ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20211115041517
Original scientific paper

SiamLST: Learning Spatial and Channel-wise Transform for Visual Tracking

Jun WANG, Limin ZHANG, Yuanyun WANG*, Changwang LAI, Wenhui YANG, Chengzhi DENG

Abstract: Siamese network based trackers regard visual tracking as a similarity matching task between the target template and search region patches, and achieve a good balance between accuracy and speed in recent years. However, existing trackers do not effectively exploit the spatial and inter-channel cues, which lead to the redundancy of pre-trained model parameters. In this paper, we design a novel visual tracker based on a Learnable Spatial and Channel-wise Transform in Siamese network (SiamLST). The SiamLST tracker includes a powerful feature extraction backbone and an efficient cross-correlation method. The proposed algorithm takes full advantages of CNN and the learnable sparse transform module to represent the template and search patches, which effectively exploit the spatial and channel-wise correlations to deal with complicated scenarios, such as motion blur, in-plane rotation and partial occlusion. Experimental results conducted on multiple tracking benchmarks including OTB2015, VOT2016, GOT-10k and VOT2018 demonstrate that the proposed SiamLST has excellent tracking performances.

Keywords: deep learning; siamese network; sparse transform; visual tracking

1 INTRODUCTION

Visual tracking aims to predict the trajectory and scale variations of a target in subsequent frames with the given target state in the initial frame. As an important branch of computer vision, visual tracking has a variety of applications, such as intelligent transportation system, augmented reality and human-computer interaction, to name a few. Although the performance of visual tracking has been greatly improved recently, well-balanced visual tracking still remains enormous challenge due to complicated scenarios, such as low resolution, partial occlusion, illumination variation, motion blur, in-plane and out-of-plane rotations, and so on.

In recent years, based on the techniques of deep learning, visual trackers obtain well-balanced tracking performances between accuracy and real-time speed. The typical tracking methods based on deep learning include two core components: feature extraction backbone based on Convolutional Neural Network (CNN) and similarity computing based on cross-correlation. These trackers have a powerful depth feature extraction ability to promote the tracking performance when the targets suffer from serious appearance variations. As a pioneering work [1], Siamese network is used for visual tracking. Many state-of-the-art algorithms based on Siamese network are proposed, such as SiamBAN [2].

Siamese network based visual algorithms regard the target tracking task as a similarity learning problem between the target template and search patches, which achieve real-time tracking performances. At first, the convolutional neural network adopts offline training manner to learn a similarity function on a large number of video sequences. Next, the similarity scores between the target template and search patches are computed. Lastly, the target position and scale offset are evaluated based on the score map in the next frame.

Despite the great success, Siamese network based visual trackers still have some disadvantages as follows: 1) in target feature extracting, the shallow convolutional layer of CNN tends to cause the lack of generalization ability, such as VGG16 [3]. Visual tracker based on a shallow feature extraction backbone is easy to shift when there is serious noises or local damages in an input image. 2) In deep CNN feature extracting, the offline training is often time-consuming. Meantime, the tracking performance will decline when the network depth is very deep, such as ResNet152 [4].

Recently, attention mechanism becomes a hot topic for the powerful ability in highlighting region of interest (ROI), such as FcaNet [5]. Attention has been widely applied in the fields of anomaly detection [6], semantic analysis [7] and face recognition [8]. Also, it is introduced to visual tracking. Wang et al. [9] design a deep architecture consisting of residual attention, general attention and channel attention for visual tracking. Zhang et al. [10] design an end-to-end framework to exploit the contextual information in consecutive frames, and the proposed tracking algorithm achieve robust tracking performance. However, the above works only pay more attention to the ROI, and the computing resources cost is expensive.

In this paper, in order to address the above-mentioned problem, we design a learnable feature extraction backbone that effectively exploits spatial and channel-wise correlations. Our algorithm achieves well-balanced tracking performance between tracking accuracy and average overlap. In addition, the SiamLST has surpassed some SOTA algorithms in some complicated situations, such as out-of-view, partial occlusion, and scale variation. The main contributions include three folds as follows:

- We propose a novel end-to-end deep model to extract more discriminative features. Comparison with other competing algorithms, our tracker effectively exploits dependency of channel-wise features and decreases the number of parameters by combining the advantages of CNN and the learnable sparse module.
- We design a visual tracking algorithm based on Siamese network. It consists of the proposed deep model and an efficient cross-correlation method. Our algorithm takes full advantages of spatial and channel-wise correlations, which greatly alleviate the influences of appearance variations, such as occlusion and motion blur.
- Extensive experiments demonstrate that the SiamLST algorithm outperforms SOTA works while running at real-time speed on OTB2015, GOT-10k, VOT2016 and VOT2018 benchmarks.

For the rest of this paper, we will describe this work according to the following arrangement. In Section II, we will review relevant tracking techniques and algorithms. The details of the designed SiamLST are described in Section III. The extensive experiments conducted on multiple benchmarks are presented in Section IV. In Section V, we will draw a conclusion.

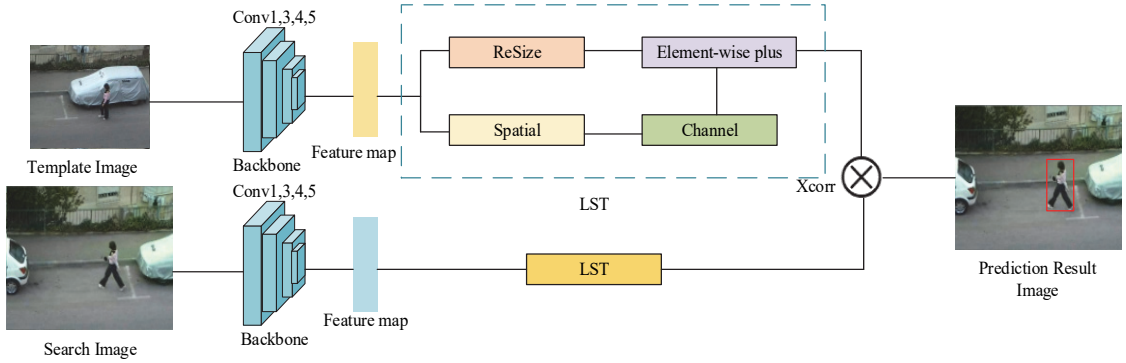

Figure 1 The SiamLST workflow.

2 LITERATURE REVIEW

In this section, we will review some relevant tracking techniques and algorithms. In particular, the end-to-end visual trackers based on Siamese network and attention mechanism are reviewed.

2.1 Visual Trackers Based on Correlation Filter

According to appearance models, visual tracking algorithms are roughly divided into generative and discriminative algorithms. Among them, the typical generative algorithms include mean shift [11] and sparse representation [12], while the representative works of discriminative algorithms are correlation filter based trackers [16] and trackers based on deep learning [17].

In the past decades, visual trackers based on correlation filter have received extensive attention because of the simple structure and expansibility. In [13], the correlation filter is applied to visual tracking task for the first time, and the speed reaches 669 frames per second. On the basis of MOSSE, Henriques et al. [14] add a regularization term to avoid overfitting, and introduce circulant matrix and kernel function to improve the tracking speed. At the same time, Henriques et al. further propose kernelized correlation filters (KCF) based tracker [15].

2.2 Siamese Based Tracking Algorithms

Recently, visual trackers based on Siamese network have received considerable attention due to the achieved good balance between accuracy and real-time speed. Siamese network based tracking algorithms usually learn a similarity matching function by off-line training from a large number of labelled sequences, which improve the speed and accuracy of online tracking.

In [1], a fully convolutional neural network for computing the similarity between the template and a search region is designed. Guo et al. [21] propose a dynamic Siamese network based tracker that can online learn the target appearance variation and suppress noise from the previous frame. In addition, this tracker uses continuous video sequences instead of image pairs for training. In [18], the category semantic information branch is added to the tracking framework, and this method achieves robust tracking performance between accuracy and overlap rate.

Li et al. [19] propose a novel Siamese network framework consisting of Siamese network and region proposal network (RPN). RPN obtains a similarity score map by a classification branch and a regression branch. SiamRPN++ [22] aims to improve tracking accuracy, and develops different frameworks to obtain superior tracking performances. Yu et al. [23] develop a novel deformable Siamese attention network consisting of a self-attention to exploit richness of semantic information and a cross-attention to enhance spatial and channel-wise correlations.

2.3 Attention Mechanism

Attention mechanism makes the convolutional neural network focusing on the ROI to better highlight it. Attention mechanism is widely used in deep learning, such as object detection, face recognition, and semantic segmentation. FcaNet [5] uses GAP from a frequency domain perspective to compensate for the lack of feature information in existing channel attention methods. FcaNet extends GAP to a more general 2-dimensional discrete cosine transform (DCT) form, and introduces more frequency components to fully utilize the information. In [7], a residual attention network captures mixed attention to train very deep residual attention networks.

Recently, attention has been applied to visual tracking for enhancing the feature representation ability. Wang et al. [4] propose a novel deep architecture that consists of channel attention, residual attention and general attention to promote tracking performance. Yu et al. [23] design a deformable Siamese attention network to enhance target discriminative ability. Zhu et al. [24] develop an end-to-end algorithm for visual tracking, which takes full advantages of rich flow information. Although these methods increase the weight of the ROI, they result in the increase of the number of parameters. Inspired by the above-mentioned works, we propose a novel attention mechanism based Siamese network that effectively reduces the number of pre-trained model parameters and exploits inter-channels feature dependency.

3 RESEARCH METHOD

In this section, we will describe our SiamLST algorithm in details. Inspired by SiamFC, our algorithm obtains more richness and sparse representation features by designed deep model. In Fig. 1, the SiamLST includes two core components: a novel deep model for extracting the template and search region features, and an efficient similarity matching method. Among them, Xcorr is used to measure the similarity between the target template and the search region patches, which calculates the range of variation of the target position by the similarity score map in the current frame.

3.1 Siamese Network Backbone

In recent years, tracking algorithms based on deep learning have attracted great attention due to good balance between accuracy and real-time speed. Visual trackers based on Siamese architecture consist of a features extraction backbone and an efficient cross-correlation method. In [1], Siamese network is introduced to visual tracking, which regards visual tracking as a similarity learning task, and achieves competing tracking performance. Inspired by SiamFC, some algorithms achieve more robust tracking performance and real-time speed by introducing extra subnetworks or ensembling multiple subnetworks [5].

Most trackers based on Siamese network adopt shallow convolutional layers to extract target features, such as VGG16 [3]. At present, the deep convolutional neural network plays a leading role in visual tracking and has achieved excellent image classification performance, such as ResNet [4]. Although deep learning based visual trackers have made significant progress, the pre-trained model of Siamese network based tracking algorithms have a large number of parameters, and hardly take full advantages of spatial and channel-wise correlation. The convolution layer usually needs the vast convolution kernels to extract semantic information of the input image and distinguishes the foreground from its background. Unfortunately, this method is easy to cause the vast number of offline model parameters. Meantime, all pixels participating in the sliding window operation have the same weight coefficient. The ROI is not highlighted and the background information is not suppressed. This learning method has some disadvantages in locating the target in the next frame.

3.2 Learnable Feature Extraction Backbone

To solve the mentioned problem of feature extraction stage in visual tracking, we propose a novel Learnable Spatial and Channel-wise Transform model based on attention mechanism. Before describing our model in detail, let's review the operation of the convolution layer in traditional features extraction backbone. The output $O_{i,j,k}$ by convolving is computed as follows:

$$O_{i,j,k} = \sum_{x=1}^{s}\sum_{y=1}^{s}\sum_{z=1}^{c_{in}} \Omega\left(\mathcal{J}; i, j\right) \cdot \mathcal{K}_{x,y}^{(k)} \tag{1}$$

where $\left(\mathcal{J}; i, j\right)_{x,y,z}$ means a tensor, which is extracted from I at pixel position $\left(x, y, z\right)$, Ω is the sliding window in convolution layers, and $\mathcal{K}_{x,y}^{(k)}$ represents the pixel at $\left(x, y\right)$ of $\mathcal{K}^{(k)}$.

We observe that the traditional convolution layers have two disadvantages. At first, all the pixels in a slide window are needed to participate in calculation in the spatial position $\left(x, y\right)$. It is helpful for the extraction of high-frequency features, but causes redundancy for the extraction of low-frequency features that are the majority in the whole image. Next, all pixel in positions have the same weight for each channel, which is disadvantageous to distinguish the foreground from surround background. Therefore, we intend to discuss the proposed deep model from spatial and channel-wise, respectively.

The designed algorithm SiamLST processes the feature maps from spatial and channel-wise, respectively. Our model effectively exploits feature dependency to alleviate the local feature redundancy by transforming to sparser domain. In order to retain more details of the input image as much as possible, similar to residual network, we use the point-wise convolution operator to adapt the feature map scale. The general process is as follows:

$$T_{LST} \circ \mathcal{J} = T_s \circ T_c \circ T_r \circ \mathcal{J} \tag{2}$$

where T_s and T_c represent the spatial and channel-wise transform in the learning sparse transform module, respectively. T_r means a down-sampling operation, which ensures that the input image features can concatenate with the convolutional features. T_{LST} means *LST* module in Fig. 1 and $\mathcal{J}$ corresponds to the output depth feature of the backbone network.

The learnable spatial transform. The spatial feature redundancy is a serious problem in training lightweight networks. To alleviate this problem, we develop an efficient spatial transform module T_s by combining a learnable weight $\mathcal{W}_s$. The specific detail is to disrupt the input image to different frequency bands through continuous row and column transform. The learnable weight can be described as follows:

$$\mathcal{W}_s = \mathcal{W}_{\text{column}} \otimes \mathcal{W}_{\text{row}} \tag{3}$$

where $\otimes$ means the Kronecker product. The row and column transforms correspond to different learnable weights $\mathcal{W}_{\text{column}}$ and $\mathcal{W}_{\text{row}}$.

The learnable channel transform. We exploit the channel dimension dependency by T_c, which plays a key role in suppressing inter-channels redundancy. The maps of the channel-wise transform module input features to a sparser domain, and resizes scale by T_r. Similar to the above spatial transform, we reweight the feature map to highlight ROI. Moreover, we effectively use the point-wise convolutional layer to obtain inter-channel cues.

The learnable resize transform. In a traditional feature extraction backbone, each pixel on the sliding window has the same weight coefficient, which hardly highlights ROI and suppresses noisy inference. To alleviate this problem,

we develop a learnable weight mapping operator. The proposed model obtains sparser template features by spatial and channel-wise transforms, and exploits inter-channel cues to highlight ROI. Meantime, the tracking performance was improved by an efficient resize transform module, and enhanced the robustness of the designed Siam LST.

Activation Scheme. After the above steps, the designed model obtains sparser features that are helpful to deal with challenging in video sequences. In addition, our trained model becomes more lightweight by exploiting inter-channel cues. The nonlinear activation function can reduce the negative sample feature to zero and keep the positive sample feature unchanged. However, the existing nonlinear activation functions are not robust enough in some scenes when the input image is noisy or partially occluded. To further exploit nonlinear semantic information, based on ReLU activation function, a novel activation function is designed as follows:

$$y = \begin{cases} \operatorname{sgn}(x)(|x| - \tau), & |x| \geq \tau \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $\operatorname{sgn}(x)$ is the symbolic function and τ is the hyper parameter threshold. When x is greater than τ, y value is 1; x is less than τ, y value is −1, the rest of the case y value is 0.

Specifically, our model achieves more robust tracking performance when there are partial occlusions or motion blurred in the input images. Moreover, the developed activation function can compress the trivial features into sparse domain to make the trained model more lightweight.

4 RESULTS AND DISCUSSION

In this section, we will introduce environment configurations and the evaluation metric. Next, we compare our algorithm with state-of-the-art tracking algorithms in quantitative and qualitative ways. Finally, we analyse the limitations of the proposed algorithm and the future works.

4.1 Implementation Details

We develop an end-to-end learnable model as a feature extraction backbone by taking full advantages of the CNN learnable sparse transform, and we select various convolution layers to enhance the deep model representation capacity. Finally, we select the best model that is verified in the ablation study. In addition, the target template is cropped to 127 × 127 × 3 and the search region is cropped to 255 × 255 × 3. There are totally 50 epochs performed with stochastic gradient descent. The designed SiamLST is compared with some competing visual trackers including SiamFC [1], KCF [15], LDES [28], AFCN [29], LMCF [30], DSST [26], and so on.

Environment configuration. Our algorithm is implemented with a NVIDIA Quadro P4000 GPU and Intel Xecon E5-2600 v4 CPU (2.00 GHz), 32 GB RAM, and tested in Python 3.6 with Pytorch 1.4.0 during the offline training stage. We utilize 9335 video sequences in GOT-10 k test split to generalize model. In addition, we evaluate the best model on multiple benchmarks.

Evaluation metric. The one-pass evaluation (OPE) metric of precision and success rate are adopted to evaluate the tracking performance. The center location error (CLE) between the bounding boxes and the ground-truth is adopted to evaluate the performance of accuracy. Moreover, the intersection over union (IoU) between the predicted bounding boxes and the ground truth is adopted to evaluate the trackers performance of success rate. Finally, we analyse the tracking performance in quantitative and qualitative comparisons.

4.2 Ablation Study

As shown in Tab. 1, we select different AlexNet layers to design various deep models, and the experimental results demonstrate that the best method is the second model on OTB2015. At the features extraction stage, the best strategy should be initialized with some smaller convolutional kernels, and the receptive field will increase when the network is deepening. On the one hand, when a feature map has more detailed information in the shallow layers, the smaller convolutional kernel effectively captures the information. On the other hand, when a feature map has more rich semantic information in the deep layers; the larger receptive field takes full advantage of exploited semantic information. In investigating the impact of different convolutional layers, we obtain the best model, i.e., the best feature extraction backbone consists of AlexNet 1, 3, 4, 5 convolution layers and learnable spatial and channel-wise modules.

Table 1 Different feature extraction backbones by combining various convolution layers and learnable sparse transform model.

Dataset	Backbone	Embedding type	Suc.	Pre.
OTB2015	Conv2,3,4,5+LST	Xcorr	0.787	0.590
	Conv1,3,4,5+LST		0.806	0.589
	Conv1,2,4,5+LST		0.790	0.588
	Conv1,2,3,5+LST		0.755	0.568
	Conv1,2,3,4+LST		0.769	0.580

4.3 Comparison With State-of-the-art Trackers

OTB2015. OTB2015 is a typical tracking benchmark including 98 challenging sequences. Each sequence has one or more attributes such as fast motion (FM), deformation (DEF), out-of-view (OV), occlusion (OCC), low resolution (LR), out-of-plane rotation (OPR), in-plane rotation (IPR), background clutter (BC), motion blur (MB), illumination variation (IV) and scale variation (SV). In Fig. 2 and Tab. 2, our algorithm obtains the smallest center location error and the second overlap. On the one hand, the spatial and channel-wise transform is helpful to improve tracking performances. On the other hand, we calculate the similarity between the template and search patches by an efficient cross-correlation method.

GOT - 10 k. It is a large-scale challenging tracking benchmark. It contains more than 10,000 videos and has more than 1.5 million manually labelled bounding boxes, divided into 563 target categories and 87 movement patterns. In Tab. 3, we compare SiamLST against superior trackers on GOT - 10 k, and our tracker SiamLST achieves

the best results in SR0.5 and SR0.75. In addition, the result of our algorithm obtains the second result in AO.

Table 2 Comparisons between SiamLST and other SOTA trackers on OTB2015. We highlight the best three results in red, blue and green.

Trackers	Precision	Success	FPS
KCF [15]	69.6	47.4	223.8
DSST [26]	68.0	51.3	25.4
MEEM [27]	78.1	53.0	19.5
ACFN [29]	79.9	57.4	15
LCT [31]	76.2	56.2	27
LMCF [30]	78.9	58.0	85
CFNet [17]	74.8	56.8	75
SiamFC [1]	77.1	58.2	86
LDES [28]	78.5	61.5	20
SiamLST	80.6	58.9	71

VOT2016. VOT2016 expands the test sets to 60 groups, and realizes automatic sample labelling. In Fig. 3, our algorithm SiamLST achieves the best EAO result compared with other SOTA trackers. Due to the learnable sparse transform model being added to CNN, we achieve more robust and accurate performance when input image appears with corruption or noisy.

Table 3 Comparison with SOTA algorithms on GOT-10k.

Tracker	AO	SR0.5	SR0.75
SiamFC [1]	34.8	35.3	9.8
CFNet [17]	29.3	26.5	8.7
MDNet [25]	29.9	30.3	9.9
BACF [16]	26.0	26.2	10.1
SiamLST	33.6	36.8	10.2

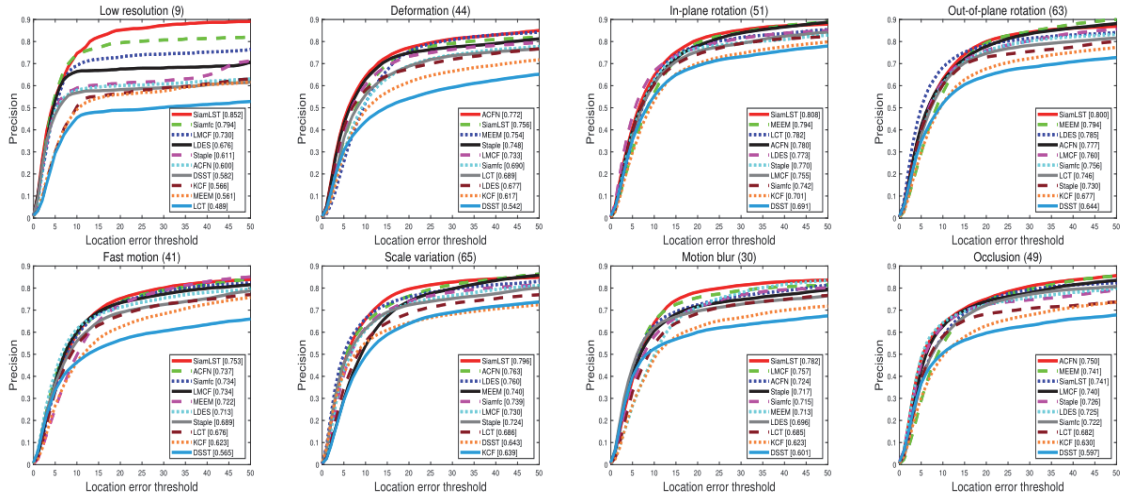

Figure 3 Precison plot of OPE.

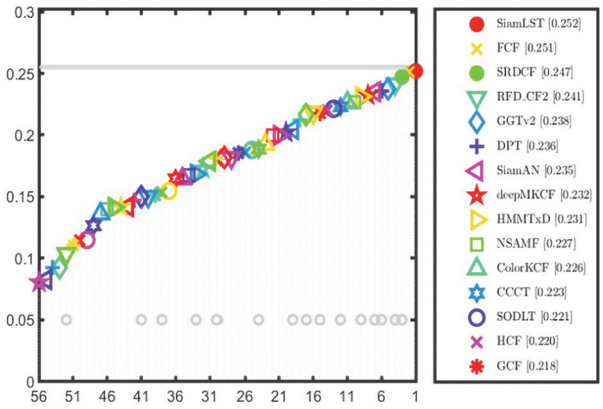

Figure 4 Expected averaged overlap performance on VOT2016 benchmark.

Table 4 Comparison with competitive trackers on VOT2018 benchmark. The best three results are highlighted in red, blue and green colors.

Trackers	A	R	EAO
DSST [26]	39.5	145.2	7.9
KCF [15]	44.7	77.3	13.5
DCFNet [32]	47.0	54.3	18.2
SiamFC [1]	50.3	58.5	18.8
DSiam [21]	51.2	64.6	19.6
CSTEM [35]	46.7	41.2	22.6
MEEM [27]	46.3	53.4	19.2
STST [21]	46.4	62.1	18.7
UpdateNet [34]	51.8	45.4	24.4
SiamLST	52.9	49.1	21.7

VOT2018. VOT2018 has added a long-term tracking competition. In comparison with short-term tracking algorithms, VOT2018 increases two challenging factors: full occlusion and out-of-view. In this case, the target completely disappears in the video frames, and the tracker needs to judge whether the target disappears and re-detects it when it appears. In Tab. 4, the designed SiamLST achieves the SOTA result in accuracy. We analyse that the learnable sparse transform is a benefit in locating the target center position.

4.3.1 Quantitative Evaluation

In this section, we compare the SiamLST with seven outstanding algorithms on OTB2015. In Tab. 4, our algorithm achieves SOTA precision and success rates according to the one-pass evaluation. The proposed SiamLST is superior to other competing algorithms in accuracy, and achieves the second result in success rate.

These results demonstrate the effectiveness of the proposed learnable feature extraction backbone. It is worth noting that SiamLST improves the precision by 0.035 compared with SiamFC, and only increases little parameters. In addition, the advantage of SiamLST is that it effectively alleviates feature redundancy. Fig. 4 and Fig. 5 show the center location error and overlap rate of each attribute. We select eight challenging attributes to evaluate the tracking performance on OTB2015.

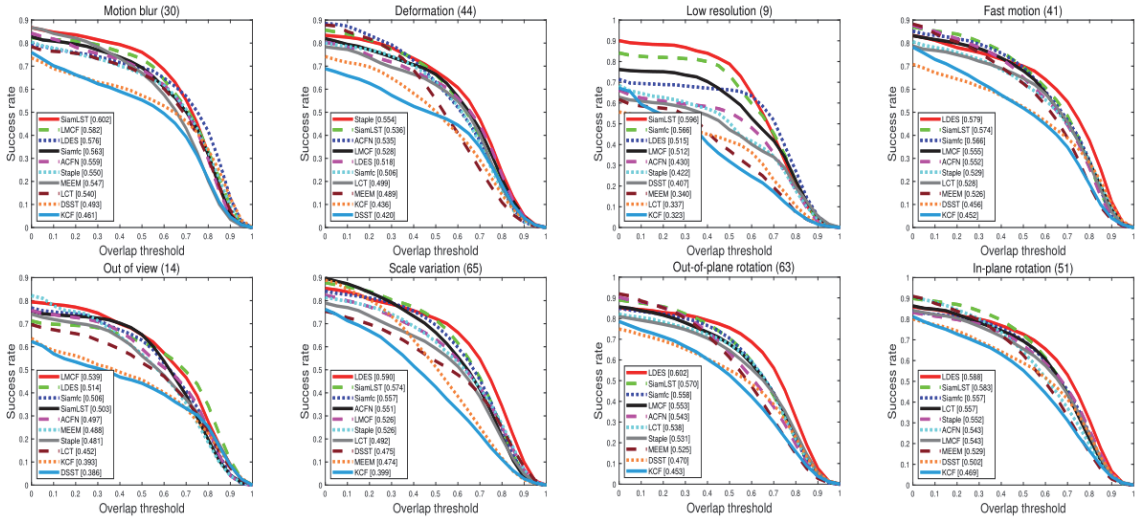

Figure 5 Overlap success (OS) rates of OPE.

4.3.2 Qualitative Evaluation

In Fig. 6, we show some tracking results compared with some outstanding tracking algorithms. Next, we will analyse the advantages and disadvantages of the designed SiamLST with specific sequences.

Illumination variation. In Fig. 6, the targets in the sequences Car1, Human2 and Skiing suffer illumination variations, and our algorithm achieves the best tracking performance compared with KCF, LDES, SiamFC. In the Car1 sequence, there are similar targets and low resolution interference in the scene. We can see that our algorithm can effectively locate the target and other trackers cannot adapt to the scale variation well. In the

Skiing sequence, the target suffers scale variation, in-plane and out-of-plane rotations. We can see that only our algorithm and LDES tracker can precisely track the target due to the SiamLST extracting depth features with semantic information. In Human2 sequence, the target undergoes motion blur and illumination changes. Due to sparser and richer template features, the proposed algorithm can track the target when input image appears interrupted or noisy. Finally, due to lacking online template updating module, the proposed tracking algorithm drifts away when suffering the target long-term occlusion.

Figure 6 Qualitative evaluation on OTB2015.

Occlusion. As shown in Fig. 6, we can see that some scenes of Basketball, Biker, Football, Human7 and Tiger2 sequences suffer partial or short-time occlusion. Occlusion is a typical challenge factor in video sequences. In Basketball video sequence, our SiamLST algorithm can accurately track the target and adapt the scale variation

when there are many similar objects and partial occlusion in the scene. In Bikervideo sequence, due to the influence of fast motion and in-plane rotation, KCF and LDES fail to track the targets. SiamLST can distinguish the foreground from the surround background well. The same situation occurs in Football, Human7 and Tiger2 video sequences, when the target suffers partial or short-time occlusion in the scene, our SiamLST algorithm performs better than the other three SOTA algorithms. The source of our SiamLST tracker obtains outstanding tracking performances. At first, the SiamLST effectively exploits inter-channel cues to represent the target and search patches, which largely reduce pre-trained model parameters. Meantime, the learnable feature extraction backbone obtains richer and sparser features. Finally, our tracker SiamLST achieves leading performances when the target suffers occlusions.

4.4 Limitations and Future Works

4.4.1 Limitations

In Fig. 7, we show some cases of tracking failure. The tracker SiamLST drifts under the condition of long-term tracking and serious occlusions. In Bird1 sequence, when the target completely disappears for a period of time, the tracker SiamLST cannot track the target because our tracker has no re-detection scheme. In the 378th frame, when the background specks in the scene, our tracker loses the target again because our tracker has no online update module.

Figure 7 Two failure cases. (a) Bird1. (b) CarDark. OurSiamLST and ground truth results are shown in red and green boxes.

4.4.2 Future Works

Our algorithm is inspired by SiamFC and takes full advantages of CNN and LST to represent a target, and achieves competing tracking performances. Next, we will combine LST in various deep feature extraction backbones, such as ResNet152. Moreover, we will consider some online template update schemes to adapt appearance variations and enhance the tracking accuracy.

5 CONCLUSION

In this paper, we design a simple and efficient visual tracker based on Siamese network, which effectively reduces the number of pre-trained model parameters and exploits the dependency of inter-channels features. In the designed model, the feature maps become sparser and richer by taking the advantages of CNN and the learnable sparse transform module. The model exploits spatial and channel-wise correlation, which tends to handle challenging scenarios, such as scale variation, in-plane and out-of-plane rotations and partial occlusion. Extensive results demonstrate that the SiamLST algorithm obtains leading performances in multiple benchmarks including OTB2015, GOT-10k, VOT2016 and VOT2018.

Acknowledgements

This work is supported by the Jiangxi Science and Technology Research Project of Education Department (No: GJJ190955), the National Natural Science Foundation of China (No: 61861032 and 61865012), and the Innovative Projects of NIT (No: YC2021-S815).

6 REFERENCES

[1] Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., & Torr, P. H. (2016). Fully-convolutional siamese networks for object tracking. *European conference on computer vision*, 850-865. https://doi.org/10.1007/978-3-319-48881-3_56

[2] Chen, Z., Zhong, B., Li, G., Zhang, S., & Ji, R. (2020). Siamese box adaptive network for visual tracking. *IEEE Conference on Computer Vision and Pattern Recognition*, 6668-6677. https://doi.org/10.1109/CVPR42600.2020.00670

[3] Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprintarXiv:1409.1556.

[4] Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., & Tang, X. (2017). Residual attention network for image classification. *IEEE conference on computer vision and pattern recognition*, 3156-3164. https://doi.org/10.1109/CVPR.2017.683

[5] Qin, Z., Zhang, P., Wu, F., & Li, X. (2021). Fcanet: Frequency channel attention networks. *IEEE International Conference on Computer Vision*, 783-792. https://doi.org/10.1109/ICCV48922.2021.0008

[6] Jing, L., Peng, Y., & Zhang, D. (2021). Anomaly detection based on multiple streams clustering for train real-time ethernet. *Tehnički vjesnik*, 1353-1361. https://doi.org/10.17559/TV-20210414030120

[7] Song, Y. (2021). Construction of event knowledge graph based on semantic analysis. *Tehnički vjesnik*, 1640-1646. https://doi.org/10.17559/TV-20210427063132

[8] Chunquan, D., Quanlei, W., Kun, J., Tao, Z. (2021). Ultimate support force of excavation face in curved shield tunnels in composite strata. *Tehnički vjesnik*, 708-717. https://doi.org/10.17559/TV-20180720050519

[9] 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. *IEEE conference on computer vision and pattern recognition*, 4854-4863. https://doi.org/10.1109/CVPR.2018.00510

[10] Zhang, Y., Wang, L., Qi, J., Wang, D., Feng, M., & Lu. H. (2018). Structured siamese network for real-time visual tracking. *European conference on computer vision*, 351-366. https://doi.org/10.1007/978-3-030-01240-3_22

[11] Collins, R. T. (2003). Mean-shift blob tracking through scale space. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, II-234.

[12] Zhong, W., Lu, H., & Yang, M.-H. (2014). Robust object tracking viasparse collaborative appearance model. *IEEE Transactions on Image Processing*, 2356-2368. https://doi.org/10.1109/TIP.2014.2313227

[13] Bolme, D. S., Beveridge, J. R., Draper, B. A., & Lui, Y. M. (2010). Visual object tracking using adaptive correlation filters. *IEEE computer society conference on computer vision and pattern recognition*, 2544-2550. https://doi.org/10.1109/CVPR.2010.5539960

[14] Henriques, J. F., Caseiro, R., Martins, P., & Batista, J. (2012). Exploiting the circulant structure of tracking-by-detection with kernels. *European conference on computer vision*, 702-715. https://doi.org/10.1007/978-3-642-33765-9_50

[15] Henriques, J. F., Caseiro, R., & Batista, J. (2014). High-speedtracking with kernelized correlation filters. *IEEE transactionson pattern analysis and machine intelligence*, 583-596. https://doi.org/10.1109/TPAMI.2014.2345390

[16] Kiani, G. H., Fagg, A., & Lucey, S. (2017). Learning background-aware correlation filters for visual tracking. *IEEE international conference on computer vision*., 1135-1143. https://doi.org/10.1109/ICCV.2017.129

[17] Valmadre, J., Bertinetto, L., Henriques, J., et al. (2017). End-to-end representation learning for correlation filter based tracking. *IEEE conference on computer vision and pattern recognition*. 2805-2813.

[18] He, A., Luo, C., Tian, X., & Zeng, W. (2018). A twofold Siamese network for real-time object tracking. *IEEE Conference on Computer Vision and Pattern Recognition*, 4834-4843. https://doi.org/10.1109/CVPR.2018.00508

[19] Li, B., Yan, J., Wu, W., Zhu, Z., & Hu, X. (2018). High performance visual tracking with siamese region proposal network. *IEEE Conference on Computer Vision and Pattern Recognition*, 8971-8980. https://doi.org/10.1109/CVPR.2018.00935

[20] Wang, Q., Zhang, L., Bertinetto, L., Hu, W., & Torr, P. H. (2019). Fastonline object tracking and segmentation: A unifying approach. *IEEE Conference on Computer Vision and Pattern Recognition*, 1328-1338. https://doi.org/10.1109/CVPR.2019.00142

[21] Guo, Q., Feng, W., Zhou, C., Huang, R., Wan, L., & Wang, S. (2017). Learning dynamic siamese network for visual object tracking. *IEEE International Conference on Computer Vision*, 1781-1789. https://doi.org/10.1109/ICCV.2017.196

[22] Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., & Yan, J. (2019). Siamrpn++: Evolution of siamese visual tracking with very deep networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 4282-4291. https://doi.org/10.1109/CVPR.2019.00441

[23] Yu, Y., Xiong, Y., Huang, W., & Scott, M. R. (2020). Deformable siamese attention networks for visual object tracking. *IEEE Conference on Computer Vision and Pattern Recognition*, 6728-6737. https://doi.org/10.1109/CVPR42600.2020.00676

[24] Zhu, Z., Wu, W., Zou, W., & Yan, J. (2018). End-to-end flow correlation tracking with spatial-temporal attention. *IEEE conference on computer vision and pattern recognition*, 548-557. https://doi.org/10.1109/CVPR.2018.00064

[25] Nam, H. & Han, B. (2016). Learning multi-domain convolutional neural networks for visual tracking. *IEEE conference on computer vision and pattern recognition*., 4293-4302. https://doi.org/10.1109/CVPR.2016.465

[26] Danelljan, M., Häger, G., Khan, F. S., & Michael, F. (2016). Discriminative scale space tracking. *IEEE transactions on pattern analysis and machine intelligence*, *39*(8), 1561-1575. https://doi.org/10.1109/TPAMI.2016.2609928

[27] Zhang, J., Ma, S., & Sclaroff, S. (2014). MEEM: robust tracking via multiple experts using entropy minimization. *European conference on computer vision*, 188-203. https://doi.org/10.1007/978-3-319-10599-4_13

[28] Li, Y., Zhu, J., Hoi, S. C. H., Wenjie, S., Zhefeng, W., & Hantang, L. (2019). Robust estimation of similarity transformation for visual object tracking. *AAAI conference on artificial intelligence*., *33*(01), 8666-8673. https://doi.org/10.1609/aaai.v33i01.33018666

[29] Jongwon, C., Hyung, J. C., Sangdoo, Y., Tobias, F., Yiannis, D., & Jin, Y. C. (2017). Attentional correlation filter network for adaptive visual tracking. *IEEE conference on computer vision and pattern recognition*., 4807-4816. https://doi.org/10.1109/CVPR.2017.513

[30] Wang, M., Liu, Y., & Huang, Z. Large margin object tracking with circulant feature maps. *IEEE conference on computer vision and pattern recognition*., 4021-4029. https://doi.org/10.1109/CVPR.2017.510

[31] Ma, C., Yang, X., Chongyang, Z., & Yang, M. (2015). Long-termcorrelation tracking. *IEEE Conference on ComputerVision and Pattern Recognition*, 5388-5396. https://doi.org/10.1109/CVPR.2015.7299177

[32] Wang, Q., Gao, J., Xing, J., et al. (2017). Dcfnet: Discriminant correlation filters network for visual tracking. *arXiv preprint arXiv*:1704.04057.

[33] Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Cehovin L., Zajc, T., Vojir, G., Bhat, A., & Lukezic, A. (2018). Eldesokeyet al..The sixth visual object tracking vot2018 challenge results. *European Conference on ComputerVision Workshops*, 3-53.

[34] Zhang, L., Gonzalez-Garcia, A., Weijer, J., Martin, D., & Fahad, S. K. (2019). Learning the model update for siamese trackers. *IEEE International Conference on Computer Vision.*, 4010-4019. https://doi.org/10.1109/ICCV.2019.00411

[35] Zhang, Z. & Wong, T. T. Effective occlusion handling for fast correlation filter-based trackers. *arXiv preprint arXiv*:1807.04880, 2018. https://doi.org/10.22161/eec.562

Contact information:

Jun WANG
School of Information Engineering,
Nanchang Institute of Technology, Nanchang, China

Limin ZHANG
School of Information Engineering,
Nanchang Institute of Technology, Nanchang, China

Yuanyun WANG
(Corresponding author)
School of Information Engineering,
Nanchang Institute of Technology, Nanchang, China
E-mail: Wangyy_abc@163.com

Changwang LAI
School of Information Engineering,
Nanchang Institute of Technology, Nanchang, China

Wenhui YANG
School of Information Engineering,
Nanchang Institute of Technology, Nanchang, China

Chengzhi DENG
School of Information Engineering,
Nanchang Institute of Technology, Nanchang, China