Dual Siamese Channel Attention Networks for Visual Object Tracking

Wenxing Gao, Xiaolin Tian, Yifan Zhang, Nan Jia, Ting Yang, Licheng Jiao

Dual Siamese Channel Attention Networks for Visual Object Tracking*

Wenxing Gao, Xiaolin Tian, Yifan Zhang, Nan Jia, Ting Yang, and Licheng Jiao

School of Artificial Intelligence Xidian University, Xi'an, 710071, China
{gaowxing,z1015367971,jiananxidian}@163.com
20171213676@stu.xidian.edu.cn
{xltian,lchjiao}@mail.xidian.edu.cn

Abstract. Siamese network based trackers have achieved remarkable performance on visual object tracking. The target position is determined by the similarity map produced via cross-correlation over features generated from template branch and search branch. The interaction between the template and search branchs is essential for achieving high-performance object tracking task, which is neglected in previous works as features of the two branchs are computed separately. In this paper, we propose Dual Siamese Channel Attentions Networks, referred as SiamDCA, which exploits the channel attentions to further improve tracking robustness. Firstly, a convolutional version of Squeeze and Excitation Networks (CSENet) is embedded in backbone to explicitly formulate interdependencies between channels to recalibrate channel-wise feature responses adaptively. Meanwhile, we propose a novel Global Channel Enhancement (GCE) module, which is capable of capturing attention weights of each channel in template branch, so as to normalize the channel characteristics in search branch. We experiment on benchmark OTB2015, VOT2016 and UAV123 where our algorithm demonstrates competitive performance versus other state-of-the-art trackers.

Keywords: Siamese network · Visual object tracking · Channel Attentions.

1 Introduction

Visual object tracking is a fundamental but challenging task in computer vision, which has a wide range of applications, such as visual surveillance [22], human-computer interactions [15], automatic driving, robot sensing, etc. Given an arbitrary target location in the initial frame, the tracker needs to infer the location of the target in each subsequent frame. Although visual object tracking has received extensive attention over the last decades, it still faces challenges due to numerous factors such as occlusion, scale variation, background clutters, fast motion, appearance variations.

The Siamese based trackers [1, 24] are trained completely offline by using massive frame pairs collected from videos. Both the features of template branch and search branch are extracted through the backbone independently. In addition, as observed in [11], each channel map of the high-level convolutional features

* Supported by the National Natural Science Foundation of China (No. 61977052).

usually responses for a specific object class. There is no information interaction between the template branch and the search branch, which limits the potential performance of Siamese architecture.SiamAttn [23] introduce a new Siamese attention mechanism which computes deformable self-attention and cross-attention jointly to improve discriminability, however, it imposes a heavy computational burden.

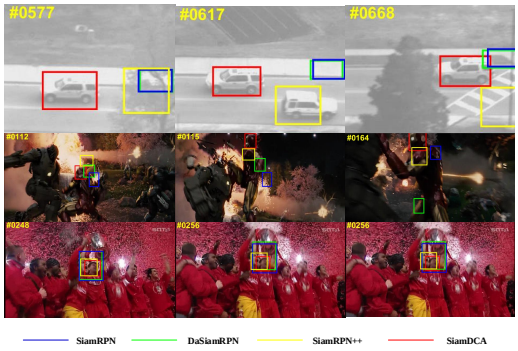


Fig. 1. Tracking results of our SiamDCA with three state-of-the-art trackers.Our results are more accurate,and are robust to appearance changes, complex background and close distractors with occlusions.

In this work, we first introduce the convolutional version of Squeeze and Excitation Networks (CSENet) [9] in the backbone to adaptively recalibrate channel-wise feature responses by explicitly modeling interdependencies between channels in template branch and search branch. Besides, we propose a Global Channel Enhancement (GCE) Module, which can capture rich global context information in each channel of template features, so as to enhance the obtained information and standardize the search feature of corresponding channel in the meanwhile. Extensive experiments on OTB2015, VOT2016 and UAV123 have illustrated the SiamDCA has effectively improved robustness of visual object tracking in scenes such as fast motion, background clutters, as illustrated in Figure 1.

2 Related Work

2.1 Siamese based trackers

Siamese based trackers [1, 11] has dominated tracking performance in recent years. A Siamese network takes an image pair as input, comprising a template branch and a search branch. SiamFC [1] first introduces correlation operators between the template branch and the search branch and highly improves the performance of trackers. Inspired by the Region Proposal Network (RPN) that is first proposed in Faster R-CNN [18], SiamRPN [12] performs the RPN extraction after the Siamese backbone to generate anchor boxes, which avoids the trouble of multi-scale testing in SiamFC. SiamRPN++ [11] uses ResNet [8] instead of AlexNet [10] in SiamRPN, so that the backbone can extract richer features. In addition, SiamRPN++ introduces depthwise separable convolution to reduce the amount of parameters in SiamRPN and futher improves the performance of Siamese architecture.

2.2 Attention mechanism

Recently, with the attention mechanism, many tasks in the field of computer vision and natural language processing have achieved better performance.

SENet[9] is an effective, lightweight attention mechanism that can self-recalibrate the feature map via channel-wise importance. It can effectively improve the representative quality of the network by explicitly modeling the interdependence between channels. As stated in SiamRPN++[11], each channel map of the high-level convolutional features usually responses for a specific object class. Therefore, we focus on improving the robustness of tracker from the perspective of channel attention. Extensive experiments have proved that our framework can improve the visual object tracking performance.

In Non-local [20], for each point on the feature map, it needs to dot product with all other points to obtain the spatial attention vector. This attention mechanism has recently been used in many fields such as instance segmentation and visual object tracking, which have achieved excellent performance. GCNet [2] is a simplified version of Non-Local, which calculates the spatial attention that is common to all points on the feature map, hence this attention only needs to be calculated once.

3 Method

Overview. We apply a ResNet50 embedded with the proposed CSENet as our backbone network of Siamese construction. On the one hand, as the layers become deeper, it computes increasingly high level features. On the other hand, under the influence of CSENet, high level features of the last three stages of the two branches will be transformed into features with channel attention. Then features with channel attention of template branch are enhanced by the proposed GCE module, generating weights of importance on each channel. Meanwhile, these weights are applied to the corresponding channel of the search branch to normalize the characteristics of the search branch. Finally, three Siamese RPN blocks described in [11] were feed the last three features with channel attention of template branch and the last three normalized features of search branch, generating dense response maps, which are further processed by a classification head and a bounding box regression head to predict the location of the target.

3.1 Siamese-based Trackers

The Siamese network based trachers formulate the visual object tracking as a similarity matching problem. There are two branchs in Siamese construction, the template branch and the search branch. They pass through a backbone with shared parameters, obtaining the target's feature (Z) and the search area (X) in a common embedding space. A cross-correlation operation between template features and search features is performed in the embedding space, generating a similarity map. Hence, this tracking process can be expressed as,

$$f_i(Z, X) = \phi(Z) \star \phi(X), i \in \{cls, reg\}. \tag{1}$$

where $\star$ denotes the cross-correlation operation, $\phi(\cdot)$ indicates the backbone of siamese network for feature extraction and i denotes the subtasks, where "cls" represents the classification head, "reg" represents the bounding box regression head.

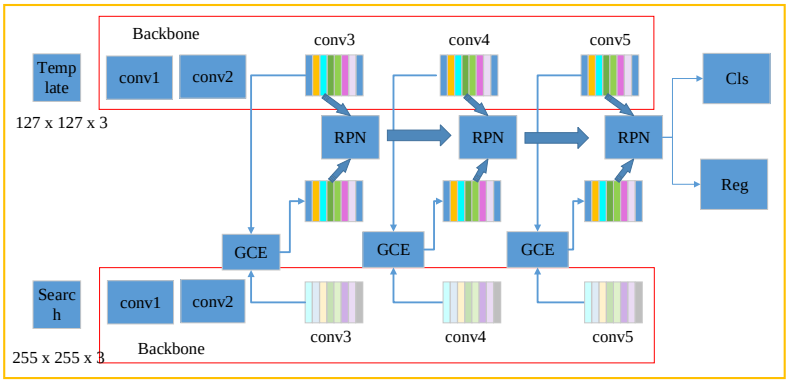


Fig. 2. An overview of the Dual Siamese Channel Attention Network (SiamDCA).

3.2 Convolutional SE Networks in backbone

As observed in SiamRPN++ [11], each channel map of the high-level convolutional features usually responses for a specific object class. Therefore, improving the correlation between the channels in the two branch features can greatly improve the accuracy and robustness of visual object tracking. Inspired by SENet [9], we embed the SE module into the last four blocks of backbone to perform dynamic channel-wise feature recalibration. In the meanwhile, in order to reduce the computation, we apply two convolution layers with 1×1 kernels instead of FC layers after squeeze operation, referred as CSENet, as illustrated in Figure 3. Consequently, the features of the search branch and the template branch are channel dependent.

In this work, we continue to follow the idea of the multi-level features in SiamRPN++ [11]. Features extracted from the last three residual blocks are used for the input of the subsequent network module. We refer these outputs as $F_3(z)$, $F_4(z)$ and $F_5(z)$, respectively. As shown in Figure 2, these three outputs are fed into three GCE modules and RPN module individually.

3.3 Global Channel Enhancement Module

As we all know, in Siamese architecture, features of the template branch and search branch are computed independently. However, different from the template branch, features in search branch is not sensitive to target. In order to distinguish the target from background, we use the template branch to guide the other branch.

As shown in Figure 4, the GCE block module contains two submodules, the first module is to strengthen the channel related features of the template branch. While the other module is to make use of enhanced channel-wise representation in the first module, then fully capture dependencies in channel level.

Specifically, in the first module, suppose the input features are $\mathbf{Z} \in \mathbf{R}^{C \times H \times W}$ in every block, we first apply two separate convolution with 1×1 kernels on $\mathbf{Z}$ to generate global features $\mathbf{G}$ and key features $\mathbf{K}$ respectively, where $\mathbf{G} \in \mathbf{R}^{C \times H \times W}$ and $\mathbf{K} \in \mathbf{R}^{1 \times H \times W}$. Then the $\mathbf{G}$ is filtered to $\widetilde{\mathbf{G}}$ that contains the crucial

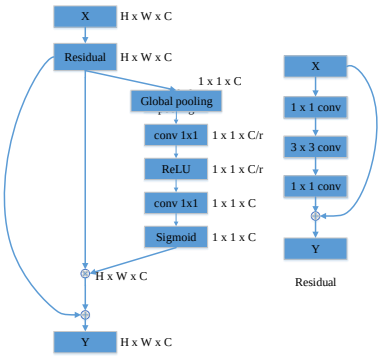


Fig. 3. Convolutional SE block used in last blocks of backbone, Residual is normal residual block in ResNet.

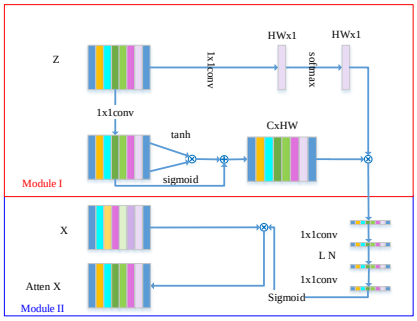


Fig. 4. An overview of the proposed Global Channel Enhancement (GCE) Module.

global information through the tanh and the sigmoid. Moreover, the significant information is added to the original information element-wise to ensure features are enhanced without losing the original feature information. This formula used for filtering as,

$$\widetilde{\mathbf{G}} = tanh(\mathbf{G}) \cdot sigmoid(\mathbf{G}) + \mathbf{G}. \tag{2}$$

Then the enhanced feature $\widetilde{\mathbf{G}}$ is reshaped to $\hat{\mathbf{G}} \in \mathbf{R}^{C \times N}$ where $N = H \times W$. Meanwhile, the $\mathbf{K}$ is reshaped to $\overline{\mathbf{K}} \in \mathbf{R}^{1 \times N}$ where $N = H \times W$, then we generate a filtered feature $\hat{K}$ via column-wise softmax operations. Finally we calculate the enhanced channel-related feature $\mathbf{O} \in \mathbf{R}^{C \times 1}$ as,

$$\mathbf{O} = \hat{\mathbf{G}}\hat{K} \in \mathbf{R}^{C \times 1}. \tag{3}$$

Then we reshape $\mathbf{O}$ to $\overline{\mathbf{O}} \in \mathbf{R}^{C \times 1 \times 1}$ and use $\overline{\mathbf{O}}$ as input for the next module.

There are two inputs in the second module, the first one is the output of the module $\overline{\mathbf{O}}$, while the other is the feature $\mathbf{X}$ of search branch. First of all, we apply a convolution layer with 1×1 kernels on $\overline{\mathbf{O}}$ to reduce the channels of $\overline{\mathbf{O}}$ to $\frac{C}{r}$ (we set r as 16). Besides, we add layer normalization to increase the convergence speed, as well as to benefit generalization ability. Finally we apply a convolution layer with 1×1 kernels to restore the channels of the strengthen feature, and use the feature to multiply $\mathbf{X}$ channel-wise to achieve the normalization of search branch features.

With our GCE module, the adjusted search features is sensitive to the target and objects are more discriminative against distractors and background.

4 Experiments

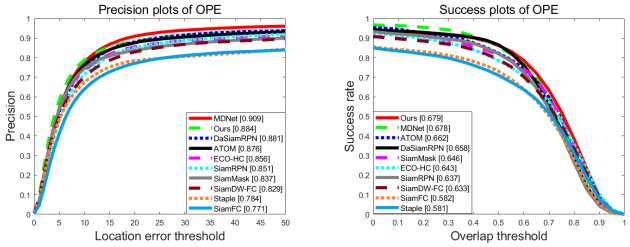
4.1 Implementation Details

The networks are trained on COCO [14], ImageNet DET [6], ImageNet VID [6], and YouTube-BoundingBoxes Dataset [17]. Our model is trained for 20 epochs, using a warmup learning rate in the first 5 epochs and a learning rate exponentially decayed in the last 15 epochs. Only the first layer of the weights of

backbone are frozen, for the first 10 epochs, then the whole networks are trained end-to-end for the last 10 epochs.

During inference, the regression network branch will predict more than one box. As [11], in order to get more accurate tracking results, we use scale change penalty to suppress large changes in target size and cosine window to suppress large displacements. These two penalties will eventually affect the classification score. We will re-rank the constrained score, and finally select the box corresponding to the maximum value of the classification score as the tracking result, this result is more accurate.

4.2 Comparisons with the State-of-the-art



(a) Precision Plot (b) Success Plot

Fig. 5. Comparisons on OTB2015.

We compare our SiamDCA tracker with the state-of-the-art trackers on three tracking benchmarks databases: OTB2015 [21], UAV123 [16], VOT2016 [7]. We tracker achieves state-of-the-art results and runs at 25 frames per second (fps).

OTB2015 [21]. OTB2015 is a commonly used benchmark, containing 100 sequences. It has two evaluation metrics, a precision score and an area under curve (AUC) of success plot respectively. As show in Figure 5, Our SiamDCA tracker is compared with numerous state-of-the art trackers include ATOM [3], ECO [4], DaSiamRPN [24] et al. In the process of comparison, our method ranks amount top-2 in the accuracy and success score.

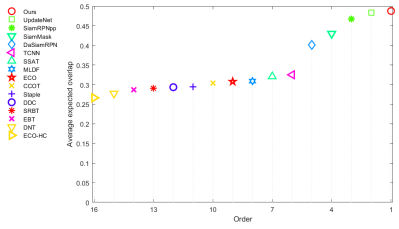


Fig. 6. Expected averaged overlap performance on VOT2016.

VOT2016 [7]. VOT2016 is a widely-used benchmarks for visual object tracking. It has three evaluation metrics, accuracy (A), robustness (R), and expected average overlap (EAO). We compare our SiamDCA tracker with the state-of-the-art trackers. As illustrated in Figure 6, our SiamDCA tracker achieves 0.485 EAO. Compared with recent SiamRPN++(SiamRPNpp)[11] and SiamMask [19], our methods increases 2.1% and 4.3% on EAO respectively.

Table 1. Result on UAV123. Red, blue represent 1st, 2nd respectively.

Tracker	AUC	Pr
SAMF [13]	0.395	0.592
SRDCF [5]	0.464	0.676
ECO [4]	0.525	0.741
SiamRPN [12]	0.527	0.748
DaSiamRPN [24]	0.586	0.796
SiamRPN++ [11]	0.613	0.807
ATOM [3]	0.644	-
Ours	0.630	0.822

Table 2. Ablation study on UAV123.

Method	AUC	Pr
Baseline	0.613	0.807
+CSE	0.614	0.814
+GCE(Ours)	0.630	0.822

UAV123 [16]. UAV123 is a new aerial video benchmark which includes 123 sequences. It has two evaluation metrics, the same with OTB2015. Table 1 illustrates the precision and success plots of the compared trackers. Specifically, our SiamDCA tracker achieves a success score of 0.630, which outperforms SiamRPN++ (0.613), DaSiamRPN (0.586), SiamRPN (0.527), ECO (0.525) and SRDCF (0.464). In addition, our tracker achieves a precision score of 0.822, and improvements of 1.5%, 2.6%, 7.4%, 8.1% and 14.6%, compared with SiamRPN++, DaSiamRPN, SiamRPN, ECO and SRDCF.

4.3 Ablation Study

We study the impact of individual components in SiamDCA on UAV123 to illustrate the role of each part. We use SiamRPN++ [11] as baseline. As illustrated in Table 2, By adding convolutional SE block to backbone of SiamRPN++, the precision score can be imported to 0.614. And the success plot score can further imported to 0.814. By adding GCE Module, the precision can be imported to 0.630, and the success score can further increased by +1.5%. Therefore, this result also proves the effectiveness of our method. With the CSE and the GCE, the SiamDCA achieve a more accurate tracking effect.

5 Conclusion

We have presented a new Dual Siamese Channel Attention Networks for visual object tracking. We introduce CSENet to backbone to extract features with channel relationship in template branch and search branch. In addition, we proposed GCE module to enhanced the channel-related features in template branch, then make use of dependencies of enhanced features in channel level to standardize the features of the search branch. Finally the robustness of tracking is improved effectively. Extensive experiments on three visual tracking benchmarks demonstrate that SiamDCA achieves state-of-the-art performance.

References

1. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.S.: Fully-convolutional siamese networks for object tracking (2016)
2. Cao, Y., Xu, J., Lin, S., Wei, Fangyun: Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops (Oct 2019)

3. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Atom: Accurate tracking by overlap maximization. In: CVPR. pp. 4660–4669 (June 2019)
4. Danelljan, M., Bhat, G., Shahbaz Khan, F., Felsberg, M.: Eco: Efficient convolution operators for tracking. In: CVPR. pp. 6638–6646 (July 2017)
5. Danelljan, M., Hager, G., Shahbaz Khan, F., Felsberg, M.: Learning spatially regularized correlation filters for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4310–4318 (December 2015)
6. Deng, Jia, Khosla, Aditya, Bernstein, Michael, Huang, Zhiheng, Krause, Jonathan, Su, Hao, Ma, Sean, Fei-Fei, Li, Karpathy, Andrej, Berg, C., A., Russakovsky, Olga, Satheesh, Sanjeev: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (5 2015)
7. Hadfield, S., Bowden, R., Lebeda, K.: The visual object tracking vot2016 challenge results. In: ECCV workshops. vol. 9914, pp. 777–823 (October 2016)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (June 2016)
9. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (99), 7132–7141 (2017)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (May 2017)
11. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: Siamrpn++: Evolution of siamese visual tracking with very deep networks. In: CVPR. pp. 4282–4291 (2019)
12. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with siamese region proposal network. In: CVPR. pp. 8971–8980 (June 2018)
13. Li, Y., Zhu, Jianke, e.L., Bronstein: A scale adaptive kernel correlation filter tracker with feature integration. In: ECCV Workshops. pp. 254–265 (2015)
14. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755 (2014)
15. Liu, L., Xing, J., Ai, H., Ruan, X.: Hand posture recognition using finger geometric feature. In: ICPR, pp. 565–568 (2013)
16. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for uav tracking. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV. pp. 445–461. Springer International Publishing, Cham (2016)
17. Real, E., Shlens, J., Mazzocchi, S., Pan, X., Vanhoucke, V.: Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In: CVPR. pp. 5296–5305 (July 2017)
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6), 1137–1149 (2017)
19. Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.: Fast online object tracking and segmentation: A unifying approach. In: CVPR. pp. 1328–1338 (June 2019)
20. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR. pp. 7794–7803 (June 2018)
21. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. TPAMI **37**(9), 1834–1848 (2015)
22. Xing, J., Ai, H., Lao, S.: Multiple human tracking based on multi-view upper-body detection and discriminative learning. In: ICPR, pp. 1698–1701 (2010)
23. Yu, Y., Xiong, Y., Huang, W., Scott, M.R.: Deformable siamese attention networks for visual object tracking. In: CVPR. pp. 6728–6737 (June 2020)
24. Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., Hu, W.: Distractor-aware siamese networks for visual object tracking. In: ECCV. pp. 103–119 (2018)