PG-Net: Pixel to Global Matching Network for Visual Tracking

Bingyan Liao*, Chenye Wang*, Yayun Wang, Yaonong Wang, and Jun Yin

ZheJiang Dahua Technology CO.,LTD, Hangzhou, China.
bingyanliao@outlook.com,
{wang_chenye, wang_yayun, wang_yaonong, yin_jun}@dahuatech.com

Abstract. Siamese neural network has been well investigated by tracking frameworks due to its fast speed and high accuracy. However, very few efforts were spent on background-extraction by those approaches. In this paper, a Pixel to Global Matching Network (PG-Net) is proposed to suppress the influence of background in search image while achieving state-of-the-art tracking performance. To achieve this purpose, each pixel on search feature is utilized to calculate the similarity with global template feature. This calculation method can appropriately reduce the matching area, thus introducing less background interference. In addition, we propose a new tracking framework to perform correlation-shared tracking and multiple losses for training, which not only reduce the computational burden but also improve the performance. We conduct comparison experiments on various public tracking datasets, which obtains state-of-the-art performance while running with fast speed.

1 Introduction

Visual object tracking is one of the fundamental problems in computer vision. It has been widely adopted in the field of intelligent transportation [25], robotics [12], video surveillance [31] and human-computer interactions [21], etc. Despite its rapid progress in recent decades, problems such as scene occlusion, target deformation and background interference still remain to be investigated. Recent years, convolutional neural network (CNN) has further improved the performance of trackers. Among them, Siamese network based trackers [2, 13, 18, 26–28, 38, 17] have drawn much attention in the community. The basic framework is proposed by Bertinetto *et al.* [2]: features of search image and target template are extracted by the same backbone network firstly, and then the cross-correlation is calculated based on features. To get more precise positions, SiamRPN [18] introduces RPN module to regress the bounding boxes. Based on SiamRPN, Bo Li *et al.* [17] design a deeper network to perform layer-wise and depthwise aggregations, which achieves higher accuracy and reduces the model size simultaneously.

Although Siamese network based algorithms excel in both accuracy and speed, those trackers cannot resist background interference effectively. We find

* Equal contribution

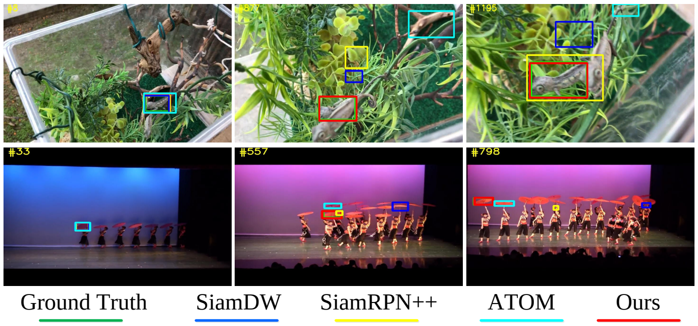


Fig. 1. Comparison result of SiamDW [36], SiamRPN++ [17], ATOM [5] and our method on two challenging sequences. PG-Net is able to distinguish the targets in *chameleon* and *umbrella*, even with strong background interference. The frame number is marked at the upper left corner of the image.

the primary reason comes from similarity calculation. Almost all Siamese trackers, such as SiamRPN [18], implement the similarity matching with a simple convolution operation on deep features. This results in matching region much larger than target area, and thus introduces a great deal of noise from background. The noise may overwhelm the target feature and lead to inaccurate matching.

To address these issues, we propose a Pixel to Global Matching Network (PG-Net), which resists background interference and finds a more accurate location of the target.Fig. 1 demonstrates such improvement — our PG-Net gives the most similar results to the ground truth. Specifically, we design a Pixel to Global Module (PGM) to realize similarity matching between template and search regions. Instead of using large matching regions, we utilize spatial pixels to calculate similarity of the template in feature domain. This operation reduces the size of matching area effectively, so that less background information is brought in and the network focuses more on target.

Further, we designed a new tracking framework to perform efficient and accurate tracking. We replace the crosss-correlation with proposed Pixel to Global matching correlation (PG-corr) to calculate the similarity of deep features. In order to reduce the calculation burden brought by the similarity calculation module, we calculate the classification and location with shared similarity maps. In the training phase, multiple loss functions are applied to different stages of backbone network to promote the tracking results.

Finally, we evaluate our PG-Net on four benchmark datasets, including VOT-2018 [16], VOT2018-LT [16], LaSOT [7] and OTB2015 [30]. And it performs best among other state-of-the-art trackers. In summary, the contributions of our work mainly include the following aspects.

- We propose a pixel to global similarity matching module to suppress background interference during tracking process.

- We design a new tracking framework based on proposed PGM, which not only reduces the computational burden but also improves the performance.
- We conduct comparison experiments on various tracking datasets, the results demonstrate that our method achieves state-of-the-art performance and fast speed.

The remaining parts of the paper are organized as follows: Section 2 briefly shows some relavant works in visual object tracking; Section 3 describes our proposed PG-Net; Section 4 evaluates PG-Net on four benchmarks; Section 5 concludes the paper.

2 Related works

In this section, some typical visual trackers proposed in recent years are reviewed. Existing tracking methods can be divided into: (i) correlation filter based [4, 15, 6] and (ii) deep learning based [2, 13, 18, 26–28, 38, 17]. The correlation filter based trackers includes: MOSSE [4] KCF [15], DSST [6], etc. They typically employ correlation filters to locate the targets based on handcrafted features. Compared with deep learning based counterparts, they are computationally efficient but less accurate.

With the development of deep learning technology and the establishment of large tracking datasets, many deep learning based tracking algorithms have emerged. Different from handcrafted features, features extracted by CNNs are more robust and contain more semantic information. Ran Tao *et al.* [26] first apply Siamese network to visual tracking tasks. The tracker simply finds the patch that matches best to the original patch of the target in the first frame. After that, Bertinetto *et al.* [2] propose a fully-convolutional Siamese network (SiamFC) to search the target from search image. Owing to the lightweight structure and end-to-end training manner, SiamFC receives significant attentions once it was proposed. Based on SiamFC, Valmadre *et al.* [27] embed a trainable correlation filter into the Siamese network, so that the correlation filter can be trained as part of the network. Qing Guo *et al.* [11] also propose a algorithm based on SiamFC. It learns appearance variation transformation and background suppression transformation online, which gets a better result. However, all of these methods locate target by searching the maximum value in the whole response map with no restriction of bounding box, leading to inaccuracy and lack of robustness.

Recently, some researchers are committed to applying detection technology to tracking tasks. Li *et al.* [18] combine Siamese network with Region Proposal Network (RPN) and significantly improve the accuracy of bounding box. And then, Zheng Zhu *et al.* [38] further increase the performance of SiamRPN by balancing distribution of training data. Heng Fan *et al.* [9] extend this approach by training a cascade of RPNs to solve the problem of class imbalance. The cascade of RPNs focuses more on hard samples by filtering out simple ones and makes the predicted bounding box more precise. Qiang Wang *et al.* [29] add a semantic segmentation subnetwork to RPN module and get pixel-level tracking

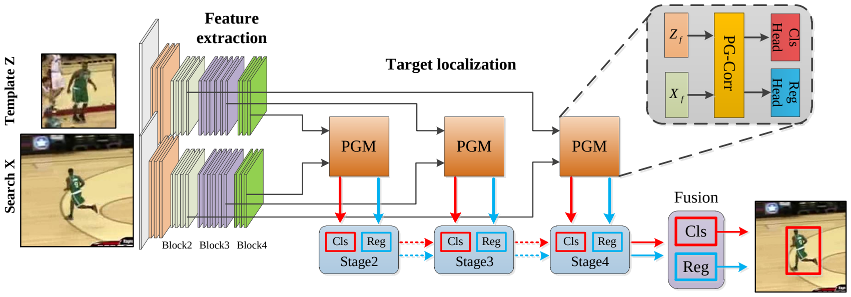


Fig. 2. Architecture of proposed network. It is composed of feature extraction subnetwork and target localization subnetwork. PGM is the Pixel to Global matching Module for similarity calculation. We utilize multiple PGMs to search target from different levels of features.

results. After that, Bo Li *et al.* [17] further propose SiamRPN++ which employs a deeper feature extraction network and more RPN modules, achieving state-of-the-art performance.

Although great progress has been made, these methods cannot suppress the background effectively. Besides, multiple cross-correlation layers in RPN lead to large computational complexity. In this work, we argue that the proposed PG-Net can reduce background interference effectively and achieve significant improvement on accuracy while reducing computational cost.

3 Pixel to Global Matching Network

In this section, we describe the proposed PG-Net in detail. First of all, we give the overview on the whole architecture of PG-Net. Secondly, we analyze how background interferes with object tracking, and propose PGM to mitigate interference. And then we elaborate the designed lightweight cross-correlation structure, which is good at reducing time consumption. Finally, multiple loss mechanism is introduced.

3.1 Overview

We design a Siamese network based tracker in this paper and its whole structure is shown in Fig. 2. The proposed network is composed of feature extraction subnetwork and target localization subnetwork. Here we employ ResNet50 [14] as the feature extraction subnetwork, and 3 PGMs to compose the localization subnetwork. For the loss function, we use multiple losses mechanism to further imporve the tracking accuracy.

ResNet50 has been proved to be a robust feature extractor in many computer vision tasks, such as object detection [37], classification [14] and semantic segmentation [10]. We modify the ResNet50 according to siamRPN++ [17] to make it more suitable for tracking tasks. The modified ResNet50 contains four residual blocks and each residual block is composed of a series convolution layers, batch normalization layers and activation layers. Different from original ResNet50, Block2, Block3 and Block4 have the same resolution in this version. In order to make pretrained weights available, Block3 and Block4 utilize dilated convolution layers to replace some convolution layers. It is noticed that the template branch and search branch have the same structure and share the same weights.

The localization subnetwork consists of 3 PGMs and each PGM outputs a set of classification (Cls) and regression (Reg) results based on densely distributed anchors. To increase the accuracy and robustness of the proposed algorithm, we select shallow, middle and deep level features as the input of the corresponding PGMs. These features are extracted from the last layers of Block2, Block3 and Block4 of the adapted ResNet50.

3.2 Pixel to Global matching Module

Cross-correlation is the core operation in Siamese tracker. Therefore, we first give a deep analysis on the cross-correlation, and illustrate several defects of it. Then we proposed PGM to address these issues.

Disadvantages of cross-correlation In existing methods, we observe that cross-correlation operation brings lots of background information in deep network. In tracking tasks, the given template is usually large to support backbone network to extract robust target features, which leads to a large output size of template subnetwork and might cause potential problems. Fig. 3(a) shows the process of searching a target and illustrates such problems. Directly mapping the coordinates of template feature to search image is expected to produce ideal matching region (green box), which has the same scale with target. However, this matching method ignores the influence of the receptive field which is one of the main factors to decide the real matching region (red box). With the network depth increasing, especially in deep network such as ResNet50, even a feature point in the final output corresponds a large receptive field of the input. Considering the large size of template feature, as shown in Fig. 3(a), the corresponding real matching region (red box) is much larger than ideal matching region. And thus, lots of background information will be brought in and overwhelms the feature of the target, making it hard to distinguish the target from similar objects in background.

We further find that the large matching region generates distributed response points, which increases the uncertainty of target localization. As demonstrated in Fig. 3(b), when searching the target (marked by the blue mask), a series of matching regions will be generated in search image. Here we take three matching regions as examples. The response (red point) is expected to appear only when the target locates in the center of matching region (red box), since this point can describe the location of the target best. However, the response points (yellow ones) are still generated even the region shifts a large range (yellow dotted box),

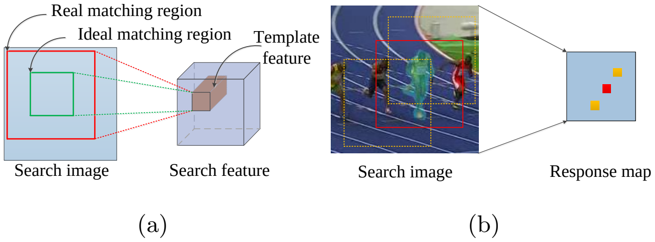


Fig. 3. We demonstrate how large matching region influence the tracking results. (a) explains why the real matching region is much larger than target. And (b) shows the influence of large matching region on results.

because the target is still in matching region. And more response points can be produced in a larger matching region, leading to inaccurate localization.

To avoid these problems, we propose the Pixel to Global correlation (PG-corr) to calculate the similarity, which can replace cross-correlation operation. And the improvements are visualized in next subsection.

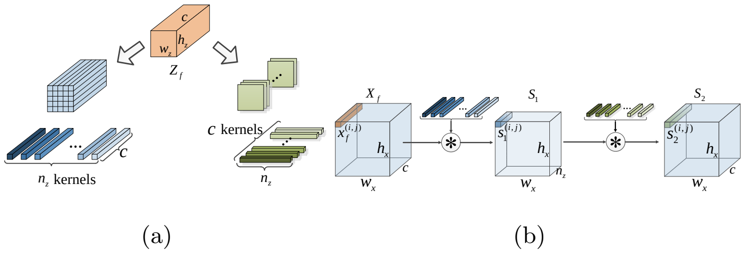
Pixel to Global matching Module Based on analysis above, we propose a Pixel to Global matching Module to calculate the similarity between each pixel on search feature and global template feature. Specifically, a pixel is a point whose length equals to the channel number of feature at a certain position. This module is composed of PG-corr and detector head for bounding box generation.

PG-corr has strong ability to suppress the interference of background, which outperforms the existing cross-correlation operation. This is mainly achieved by narrowing matching area in each search operation. Down-sampling the target feature is a straightforward method to reduce the area, but it causes substantial performance drop. In this paper, we reduce the match region by decomposing the template feature into spatial and channel kernels with size of 1 × 1, which suppresses the background interference effectively and gathers the response points on the target area accurately. This further improves the accuracy of predicted bounding box.

As shown in Fig. 4(a), the template feature Z_f is cut in height and width, forming a set of $Z_{f_S} = \{z_{f_S}^1, z_{f_S}^2, \ldots, z_{f_S}^{n_z}\}$, which has n_z kernels with length of c in spatial dimension, with

$$n_z = w_z \times h_z. \tag{1}$$

w_z and h_z are the width and height of template feature. Meanwhile, to enhance the channel correlation, the template feature is also cut in channel dimension, generating a set of $Z_{f_C} = \{z_{f_C}^1, z_{f_C}^2, \ldots, z_{f_C}^{c}\}$, which has c kernels with size of $1 \times 1 \times n_z$. The similarity calculation process is shown in 4(b), w_x and h_x are the width and height of search features X_f. For the position $x_f^{(i,j)}$ at rows j and columns i in X_f, we first calculate its similarity with spatial kernels. The *m-th*



(a) (b)

Fig. 4. The process of template feature decomposing and similarity matching. (a) shows the template feature is decomposed in spatial and channel dimensions respectively. (b) explains the matching process with decomposed kernels.

value in produced responses $S_1^{(i,j)}$ represents the similarity between $x_f^{(i,j)}$ and the *m-th* positions in spatial dimension of Z_f, which can be represented by

$$S_1^{(i,j)}[m] = x_f^{(i,j)} \cdot z_{f_S}^m \qquad m = 1, 2, \ldots, n_z. \tag{2}$$

To further acquire the similarity between $x_f^{(i,j)}$ and global template Z_f, we utilize channel kernels Z_{f_C} to unify the local positions similarity. After calculating similarity of all positions in X_f, the similarity map S_2 is obtained as

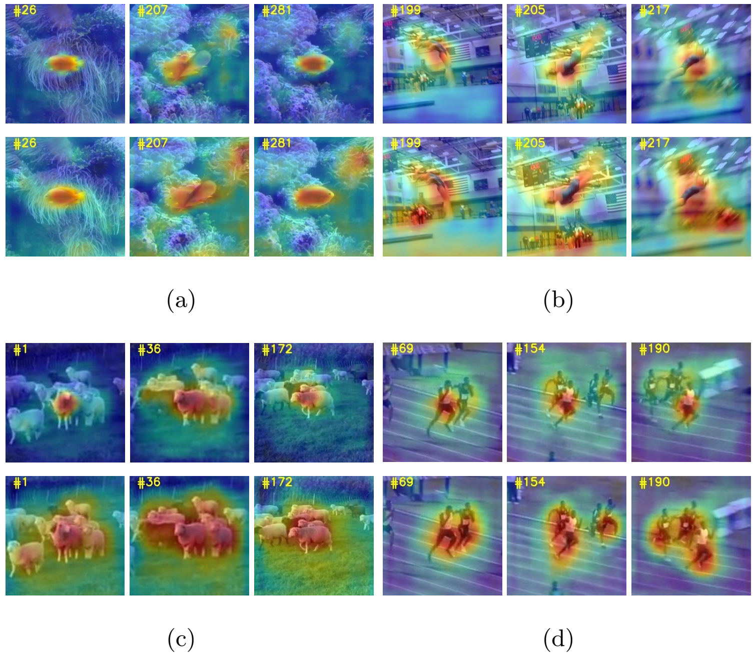
$$S_2^{(i,j)}[n] = S_1^{(i,j)} \cdot z_{f_C}^n \qquad n = 1, 2, \ldots, c. \tag{3}$$

For convenience, we define the PG-corr operation $S_2 = PG(X_f, Z_f)$ as

$$S_2[i,j,n] = \sum_{p,q,k} X_f[i,j,k] Z_f[p,q,k] Z_f[p,q,n]. \tag{4}$$

Where S_2 is the output feature of PG-corr with the same size of X_f. In order to reduce the difficulty of training, we concatenate search feature X_f and similarity feature S_2 in the channel dimension, and an 1×1 convolution layer follows to reduce the dimension. Following the PG-corr, to generate the target bounding box, we use fully convolutional layers to assemble the detector head.

In order to intuitively present the improvement of PGM on background suppression, we visualize the classification score map produced by different similarity matching methods. Examples from comparison results are shown in Fig. 5. The top rows are produced by PGM and the bottom rows based on depthwise correlation (DW-corr) operation. In Fig. 5(a) and 5(b), we obverse that the response region in score map of PGM is concentrated on target itself and the response in non-target areas is weak, while the response of DW-corr based module is strong in non-target areas. Especially when the background is complex as shown in Fig. 5(b), the response intensity in non-target areas is so close to it in the target



(a) (b)

(c) (d)

Fig. 5. Classification response maps of different sequences generated by PGM (top rows) and DW-corr based similarity matching module (bottom rows) respectively. We can find that the responses produced by PGM are more concentrated, and the responses of background are weak.

area, that the wrong detection is hard to avoid in this situation. According to Fig. 5(c) and 5(d), DW-corr based module is confused when there are similar objects in background, while PGM is still able to distinguish the targets. This mainly benefits from the pixel-level similarity matching method in PGM, which reduces the matching region to achieve precise matching.

3.3 Shared correlation architecture

Some Siamese network based trackers, such as SiamRPN [18] and Siamese Cascade RPN [9], apply regression branches to increase the accuracy of bounding box. As is shown in Fig. 6(a), the most popular mode is to perform a specific cross-correlation operation for each branch. For example, there are two cross-correlation layers in SiamRPN for classification and regression respectively. And the following SiamMASK [29] employs three cross-correlation layers for tracking and segmentation. The structure with multiple cross-correlation layers is computationally expensive.

Focus on this problem, we propose a shared correlation architecture to reduce time consumption. Different from existing methods in which each branch has the individual cross-correlation layer, we use just one shared correlation for both classification and regression branches. As is shown in Fig. 6(b), the extracted template and search features are first adjusted to squeeze the channel number by an 1×1 convolution layer (adjust layer). Then, the adjusted features are sent into PG-corr to perform similarity matching. In the final, we use the similarity map as the input of regression branch and classification branch to generate bounding box.

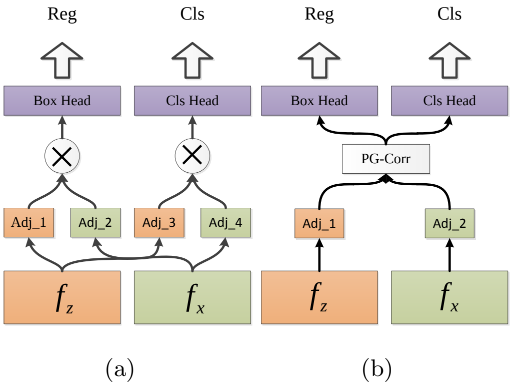


Fig. 6. Different connected methods between cross-correlation layers and two predict branches. (a) shows the connected method used in existing tracking network in which regression and classification branches have the individual cross-correlation layers. (b) is the shared connection method used in PGM. The regression and classification branches share the same PG-corr.

3.4 Multiple losses mechanism

Most existing methods just constrain the final feature with the corresponding loss function. However, only one optimized feature is difficult to perform perfectly though fusing multi-level features. In our method, we propose multiple losses mechanism to improve tracking performance. As is shown in Fig. 2, we apply loss functions on different stages of backbone network. Each loss function promotes corresponding features to be more robust and output more accurate regresses bounding boxes and classification scores. The loss function of single stage i is defined as

$$\mathcal{L}_{stage_i} = \mathcal{L}_{cls}(P_i, P^*) + \lambda \mathcal{L}_{reg}(B_i, B^*), \tag{5}$$

where λ is a hyper-parameter used to balance the two parts. $\mathcal{L}_{cls}$ is the Cross Entropy loss and $\mathcal{L}_{reg}$ is the Smooth L1 loss [18]. P_i and B_i represent the classification possibility and the predicted bounding box of the corresponding stage i.

P^* and B^* are the ground truth of classification and bounding box. Besides, we fusing the preliminary tracking results as our final output and optimize the fusion results as

$$\mathcal{L}_{fusion} = \mathcal{L}_{cls}(P_{fusion}, P^*) + \lambda \mathcal{L}_{reg}(B_{fusion}, B^*), \tag{6}$$

where

$$P_{fusion} = \sum_i \delta_i \ P_i \tag{7}$$

$$B_{fusion} = \sum_i \gamma_i \ B_i. \tag{8}$$

δ_i and γ_i are hyper-parameters optimized in training stage automatically. Our final loss function can be defined as follows

$$loss = \mathcal{L}_{fusion} + \sum_i \mathcal{L}_{stage_i}. \tag{9}$$

In inference stage, we only take the fusion output as our final outputs when tracking object.

4 Experiment

In this section, we first introduce the training details and hyper-parameters setting of the proposed method. After that, we evaluate our network on four public tracking test datasets, including VOT2018 [16], VOT2018-LT [16], LaSOT [7] and OTB2015 [30], and compare it with state-of-the-art trackers.

4.1 Implementation details

In training stage, the input size of search images is set to 255×255. And the template image size is set to 127×127, which is much larger than target area. After processing template image with feature extract network, we crop the center 7×7 regions as the template feature to reduce the influence of padding. Besides some common data argumentations, we add 10% negative sample pairs to improve the ability of network for discriminating difficult samples. The training loss is defined as Eq. (9) with balance factor $\lambda = 1.2$. The modified ResNet50 is initialized with pre-trained parameters on ImageNet and the other parts with random parameters. We train the proposed network with 430000 iterations with batch size of 28. For the first 110000 iterations, we train target localization subnetwork with warmup learning rate of 0.001 to 0.005. The following 320000 iterations are trained with learning rate decay from 0.005 to 0.0005. For the last 215000 iterations, the whole network is trained end-to-end.

The proposed method is trained on PyTorch deep learning framework with 8 TITANV GPUs and tested one TITAN X GPU. We utilize two large tracking datasets, including ImageNet VID [24] and YouTube-BoundingBoxes [23] datasets, and two large object detection datasets, including COCO [20] and ImageNet DET [24] datasets to train the network. Specifically, we crop the same image into template and search images respectively in ImageNet DET and COCO.

4.2 Ablation experiments

To investigate the impact of different similarity calculating methods, we train networks with depthwise correlation (DW-Corr) and PG-corr respectively. Evaluation results on VOT2018 are presented in Table 1. Network with PG-corr yields a great improvement compared with DW-Corr. The evidence shows that our PG-corr is a more efficient method for similarity matching.

To verify the improvement of applying multiple losses mechanism in our network, we use features from $Stage_2$, $Stage_3$ and $Stage_4$ respectively to track targets. Table 1 shows the EAO evaluated results on VOT2018. "$Stage_i$" shows the tracking results only from the coresponding stage i. And "Output" is our final tracking output from fusion results. "PG-Net" means PG-Net only trained with loss function (6). "PG-Net-mult-loss" means our network is traind with loss function (9). Experiment shows that results of each stage have obvious improvement after constraining each stage, which further promotes the final tracking results more accurate.

Table 1. Expected Average Overlap (EAO) comparation results on VOT2018 dataset for different similarity calculating methods and training strategies.

	$Stage_2$	$Stage_3$	$Stage_4$	Output
DW-Corr	/	/	/	0.408
PG-Net	0.146	0.264	0.029	0.427
PG-Net-mult-loss	0.299	0.344	0.313	0.447

4.3 Evaluation on VOT2018

We test the proposed network on VOT 2018 test dataset and compare it with 8 state-of-the-art methods, including Siamese network based algorithms and correlation filter based algorithms. VOT2018 is a public dataset for evaluating short-term performance of trackers, which contains 60 sequences totally with different challenging factors. We compare different trackers on Expected Average Overlap (EAO), Accuracy (A) and Robustness (Ro). The detailed comparison results are presented in Table 2. From Table 2, we observe that the proposed method

Table 2. Comparison results on VOT2018 dataset with performance measures of EAO, Accuracy and Robustness.

	DaSiam	UPDT	SiamRPN	MFT	LADCF	CFS-DCF	ATOM	SiamRPN++	**Ours**
	[38]	[3]	[18]	[1]	[33]	[32]	[5]	[17]	
EAO↑	0.326	0.378	0.383	0.385	0.389	0.397	0.401	0.414	0.447
A↑	0.569	0.536	0.586	0.505	0.503	0.511	0.590	0.600	0.618
Ro↓	0.337	0.184	0.276	0.140	0.159	0.143	0.204	0.234	0.192

achieves the best performance on EAO and Accuracy compared with existing methods. Especially in EAO score, our method achieves 0.447, which outperforms the state-of-the-art tracker SiamRPN++ with 0.414. The improvement mainly comes from the intelligent design of PGM. As for Roubustness, although weaker than online updating based methods, our network outperforms all offline tracking methods. Fig. 7 shows the evaluation results of EAO on VOT2018 dataset with respect to the Frames-Per-Second (FPS). According to the plot, the proposed method achieves the best performance among the compared methods while running with 42 FPS on one TITAN X GPU.

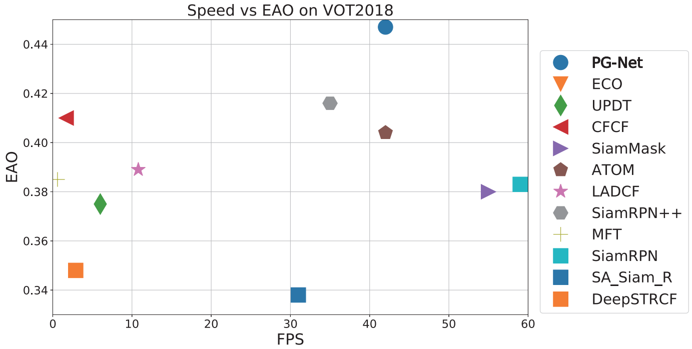


Fig. 7. The comparison of the quality and the speed with state-of-the-art trackers on VOT2018. We compare the EAO with respect to FPS.

4.4 Evaluation on VOT2018-LT

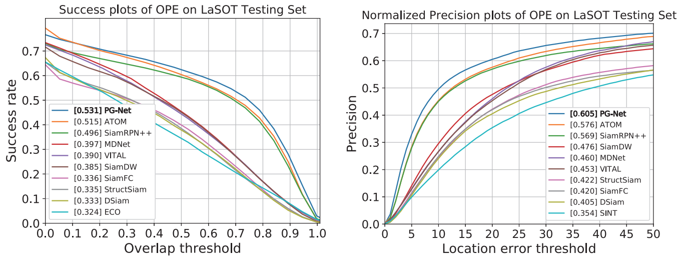
In VOT2018 challenge, a long-term tracking dataset (VOT2018-LT) is introduced. This dataset is composed of 35 long sequences. Targets in these sequences may be obscured completely or moved out of the lens for a long period. According to the statistics, the target in each video will disappear 12 times on average. There are three metrics used to evaluate the method, including Precision (P), Recall (R) and a combined F-score (F). According to the results shown in Table 3, the proposed PG-Net ranks 1st compared with state-of-the-art trackers on all metrics. Especially, our method significantly outperforms the state-of-the-art tracker SiamRPN++ by 3% on Precision.

4.5 Evaluation on LaSOT dataset

To further verify the performance of proposed method, we evaluate it on Large-scale Single Object Tracking (LaSOT). This dataset is composed of large scale high quality sequences. There are totally 280 videos and 70 categories in the

Table 3. Comparison with state-of-the-art trackers on VOT2018-LT tracking dataset.

	SiamVGG	FuCoLoT	PTAVplus	LTSINT	MMLT	DaSiam	MBMD	SPLT	SiamRPN++	**Ours**
	[19]	[22]	[8]	[16]	[16]	[38]	[35]	[34]	[17]	
F↑	0.459	0.480	0.481	0.536	0.546	0.607	0.610	0.616	0.629	0.642
P↑	0.552	0.538	0.595	0.566	0.574	0.627	0.634	0.633	0.649	0.679
R↑	0.393	0.432	0.404	0.510	0.521	0.588	0.588	0.600	0.610	0.610



(a) (b)

Fig. 8. Evaluation results on LaSOT dataset. (a) is the success rate curves and (b) is precision curves.

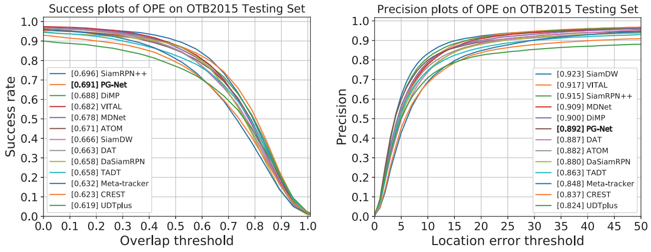
dataset. Similar to VOT2018-LT, it focuses on long-term tracking with average sequence length of 2512 frames. Fig. 8 reports the evaluation results of the proposed method and the comparison methods. We compare success rate and normalized precision among these methods. Our method achieves success rate of 53.1% which outperforms ATOM by 1.6%. And the normalized precision surpasses other method by 2.9%.

4.6 Evaluation on OTB2015

We also compare the performance with other state-of-the-art methods on OTB-2015 dataset. OTB2015 is a widely used tracking benchmark consists of 100 sequences. Notice that there is no any reset and updating in the whole tracking process, which provides a fair testbed on robustness. Here we measure success rate and precision for comparison. The evaluated results are shown in Fig. 9. Our method achieves a comparable results with state-of-the-art method SiamRPN++ on success rate.

5 Conclusion

In this paper, we present a novel Pixel to Global Matching Network to achieve high performance similarity matching by suppressing the influence of background. We show theoretical and empirical evidence that how PGM suppresses



(a) (b)

Fig. 9. Success plots and precisions plots show the comparison of our method with other state-of-the-art methods on OTB2015 dataset.

background in similarity matching. And by employing a lightweight network structure and multiple losses mechanism, our approach can reduce the computational complexity and further improve the tracking accuracy. Comprehensive experiments are conducted on VOT2018, VOT2018-LT, LaSOT and OTB2015 tracking benchmarks. The results show that our approach achieves state-of-the-art performance.

Acknowledgment

This work is generous supported by DAHUA Advanced Institute and Deep Learning Platform of Jinn. In this work, Bingyan Liao and Chenye Wang contribute equally, and Bingyan Liao is also the corresponding author.

References

1. Bai, S., He, Z., Xu, T.B., Zhu, Z., Dong, Y., Bai, H.: Multi-hierarchical independent correlation filters for visual tracking. arXiv preprint arXiv:1811.10302 (2018)
2. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 850–865. Springer (2016)
3. Bhat, G., Johnander, J., Danelljan, M., Shahbaz Khan, F., Felsberg, M.: Unveiling the power of deep tracking. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 483–498 (2018)
4. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 2544–2550. IEEE (2010)
5. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Atom: Accurate tracking by overlap maximization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4660–4669 (2019)
6. Danelljan, M., Häger, G., Khan, F., Felsberg, M.: Accurate scale estimation for robust visual tracking. In: British Machine Vision Conference, Nottingham, September 1-5, 2014. BMVA Press (2014)
7. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: Lasot: A high-quality benchmark for large-scale single object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5374–5383 (2019)
8. Fan, H., Ling, H.: Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5486–5494 (2017)
9. Fan, H., Ling, H.: Siamese cascaded region proposal networks for real-time visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7952–7961 (2019)
10. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3146–3154 (2019)
11. Guo, Q., Feng, W., Zhou, C., Huang, R., Wan, L., Wang, S.: Learning dynamic siamese network for visual object tracking. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1763–1771 (2017)
12. Gupta, M., Kumar, S., Behera, L., Subramanian, V.K.: A novel vision-based tracking algorithm for a human-following mobile robot. IEEE Transactions on Systems, Man, and Cybernetics: Systems **47**(7), 1415–1427 (2016)
13. He, A., Luo, C., Tian, X., Zeng, W.: A twofold siamese network for real-time object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4834–4843 (2018)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
15. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE transactions on pattern analysis and machine intelligence **37**(3), 583–596 (2014)
16. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Cehovin Zajc, L., Vojir, T., Bhat, G., Lukezic, A., Eldesokey, A., et al.: The sixth visual object tracking vot2018 challenge results. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018)

17. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: Siamrpn++: Evolution of siamese visual tracking with very deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4282–4291 (2019)
18. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with siamese region proposal network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8971–8980 (2018)
19. Li, Y., Zhang, X.: Siamvgg: Visual tracking using deeper siamese networks. arXiv preprint arXiv:1902.02804 (2019)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 740–755. Springer (2014)
21. Liu, L., Xing, J., Ai, H., Ruan, X.: Hand posture recognition using finger geometric feature. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). pp. 565–568. IEEE (2012)
22. Lukežič, A., Zajc, L.Č., Vojíř, T., Matas, J., Kristan, M.: Fucolot–a fully-correlational long-term tracker. In: Asian Conference on Computer Vision. pp. 595–611. Springer (2018)
23. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5296–5305 (2017)
24. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
25. Saunier, N., Sayed, T.: A feature-based tracking algorithm for vehicles in intersections. In: The 3rd Canadian Conference on Computer and Robot Vision (CRV'06). pp. 59–59. IEEE (2006)
26. Tao, R., Gavves, E., Smeulders, A.W.: Siamese instance search for tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1420–1429 (2016)
27. Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A., Torr, P.H.: End-to-end representation learning for correlation filter based tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2805–2813 (2017)
28. Wang, Q., Teng, Z., Xing, J., Gao, J., Hu, W., Maybank, S.: Learning attentions: residual attentional siamese network for high performance online visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4854–4863 (2018)
29. Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.: Fast online object tracking and segmentation: A unifying approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1328–1338 (2019)
30. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(9), 1834–1848 (2015)
31. Xing, J., Ai, H., Lao, S.: Multiple human tracking based on multi-view upper-body detection and discriminative learning. In: 2010 20th International Conference on Pattern Recognition. pp. 1698–1701. IEEE (2010)
32. Xu, T., Feng, Z.H., Wu, X.J., Kittler, J.: Joint group feature selection and discriminative filter learning for robust visual object tracking. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)

33. Xu, T., Feng, Z.H., Wu, X.J., Kittler, J.: Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking. IEEE Transactions on Image Processing (2019)
34. Yan, B., Zhao, H., Wang, D., Lu, H., Yang, X.: 'skimming-perusal' tracking: A framework for real-time and robust long-term tracking. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
35. Zhang, Y., Wang, D., Wang, L., Qi, J., Lu, H.: Learning regression and verification networks for long-term visual tracking. arXiv preprint arXiv:1809.04320 (2018)
36. Zhang, Z., Peng, H.: Deeper and wider siamese networks for real-time visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4591–4600 (2019)
37. Zhao, Q., Sheng, T., Wang, Y., Tang, Z., Chen, Y., Cai, L., Ling, H.: M2det: A single-shot object detector based on multi-level feature pyramid network. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9259–9266 (2019)
38. Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., Hu, W.: Distractor-aware siamese networks for visual object tracking. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 101–117 (2018)