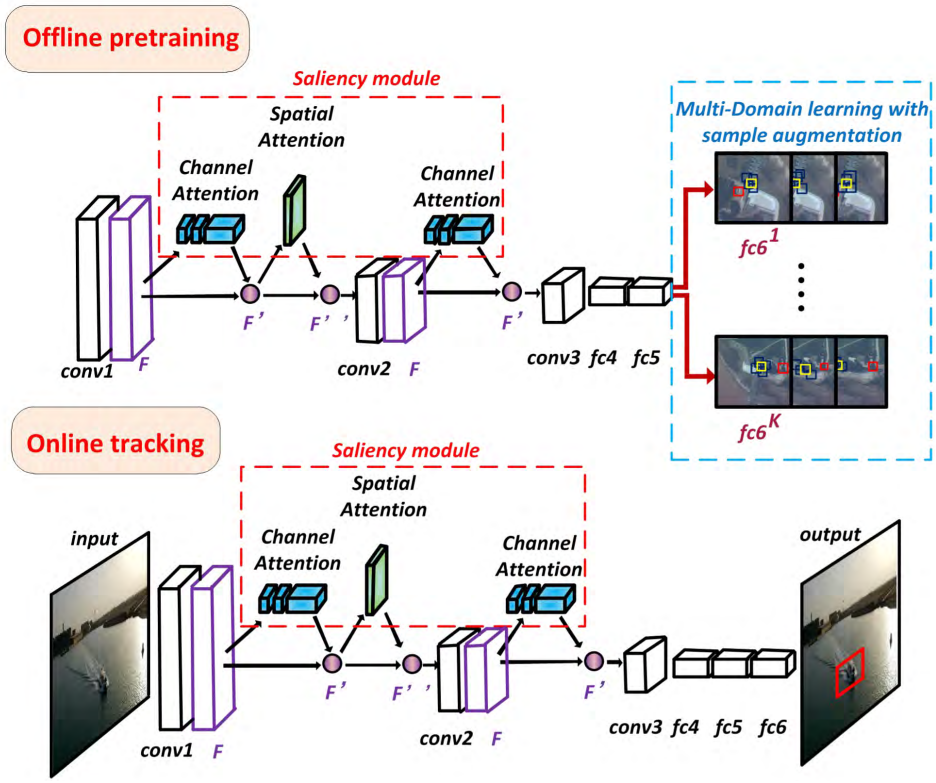

FIGURE 1. The proposed tracking network.

discriminative information separately for visual tracking. Zhang *et al.* [9] proposed simple two-layer convolutional neural networks to obtain feature maps for the tracker object for online object tracking. Both of these approaches had high tracking accuracy. However, they generally boosted performance through the design of the deep structure of the network, which greatly affects efficiency.

A popular CNN-based tracking algorithm with state-of-the-art accuracy and speed named MDNet [10] was presented recently. It pretrains a CNN using a large set of videos with tracking ground-truths to obtain a generic target representation that can enhance the adaptability of the network to various targets. However, from the perspective of UAV platforms, the target is generally small, and it can be easily confused with the background as well as become blurry from frequency scale and orientation changes. Meanwhile, MDNet [10] would produce many negative samples for training in every tracking frames, which leads to high computational costs. To address these issues, we propose a robust tracking approach for UAV videos, and the main contributions are summarized as follows.

(1) In the offline pretraining stage, we propose a typical easily confused negative sample augmentation strategy by combining LSGANs [11] with a multi-orientation Gaussian Pyramid to generate enough valid samples.

(2) For the design of tracking network structure, we embed a saliency module between convolutional layers and optimize the arrangement of its functional sub-modules to boost the saliency of the feature map, which improved the network representation power for rapid dynamic changes in the target.

(3) In the negative sample generation of tracking frames, a local weight allocation model is constructed for screening high-weight negative samples. This strategy can not only enhance the stability of tracking process but also effectively reduce the invalid samples network that needs to be learned.

Experimental results demonstrate that the proposed method can significantly improve tracking accuracy and efficiency compared to the state-of-the-art trackers in a UAV aerial video.

II. SALIENCY ENHANCED MULTI-DOMAIN CONVOLUTIONAL NEURAL NETWORK (SEMD) FOR TRACKING

A. NETWORK ARCHITECTURE

The proposed network consists of convolutional layers, a saliency module for enhancing the saliency of feature maps, and fully connected layers for binary classification as shown in Fig. 1. For offline pretraining, our algorithm pretrains a CNN to obtain a generic target representation using a large set of videos with tracking ground-truths. A lot of negative samples are generated by using typical easily confused negative sample augmentation for multi-domain learning. For online

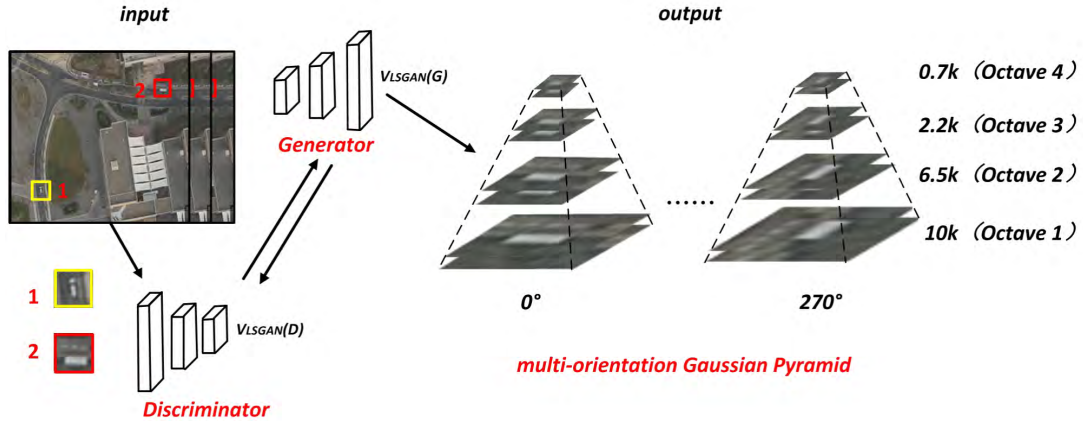

FIGURE 2. Typical easily confused negative sample augmentation. Yellow and red bounding boxes denote the positive and typical easily confused negative samples needed to augment each frame, respectively. The discriminator and generator are trained to augment negative samples. The multi-orientation Gaussian Pyramid generates negative samples in different scales and orientations.

tracking, we manually select tracking single target in the first frame, the network computes feature maps of these boxes through a single forward pass. Note that a CNN feature is refined from the saliency module, which can achieve suppression of background information effectively. Finally, the refined feature is fed to three fully connected layers for classification between targets and the background.

B. TYPICAL EASILY CONFUSED NEGATIVE SAMPLE AUGMENTATION

It is well known that the effect of pretraining has a vital influence on the success rate of tracking in a CNN-based algorithm. MDNet [10] constructs the shared layers, from which the model that is obtained has useful generic feature representations and generates positive and negative samples from different distributions. Although it performs excellently in the tracking from the widely applied head-up perspective, it has the opposite effect from the UAV perspective. The explanation for this result is that the sloshing of UAV platform will provide high frequency scale change of a target. If we use these different scale samples as pretraining directly, the tracking accuracy will be greatly compromised. Furthermore, typical easily confused targets in complex scenes are the main factors affecting tracking performance. Therefore, it is necessary to augment easily confused negative samples to enhance the robustness of our network. To deal with this problem, we use LSGANs [11] to produce similar samples with different definitions in the single frame. Compared to MDNet [10], our algorithm is outstanding at distinguishing a target from a typical ambiguous easily confused object. Fig. 2 shows the flowchart for augmentation.

Given a frame from a video sequence, the hand-crafted negative samples with easily confused frames are denoted by $\boldsymbol{x1}, \boldsymbol{x2} \ldots \boldsymbol{xt}$, and $\boldsymbol{t}$ is the number of samples that needed to be augmented. For each $\boldsymbol{x}$, LSGANs generate samples relying on discriminator $\boldsymbol{D}$ and a generator $\boldsymbol{G}$, and we train $\boldsymbol{D}$ and $\boldsymbol{G}$ through the following function:

$$\min_D V_{LSGAN}(D) = 0.5E_{(x)}[(D(x) - a)^2] + 0.5E_{(z)}[(D(G(z)) - b)^2] \quad (1)$$

$$\min_G V_{LSGAN}(G) = 0.5E_{(z)}[(D(G(z)) - c)^2] \quad (2)$$

where $\boldsymbol{b}$ and $\boldsymbol{a}$ are the labels for fake data and real data, respectively, and $\boldsymbol{c}$ denotes the value that $\boldsymbol{G}$ wants to make $\boldsymbol{D}$ believe fake data, and $\boldsymbol{z}$ is distributed normally.

In addition, we also collect samples with different sizes and orientations using a Gaussian Pyramid with the goal of simulating the UAV platform rotation, which generally appears in complex environments.

Then, the generated samples of $\mathbf{0.7}\boldsymbol{k}\textbf{-}\boldsymbol{th}$, $\mathbf{2.2}\boldsymbol{k}\textbf{-}\boldsymbol{th}$, $\mathbf{6.5}\boldsymbol{k}\textbf{-}\boldsymbol{th}$ and $\mathbf{10}\boldsymbol{k}\textbf{-}\boldsymbol{th}$ iterations were considered as an octave (first octave is $\mathbf{10}\boldsymbol{k}\textbf{-}\boldsymbol{th}$ iteration), and we use Gaussian filtering to produce another sample in every octave as follows:

$$G(\mathrm{r}) = \frac{1}{2\pi\sigma^2}\exp(-\frac{r^2}{2\sigma^2}) \quad (3)$$

$$r = \sqrt{x^2 + y^2} \quad (4)$$

where $\boldsymbol{r}$ denotes the row and column numbers of the samples, $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, and σ is set to 2.5 empirically.

Furthermore, for adjacent octaves, we employ sampling to change the size of the next octave to 1/4 of the previous octave, and we rotate all samples every 90 degrees.

Through this approach, the multi-orientation Gaussian Pyramid is constructed to collect negative samples in different scales and orientations from multiple perspectives. Specially, We mainly select the objects with similar appearance characteristics and sizes to the target for augmentation, such as the roof (red bounding box) shown in Fig. 2.

C. SALIENCY MODULE IN A NETWORK

As discussed earlier, MDNet uses a shallow network considering the real-time requirements. However, from the

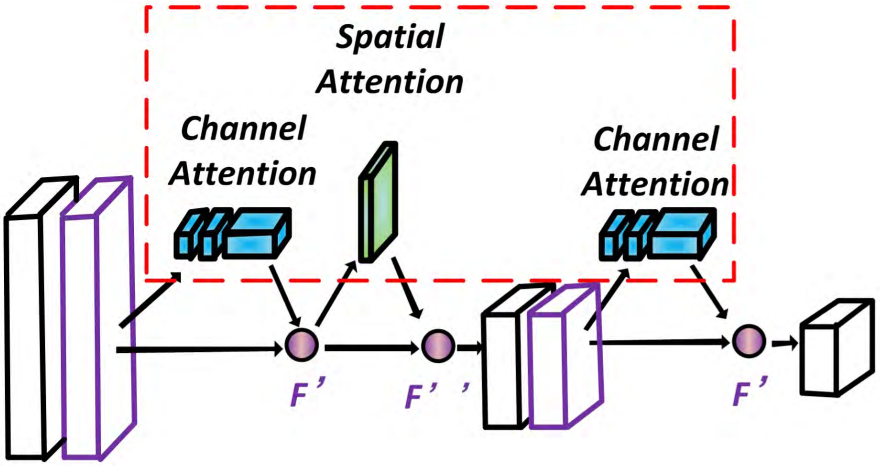

FIGURE 3. The saliency module.

perspective of a UAV platform, remote sensing scenes usually have few pixels in target region and have complex backgrounds. Therefore, it is necessary to enhance feature saliency. CBAM [12] is a lightweight and general module that can be integrated into any CNN architectures seamlessly with negligible overheads, and it can boost feature saliency. It consists of two sub-modules, namely, the Channel attention module and Spatial attention module, as shown in Fig. 3.

Given a feature map $\boldsymbol{F} \in \boldsymbol{R}^{\boldsymbol{C*H*W}}$ as input, CBAM sequentially infers a channel attention map $\boldsymbol{M_c} \in \boldsymbol{R}^{\boldsymbol{C*1*1}}$ and a spatial attention map $\boldsymbol{M_s} \in \boldsymbol{R}^{\boldsymbol{1*H*W}}$. The overall attention process can be summarized as follows:

$$F' = M_C(F) \bullet F \tag{5}$$

$$F'' = M_S(F') \bullet F' \tag{6}$$

where $\bullet$ denotes element-wise multiplication, and $\boldsymbol{M_c(F)}$ and $\boldsymbol{M_s(F)}$ is computed as follows:

$$M_C(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{7}$$

$$M_S(F) = \sigma(f^{7\times7}([AvgPool(F)); MaxPool(F)])) \tag{8}$$

where σ denotes the sigmoid function, $\boldsymbol{MLP}$ is the weight of average-pooling and max-pooling, and $\boldsymbol{f^{7\times7}}$ represents a convolution operation with a filter size of 7×7.

As validated in the general network architecture, feature saliency can be boosted through $\boldsymbol{M_c}$ and $\boldsymbol{M_s}$ in turn. However, the pixels in target region are generally very few in UAV aerial video, so it is particularly vital to enhance the initial feature saliency. We add an adaptive convolution layer after average-pooling of $\boldsymbol{M_c}$ to change the output channel of the previous convolution layer to twice the original, which can tune the channel adaptively to ensure the subsequent process is running smoothly. Meanwhile, only channel attention module $\boldsymbol{M_c}$ is embedded after the second convolution layer to give consideration to both timeliness and tracking accuracy. We provide experimental results with different arrangements in Section IV.

D. HARD NEGATIVE SAMPLE SCREENING BASED ON LOCAL WEIGHT ALLOCATION

In the sample generation stage of every tracking frame, MDNet collects 200 negative samples from the whole image. However, in the remote sensing target tracking from the UAV perspective, the relative displacement between frames is limited. On the other hand, easily confused objects around targets are generally the main factor for performance degradation rather than distant objects. To screen valuable negative samples, a local weight allocation is constructed, which accelerates the hard negative mining as follows:

$$d = e^{-(\frac{s_i}{100})^2/2\sigma^2} \tag{9}$$

where $\boldsymbol{s_i}$ denotes the distance between the ***i-th*** negative sample centre and the target centre of the previous frame, and σ is set to 2.5 empirically.

The distance $\boldsymbol{d}$ represents the value of the negative sample for network training. Setting the threshold $\boldsymbol{T}$ (taken as 0.007 empirically) when a $\boldsymbol{d}$ satisfies the following:

$$d < T \tag{10}$$

the negative sample is retained and vice versa. Finally, all valuable samples are screened out. Then, an efficient hard negative mining technique is incorporated in the learning procedure and we got the highest score of 48 negative samples and re-entered the network for fine-tuning the fully connected layers, which can make the network become more discriminative.

III. IMPLEMENTATION DETAILS

A. TARGET CANDIDATE GENERATION

In order to estimate the target position in each frame, a lot of target candidates sampled around the previous target position are evaluated using our tracking network, and we obtain their positive scores and negative scores from the network. The optimal target is given by finding the candidate with the maximum positive score.

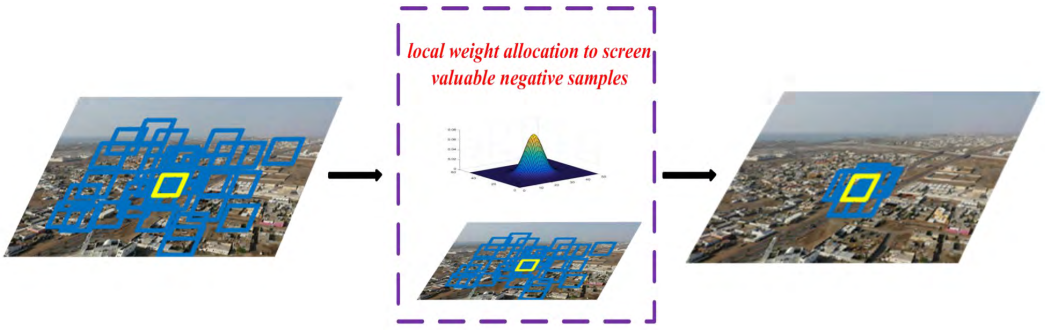

FIGURE 4. Screening of samples. The yellow and blue bounding boxes denote the ground-truth and negative samples, respectively.

B. OFFLINE PRETRAINING

For multi-domain learning with K training sequences, we train the network for 100K iterations, and the learning rates are set to 0.0002 for convolutional layers and 0.0015 for fully connected layers. For each iteration of offline pretraining, we collect 50 positive and 200 negative samples from every frame, where positive and negative samples have overlap ratio $r \geq 0.7$ and $r \leq 0.5$ IOU overlap ratios with ground-truth bounding boxes, respectively. For bounding-box regression, we use 1000 training samples with the same parameters as MDNet, and a minibatch is composed of 128 examples—32 positive and 96 negative samples. Note that our tracking network is trained by using Stochastic Gradient Descent (SGD) and receives a 107×107 RGB input.

C. ONLINE LEARNING

Similarly, for online learning, we collect 40 positive and 48 negative samples with $r \geq 0.7$ and $r \leq 0.3$ IOU overlap ratios with the estimated target bounding boxes, respectively. The weight decay and momentum are set to 0.0004 and 0.75, respectively. The learning rates are set to 0.0002 for convolutional layers and 0.0015 for fully connected layers. Note that we train the fully connected layers for 40 iterations at the initial frame of a new remote sensing test sequence.

IV. EXPERIMENTS

In this section we present our results for multiple datasets with comparisons to state-of-the-art tracking algorithms for UAV123 and a homemade dataset, and analysed the performance of our tracker through ablation studies and experiments with different arrangements of CBAM functional sub-modules.

A. EVALUATION METHODOLOGY

We evaluated our tracker, denoted by SEMD, with two datasets including (1) UAV123 [13] and (2) a homemade dataset with an average of 23 seconds per video. There are seventy fully annotated HD videos with challenging scenes captured from a UAV in the homemade dataset. Apart from a small number of the challenges of conventional datasets, these video sequences main suffer from further difficulties such as target region has few pixels, rapid changes in scale and perspective, and multiple easily confused disruptors, and the homemade dataset contains a wide variety of targets including cars, trucks and persons. For comparison, we employed several state-of-the-art trackers including ECO [14], MDNet [10], SRDCF [15], MCPF [16], BACF [17], PTAV [18], CFNet [19] and DSST [7]. These algorithms were implemented in PyTorch with a 3.5 GHz Intel Core I7-7800X and NVIDIA Titan V GPU.

We follow the evaluation protocol presented in a standard benchmark [20], where the performance of trackers is evaluated based on two criteria: centre location error and bounding box overlap ratio, and the performance is visualized using precision and success plots. The two plots are generated by computing ratios of successfully tracked frames at a set of different thresholds in the two metrics. The ranks of trackers are determined by the accuracy at the 20-pixel threshold in the precision plot. In the success plot, the Area Under Curve (AUC) scores of individual trackers are used to rank the trackers. Note that we used the same parameters for all of the tested datasets.

B. EXPERIMENTS ON DIFFERENT ARRANGEMENTS OF CBAM FUNCTIONAL SUB-MODULES

To verify the effectiveness of the proposed arrangement of channel attention and spatial attention sub-modules mentioned in Section II, we compared five different approaches embedded after the two convolution layers, as shown in Table 1. Let $\boldsymbol{J1}$ and $\boldsymbol{J2}$ denote the first and second convolution layer, respectively. Then, $\boldsymbol{C}$ and $\boldsymbol{S}$ denote channel attention and spatial attention sub-modules, respectively.

As presented in Section II, embedding $\boldsymbol{C}$ and $\boldsymbol{S}$ after $\boldsymbol{J1}$, and only embedding $\boldsymbol{C}$ after $\boldsymbol{J2}$ can achieve higher tracking accuracy and speed. Regarding the cause of this result, we think that in the subsequent tracking process after $\boldsymbol{J1}$, we should pay more attention to the target itself rather than the location of the target. In other words, this outcome occurred because two functional sub-modules, channel and spatial, compute complementary attention and focus on 'what' and 'where', respectively. Considering this result, we think in a tracking network for a remote sensing target, the key is determining

TABLE 1. Impacts of different arrangements of CBAM functional sub-modules.

	J1:C+S J2:C+S	J1:C J2:C+S	J1:S J2:C+S	J1:C+S J2:C	J1:C+S J2:S
Succ(%)	63.2	59.7	58.8	63.0	57.1
Prec(%)	78.4	69.2	67.5	78.4	66.0
FPS	34	53	50	52	44

TABLE 2. Internal comparison results.

Augmentation	Saliency module	Weight allocation	Succ(%)	Prec(%)	FPS
	√	√	59.7	75.3	52
√		√	60.3	77.0	68
√			59.9	75.6	36
√	√		61.1	77.2	33
√	√	√	63.0	78.4	52

how to enhance the saliency of target features and distinguish the target from background better, which represents 'what'. Therefore, 'where' should not be the focus of our discussion of 'two-category' tracking.

C. ABLATION STUDY

We perform several studies on UAV123 [13] to investigate the effectiveness of individual components in the proposed tracking algorithm. In this study, we used a small amount of the homemade UAV remote sensing dataset and UAV123 dataset for pretraining, and then remove them during testing.

Table 2 presents several options implemented in the network, where 'Augmentation', 'Saliency module', and 'Weight allocation' denote our proposed three mechanisms. According to our experiments, using all three mechanisms is helpful for improving success and precision rates the most. Therefore, the results prove that each component makes a meaningful contribution to tracking performance improvement.

D. EVALUATIONS ON UAV123 AND HOMEMADE UAV DATASET

We analysed our algorithm using two different datasets, including (1) a UAV123 dataset that consists of 50 fully annotated videos with various challenging attributes, and (2) a fully annotated dataset we built by collecting remote-sensing videos taken by UAV and selecting some of them for pretraining then removing them during evaluation. We added some videos that had certain characteristics, including few pixels in target region, rapid changes in scale and perspective, and multiple easily confused disruptors. Fig. 5 and Fig. 6 show the results of the nine trackers for UAV123 and the homemade dataset, respectively. The results clearly show that SEMD outperforms all the tested trackers significantly in different datasets. These outcomes can be attributed to the introduction of typical easily confused sample augmentation, a saliency module and sample screening based on three mechanisms of local weight allocation. Typical easily confused sample augmentation can enhance the capacity to distinguish between a target and background under multi-scale and multi-rotation orientations in a remote sensing scene. The saliency module boosted feature saliency by embedding lightweight and general modules. Sample screening based on local weight allocation can screen valuable negative samples for training. All three mechanisms can improve the success and precision rate to a certain extent. Note that in the homemade dataset, our algorithm is more obviously ahead of other algorithms. This result occurred because most of the collected videos in this dataset contain challenging scenes from the perspective of UAV, and our approach focuses on solving these issues as described above.

Table 3 and Table 4 present the overall performance of trackers of the experiment, including our algorithm, in terms of success rate, precision rate at 20 pixels, and speed

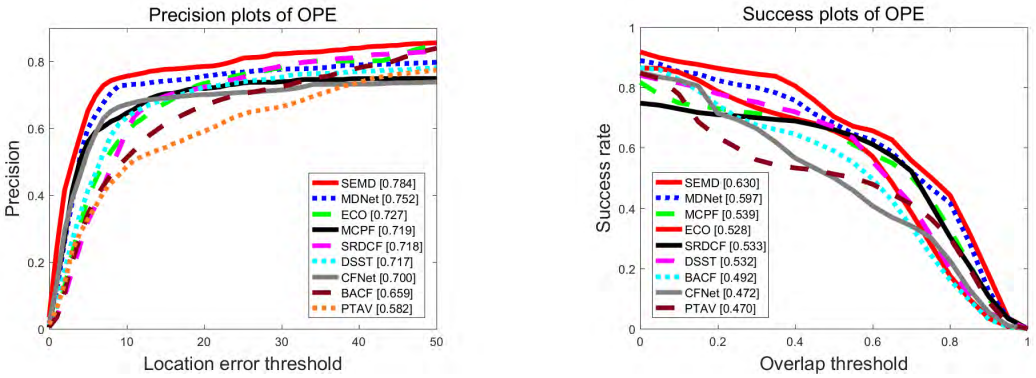

FIGURE 5. Quantitative results for UAV123 [14].

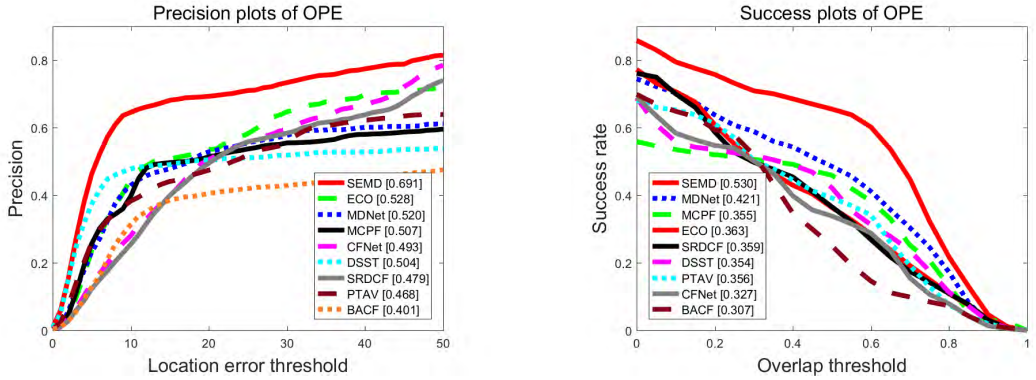

FIGURE 6. Quantitative results for a homemade dataset.

TABLE 3. Quantitative comparisons for UAV123.

	SEMD	MDnet	ECO	MCPF	SRDCF	DSST	CFNet	BACF	PTAV
Succ(%)	63.0	59.7	52.8	53.9	53.3	53.2	47.2	49.2	47.0
Prec(%)	78.4	75.2	72.7	71.9	71.8	71.7	70.0	65.9	58.2
FPS	52	45	61	13	6	17	36	29	11

TABLE 4. Quantitative comparisons for the homemade dataset.

	SEMD	MDnet	ECO	MCPF	SRDCF	DSST	CFNet	BACF	PTAV
Succ(%)	53.0	42.1	36.3	35.5	35.9	35.4	32.7	30.7	35.6
Prec(%)	69.1	52.0	52.8	50.7	47.9	50.4	49.3	40.1	46.8
FPS	47	39	70	7	3	3	31	27	9

for two datasets, respectively. Although SEMD has higher complexity than MDnet, the proposed method outperforms MDnet in the processing speed. We think the reasons can be summarized as follows: (1) we adopt two update strategies as in MDNet: short-term and long-term for improving adaptiveness and robustness of target, respectively. Long-term updates

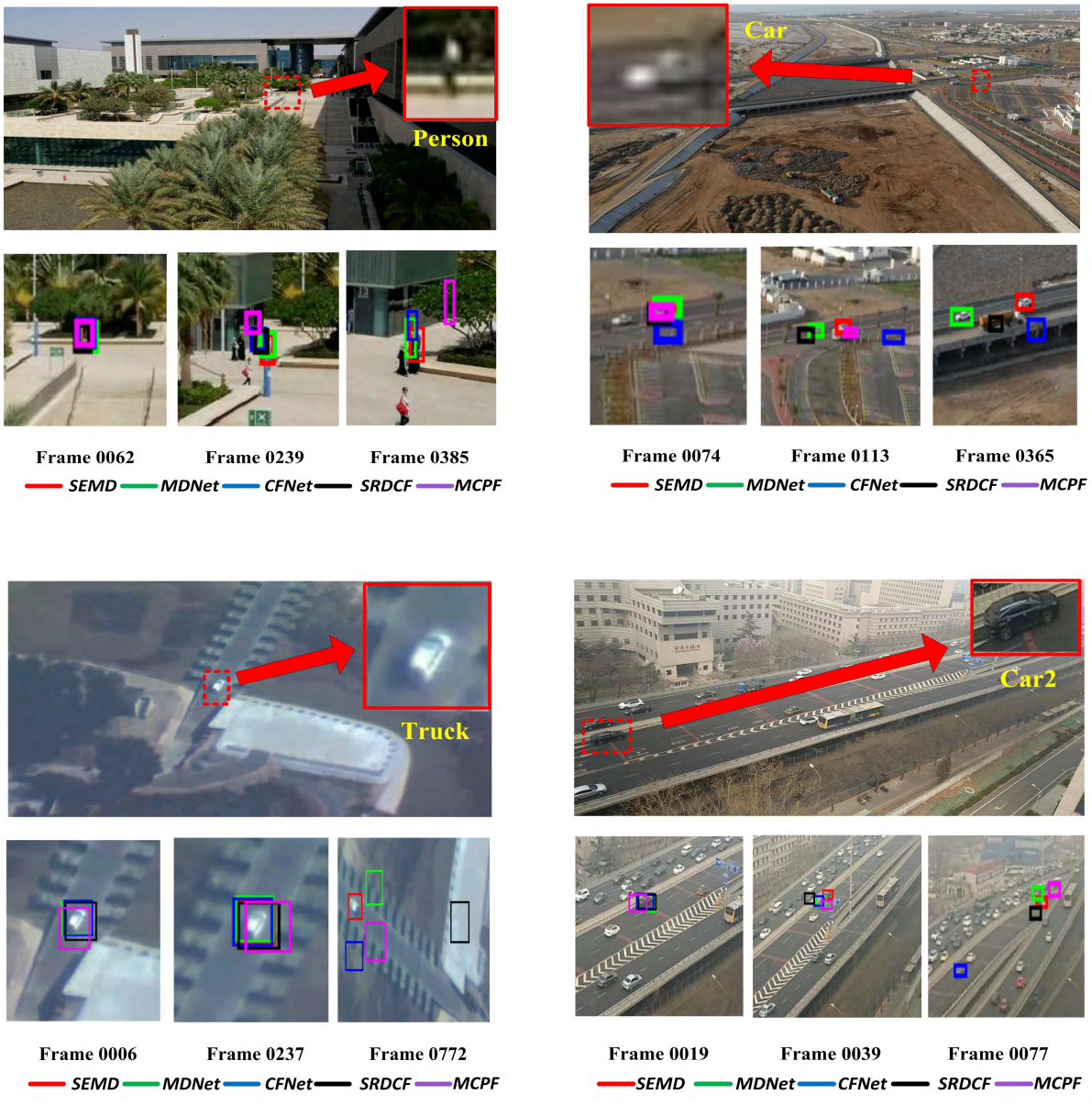

FIGURE 7. Qualitative results of the proposed method for challenging sequences (person, car, truck, car2).

are performed in regular intervals using the positive samples collected for a long period of time, and short-term updates are triggered to fine-tune the weights of the fully connected layers, whenever the score of the estimated target is below a threshold. However, UAV platforms often collect videos with short-term occlusion, rapid perspective changes and multiple similar disruptions, which can easily make the target scores below 0.5. Under this situation, there will be high computational costs for the network updating. Therefore, it is particularly important to make the network more discriminative. Benefitting from the typical easily confused negative samples augmentation and the saliency module, the network can better identify target and background. It means that the score of estimated target will be relatively high, which can avoid high computational costs producing by short-term updates. Moreover, thanks to the introduction of local weight allocation model, we can use fewer and more valuable negative samples to predict target. (2) compared with the MDNet, the extra structure of our algorithm is mainly the saliency module, which is acknowledged as a lightweight module. It has little impact on the network speed compared with the computational costs generated in other tracking stages.

Moreover, we also illustrated the qualitative results of multiple algorithms on a subset of sequences in Fig. 7. Our method shows consistently better performance for different challenging scenes, including orientation change, few pixels

FIGURE 8. Failure cases for the proposed method. Red and green bounding boxes denote the ground-truth and our result, respectively.

in target region and scale change. However, in over-complex large size scenes, the tracking task occasionally fails with a small target with many similar disturbances around, as shown in Fig. 8. We believe that this failure occurred because our three mechanisms cannot play a vital role in a situation in which the saliency of the target features is very insufficient, and other algorithms cannot solve this problem well either.

V. CONCLUSIONS

This paper presents a saliency enhanced multi-domain convolutional neural network for achieving stable and efficient target tracking in an UAV aerial video. First, we employed a typical easily confused negative samples augmentation strategy by combining LSGANs with a multi-orientation Gaussian Pyramid to make the network more discriminative. Then, a saliency module was embedded to improve network representation power. Finally, we constructed a local weight allocation model for improving timeliness. The experiments demonstrated that the proposed method can achieve high accuracy and efficiency results compared to state-of-the-art algorithms. The performance for a homemade dataset with challenging scenes from the perspective of UAV especially showed that our algorithm was more advanced than other algorithms that focus on solving these problems. However, as shown above, our method is inefficient in situations where the small target with many similar disturbances around. Therefore, we will consider introducing dynamic trajectory prediction for future research.

REFERENCES

[1] K. B. Logoglu, H. Lezki, M. K. Yucel, A. Ozturk, A. Kucukkomurler, B. Karagoz, A. Erdem, and E. Erdem, "Feature-based efficient moving object detection for low-altitude aerial platforms," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Venice, Italy, Oct. 2017, pp. 2119–2128.

[2] C. Fu, R. Duan, and D. Kircali, "Onboard robust visual tracking for UAVs using a reliable global-local object model," *Sensors.*, vol. 16, no. 9, p. 1406, Jul. 2016.

[3] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M.-M. Cheng, and S. L. Hicks, "Struck: Structured output tracking with kernels," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Barcelona, Spain, Nov. 2011, pp. 263–270.

[4] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.

[5] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Francisco, CA, USA, Jun. 2010, pp. 2244–2250.

[6] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[7] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2014, pp. 1–5.

[8] L. J. Wang, W. L. Ouyang, X. G. Wang, and H. C. Lu, "Visual tracking with fully convolutional networks," in *Proc. IEEE 15th Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3119–3127.

[9] K. Zhang, Q. Liu, Y. Wu, and M.-H. Yang, "Robust visual tracking via convolutional networks without training," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1779–1792, Apr. 2016.

[10] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. CVPR*, Jun. 2016, pp. 4293–4302.

[11] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2813–2821.

[12] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[13] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. ECCV*, Oct. 2016, pp. 445–461.

[14] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. CVPR*, Jul. 2017, pp. 6638–6646.

[15] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. ICCV*, Dec. 2015, pp. 4310–4318.

[16] T. Zhang, C. Xu, and M.-H. Yang, "Multi-task correlation particle filter for robust object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4335–4343.

[17] H. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proc. ICCV*, Oct. 2017, pp. 1135–1143.

[18] H. Fan and H. Ling, "Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking," in *Proc. ICCV*, Oct. 2017, pp. 5486–5494.

[19] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2805–2813.

[20] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.

[21] J. Wan, T. Hayat, and F. E. Alsaadi, "Adaptive neural globally asymptotic tracking control for a class of uncertain nonlinear systems," *IEEE Access*, vol. 7, pp. 19054–19062, 2019.

[22] I. Jung, J. Son, M. Baek, and B. Han, "Real-time MDNet," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 89–104.

[23] J. Ma, W. Sun, G. Yang, and D. Zhang, "Hydrological analysis using satellite remote sensing big data and CREST model," *IEEE Access*, vol. 6, pp. 9006–9016, 2018.
[24] Q. Wang, Z. Teng, J. Xing, J. Gao, W. Hu, and S. Maybank, "Learning attentions: Residual attentional siamese network for high performance online visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4854–4863.

FUKUN BI was born in Yunnan, China, in 1982. He received the B.S. and M.S. degrees in electrical engineering from Qingdao University and Yunnan University, in 2004 and 2008 respectively, and the Ph.D. degree in target detection and recognition from the Beijing Institute of Technology, in 2011.

From 2011 to 2014, he was engaged in post-doctoral research at Peking University. Since 2014, he has been an Associate Professor with the North China University of Technology. He wrote more than 30 articles and more than 10 inventions. He applied for 12 patents (three of them for national defense). His research interests include target detection and recognition, moving target tracking, and remote sensing image processing. He is an IET member, a Director of the China High-tech Industrialization Research Association, and an External Expert of Leiko Defense Strategic Development Committee.

Mr. Bi honors the National Defense Technology Invention and the National Science and Technology Invention Award.

MINGYANG LEI received the B.S. degree from the North China University of Technology, Beijing, China, in 2017, where he is currently pursuing the M.S. degree with the School of Electronic Information Engineering.

His current research interests include digital image processing and deep learning in remote sensing image. He holds two patents.

Mr. Lei received the award in the Zhongguancun Civil-Military Integration Competition. He was invited to give an oral presentation at the IET conference venue.

YANPING WANG was born in Shandong, China, in 1976. He received the B.S. and M.S. degrees in mechanical and electrical engineering from the Beijing Institute of Technology, in 2001, and the Ph.D. degree in signal processing from the Institute of Electronics, Chinese Academy of Sciences, in 2004.

From 2003 to 2014, he was a Researcher with the Institute of Electronics, Chinese Academy of Sciences. From 2015 to 2017, he was a Researcher with the China Academy of Safety Science and Technology. Since 2017, he has been a Professor with the North China University of Technology. He authored more than 70 articles and more than 30 inventions. His research interests include ground-based radar imaging, remote sensing image intelligent information processing, and synthetic aperture radar. He is a national expert in production safety, a member of signal processing branch of the China electronics society and the member of the Editorial Board of Signal Processing in china.

Dr. Wang awards and honors the China Excellent Patent Award and the first prize of Science and Technology Award of the China Occupational Safety and Health Association.

DAN HUANG was born in BaoTou, China, in 1983. She graduated from the Experimental Class of the Beijing Institute of Technology, and received the Ph.D. degree in target detection and recognition, in 2011.

Since 2011, she has been with the China Research and Development Academy of Machinery Equipment, and became an Associate Research Fellow, since 2013. Since 2014, she has been an Adjunct Master Tutor at Henan Polytechnic University. Since 2011, she has been a reviewer of IET Radar, Sonar and Navigation. She was mainly engaged in the basic theory and engineering technology research of target detection and identification. She had published more than 30 academic papers in SCI/EI journals and conferences, applied for more than 20 national invention patents, and applied for more than 20 software copyrights.

Dr. Huang was a recipient of the second prize of the Science and Technology Progress Award of the weapon group company.