BOUNDING BOX DISTRIBUTION LEARNING AND CENTER POINT CALIBRATION FOR ROBUST VISUAL TRACKING

Chihui Zhuang[1], *Yanjie Liang*[1,2], *Yan Yan*[1], *Yang Lu*[1], *Hanzi Wang*[1,⋆]

[1]Fujian Key Laboratory of Sensing and Computing for Smart City, Xiamen University, Xiamen, China [2]Peng Cheng Laboratory, Shenzhen, China

chzhuang@stu.xmu.edu.cn, yanjieliang@yeah.net, {yanyan, luyang, hanzi.wang}@xmu.edu.cn

ABSTRACT

Visual tracking aims at both robust target classification and accurate localization. However, the reliability of the target bounding box and classification score are not properly addressed by most existing trackers, resulting in inaccurate tracking performance. In this paper, we propose to learn bounding box distribution in training and calibrate the center point response in inference for robust online tracking. Specifically, we propose a simple yet effective bounding box distribution learning (BDL) module to model the target bounding box distribution and enhance the localization ability of our network. Furthermore, we propose a center point calibration (CPC) module to calibrate the origin classification score with the predicted localization uncertainty and generate an accurate target center point. The proposed tracking method is referred to as DLPC. The experimental results on four challenging datasets (i.e., OTB100, VOT2019, LaSOT, and TrackingNet) show that DLPC performs favorably against several state-of-the-art trackers while running in real-time at 60 fps.

***Index Terms*—** Visual tracking, bounding box distribution learning, center point calibration

1. INTRODUCTION

Visual tracking is a fundamental task in computer vision, which aims to automatically localize an arbitrary target object given in the first frame over a video sequence. It has numerous practical applications, such as autonomous driving, video surveillance, and human-computer interaction. Although great efforts have been dedicated to improving the tracking performance, visual tracking is still a challenging task due to numerous factors such as occlusion, illumination variations, and background clutters.

Recently, the Siamese network based trackers [1, 2, 3, 4, 5, 6] have drawn much attention due to its simple structure and promising performance. Bertinetto *et al.* [1] firstly introduce a fully convolutional Siamese network to indicate the similarity between various candidate features and the template features for visual tracking. Inspired by the regression methods in object detection, the current Siamese network based trackers formulate the tracking problem as a combination of classification and regression, which yields a significant improvement in terms of tracking accuracy. However, as most current trackers, the performance of the Siamese network based trackers is still inferior due to their unreliable classification and regression. For example, SiamRPN [2] and SiamRPN++ [3] raise a certain number of anchors in each frame and directly regress the bounding box of the target from the peak response anchor, without any confidence information. Similarly, SiamFC++ [4] performs anchor-free tracking while estimating the classification score and regressing the bounding box of the target at each searching position, which neglects the reliability of the classification and regression results. As a consequence, these trackers may result in target drift and mislocalization during the tracking process, as shown in Fig. 1.

(a) (b)

Fig. 1. Two failure cases of SiamFC++ [4] in yellow compared with DLPC in red and the ground-truth in green. It shows that improving the reliability of the bounding box and the target center point alleviates inaccurate tracking such as (a) target drift and (b) mislocalization.

To alleviate the above problem, we propose an online tracking method named DLPC, built upon SiamFC++. Mo-

⋆Corresponding Author: Hanzi Wang, hanzi.wang@xmu.edu.cn

The research work is supported in part by the National Natural Science Foundation of China under Grant 61872307; in part by the Open Research Projects of Zhejiang Lab (NO. 2021KB0AB03); in part by the National Natural Science Foundation of China under Grant No. U21A20514, 62071404, and 62002302.

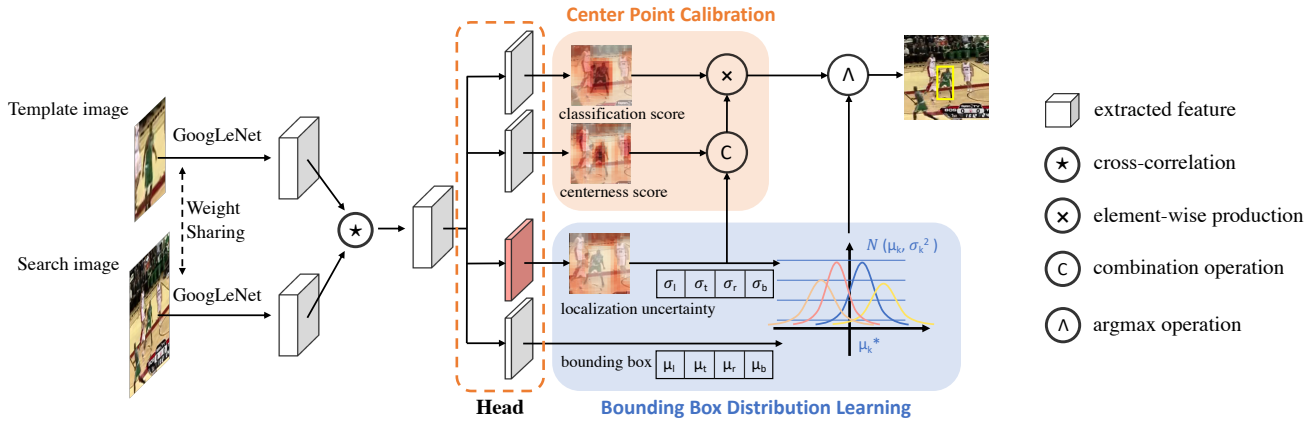

Fig. 2. Framework of the proposed DLPC, which consists of a bounding box distribution learning (BDL) module and a center point calibration (CPC) module. For visualization in the score map, deep red color represents the high response area. Note that the uncertainty σ_k is low when the bounding box prediction μ_k is closer to the ground-truth μ_k^*, where $k \in \{l, t, r, b\}$.

tivated by the object detection methods [7, 8] with reliable localization, we model the target bounding box as univariate Gaussian distributions based on anchor-free tracking methods and propose a bounding box distribution learning (BDL) module. Specifically, the BDL module involves of a localization uncertainty branch and a distribution learning loss to learn the bounding box uncertainty, which explores the reliability of the regression results. Furthermore, we propose a center point calibration (CPC) module for effectively using the uncertainty, which enhances the response peak of the target center point, thus achieving robust visual tracking. We validate the performance of the proposed DLPC on four popular benchmark datasets, and experimental results show that DLPC performs favorably against several state-of-the-art trackers while running at the real-time speed of 60 fps.

2. PROPOSED METHOD

2.1. Overall Framework

Fig. 2 illustrates the overall framework of the proposed DLPC. The input of our network consists of a template z in the initial frame and a search region x in the current frame. Firstly, we perform the same transformation on the input images and embed them into a common feature space with shared parameters between two branches. Then, the deep Siamese network learns a similarity function $F(z, x)$ to compute the similarity score, which is computed by $\varphi(z) \star \varphi(x)$, where $\star$ denotes the cross-correlation operation, and $\varphi(\cdot)$ denotes the feature embedding function. The similarity feature is employed for the classification head and the regression head.

We insert the BDL module collaborated with the bounding box branch in the regression head, jointly learning distributions of the target boxes and predicting the localization uncertainty. Simultaneously, we apply the CPC module to further utilize the localization uncertainty and generate an accurate target center point. By combining both the BDL module and the CPC module, we explore the potential of the bounding box distribution while achieving accurate classification and localization based on an anchor-free tracking framework.

2.2. Training with the BDL module

For each position in the search region, DLPC regresses four distances (left, right, top, and bottom) from the current position to the object boundaries in the bounding box prediction branch, which can be denoted as a 4D vector $\boldsymbol{p}_{x,y} = [l, t, r, b]^\top$. To alleviate the negative influence of the inaccurate classification and regression results, we firstly focus on the bounding box regression process and model the outputs as four univariate Gaussian distributions. Formally, we propose a localization uncertainty branch to estimate the localization uncertainty along with the bounding box offsets (l, t, r, b). We treat the estimated localization uncertainty as the standard deviation while considering the predicted bounding box offset as the mean value of the predicted distribution. Thus, the BDL module can learn the Gaussian probability distribution of the bounding box. Assuming that each offset is independent, the distribution is defined as:

$$G(\mu_k) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(\mu_k - \mu_k^*)^2}{2\sigma_k^2}}, k \in \{l, t, r, b\}, \quad (1)$$

where σ_k is the standard deviation corresponding to the distribution which measures the localization uncertainty. μ_k and μ_k^* are respectively the mean values of the predicted bounding box offset and the ground-truth. We train our DLPC to learn uncertainties $(\sigma_l, \sigma_t, \sigma_r, \sigma_b)$ in the localization uncertainty branch for each box offset, as shown in Fig. 2.

According to the learned distribution, we use the area under the probability density function curve to represent the probability, where the bounding box offset is within the range between the ground-truth and the current offset prediction. It is worth pointing out that this area value is proportional to the value of probability density. To obtain the area and represent the probability $P(\mu)$, we integrate the Gaussian probability density value $G(\mu)$ with the offset of k as the following form:

$$P(\mu) = 1 - 2 \int_{\min(\mu,\mu^*)}^{\max(\mu,\mu^*)} G(\mu)\, d\mu. \quad (2)$$

According to the above definition of the bounding box distribution corresponding to localization uncertainty, the distribution learning loss $\mathcal{L}_d$ in the BDL module is designed as follows:

$$\mathcal{L}_d = -\frac{1}{N_{\text{pos}}} \sum_{x,y} \sum_{k} 1_{\{c^*_{x,y}>0\}} (1 - IoU_{x,y}) \log P(\mu_{x,y,k}), \quad (3)$$

where N_{pos} denotes the number of positive samples. $1_{\{\cdot\}}$ is the indicator function that takes 1 if the condition in subscribe holds and takes 0 otherwise. $c^*_{x,y}$ is assigned to 1 if (x, y) is considered as a positive sample, and 0 if it is as a negative sample. $IoU_{x,y}$ is the intersection-over-union between the predicted bounding box and the ground-truth bounding box at location (x, y). $\mu_{x,y,k}$ is the predicted bounding box offset, where $k \in \{l, t, r, b\}$. We minimize a negative log likelihood loss constructed from a Gaussian probability density function during training. Note that our loss considers IoU to assist the prediction of uncertainty, and we expect a better drop in the distribution learning loss when the predicted box achieves a higher IoU score. By using the distribution learning loss, the BDL module can effectively explore the potential target localization, which boosts the localization performance.

2.3. Inference with the CPC module

The proposed BDL module learns the bounding box distribution and outputs the localization uncertainties corresponding to each predicted box $\boldsymbol{p}_{x,y} = [l, t, r, b]^\top$. Then, we propose the CPC module for accurate inference, as shown in Fig. 2, which employs the uncertainties to calibrate the classification score. For each location, we compute the average localization uncertainty σ_{avg}, and transform it into the center point confidence score. We multiply the center point confidence score by the centerness score to calibrate the classification score for generating a reliable center point response. The above operation is defined as:

$$Score = (1 - \sigma_{\text{avg}}) \times \delta_{\text{ctr}} \times \delta_{\text{cls}}, \quad (4)$$

where $Score$ indicates the final response in each frame, δ_{ctr} is the centerness score, and δ_{cls} is the classification score. The proposed CPC module considers both the localization uncertainty and the original centerness of the predicted target, which provides accurate classification and localization during the tracking process.

Table 1. Ablation study to evaluate the proposed BDL and CPC modules on the LaSOT dataset. The best results of each column are highlighted by **bold**.

BDL	CPC	Succ.(AUC)(%)	Norm. Prec.(%)	Prec.(%)
		55.0	58.2	54.8
✓		56.2	59.8	56.2
✓	✓	**57.4**	**60.9**	**57.6**

3. EXPERIMENTS

3.1. Implementation Details

The proposed DLPC is bulit upon SiamFC++ [4]. ILSVRC-VID/DET [9], COCO [10], LaSOT [11], GOT-10k [12] and TrackingNet [13] datasets are adopted as the training sets. We choose the stochastic gradient descent (SGD) with a momentum of 0.9 to train the proposed DLPC with the learning rate linearly increasing from 10^{-7} to 2×10^{-2}. The number of image pairs per epoch is set to 300k. 5 epochs are used for warming-up and 15 epochs are used for training. The weight of the bounding box distribution learning loss in the BDL module is experimentally set to 3.0, which is the same as the weight of the regression loss. All experiments are evaluated on an Intel(R) E5 2.1GHz CPU and a single NVIDIA RTX 2080Ti GPU with 11G memory.

3.2. Ablation study

In this section, we carry out the ablation study on the LaSOT [11] dataset to analyze the effectiveness of the BDL module and the CPC module. We compare three variants, including the baseline method, the baseline method only with the BDL module, and DLPC with both the BDL module and the CPC module.

The evaluation results are reported in Table 1. As shown in Table 1, the proposed BDL module boosts the success score and the normalized precision score of the baseline tracker by 1.2% and 1.6%, respectively, which shows that the bounding box distribution learning process assists the network to localize the target more accurately. Furthermore, by introducing the CPC module to generate a precise target center point, the method achieves a 2.4% gain on the success score, a 2.7% gain on the normalized precision score and a 2.8% gain on the precision score. The experimental results show the effectiveness of the proposed BDL and CPC modules in DLPC.

3.3. Comparison with State-of-the-art Methods

We compare our DLPC with several state-of-the-art trackers including SiamFC [1], MDNet [14], SiamRPN++ [3], ATOM [15], DiMP [16], SiamFC++ [4] on four challenging datasets. Table 2 reports the comparison results.

Table 2. Results on several benchmarks. T-Net denotes TrackingNet. The top three best results of each row are highlighted by red, green and blue, respectively.

Trackers		SiamFC (2016)	MDNet (2016)	SiamRPN++ (2019)	ATOM (2019)	DiMP (2019)	SiamFC++ (2020)	**DLPC (Ours)**
OTB100	Succ.	58.2	67.8	69.6	66.9	68.4	68.3	69.9
VOT2019	A↑	0.470	-	0.599	0.603	0.594	0.575	0.596
	R↓	0.958	-	0.482	0.411	0.278	0.406	0.414
	EAO↑	0.163	-	0.285	0.292	0.379	0.284	0.297
LaSOT	Succ.	33.6	39.7	49.6	51.5	56.9	54.4	57.4
T-Net	Prec.	51.8	56.5	69.4	64.8	68.7	70.5	71.0
	N. Prec.	65.2	70.5	80.0	77.1	80.1	80.0	81.6
	Succ.	55.9	60.6	73.3	70.3	74.0	75.4	75.0
FPS		86	1	35	30	43	90	60

OTB100 [17]. OTB100 is one of the most commonly used dataset and it consists of 100 fully annotated video sequence with various attributes. As shown in Table 2, DLPC achieves the state-of-the-art success score of 69.9%, which is the best among the competing trackers. Specifically, it notably outperforms the baseline SiamFC++ with a relative gain of 1.6%. Compared with the powerful regression based tracker DiMP, DLPC still yields a relative gain of 1.5%. The results show that DLPC can perform target classification and bounding box estimation accurately, as shown in Fig. 3.

VOT2019 [18]. VOT2019 contains 60 video sequences with various challenging factors. Without bells and whistles, DLPC achieves a competitive EAO score of 0.297. DiMP performs better due to its online updating module, which is effective to distinguish similar targets during tracking. However, DLPC achieves better performance than DiMP on the other datasets due to its robust target classification ability and accurate bounding box prediction, and it runs faster than DiMP. It is worth pointing out that DLPC outperforms the baseline tracker SiamFC++ by 1.3% on the EAO metric and 2.1% on the accuracy metric. This validates that DLPC can effectively estimate the distribution of target bounding box and accurately localize the center point during the tracking process.

LaSOT [11]. LaSOT is a large-scale dataset with 1,400 video sequences in total, and 280 video sequences in the test set. Since long sequences are common in LaSOT, there are various deformation and occlusion in the video and it provides high-quality dense annotations. The tracking results show DLPC achieves the state-of-the-art success score of 57.4%, which is the best among the competing trackers. Specifically, DLPC outperforms the baseline tracker SiamFC++ with a substantial gain of 3.0% in terms of the success score. Experimental results on LaSOT show that DLPC can achieve robust tracking with accurate classification and localization on long video sequences.

TrackingNet [13]. TrackingNet contains 30,000 video sequences with 14 million dense annotations, and we evaluate DLPC on its test set with 511 video sequences. As shown in the Table 2, DLPC achieves the best results on the precision score of 71.0% and the normalized precision score of 81.6%. Note that DLPC outperforms SiamFC++ by 1.6% on the normalized precision score, which potentially shows the capability of DLPC for accurate classification and localization. Moreover, DLPC outperforms SiamRPN++ and DiMP by 1.7% and 1.0% in terms of the success score, respectively, and DLPC runs faster than both of them (about 71% faster than SiamRPN++ and 40% faster than DiMP). We find that in more complex tracking scenarios or for longer video sequences, DLPC is less likely to lose the target by paying more attention to accurate localization.

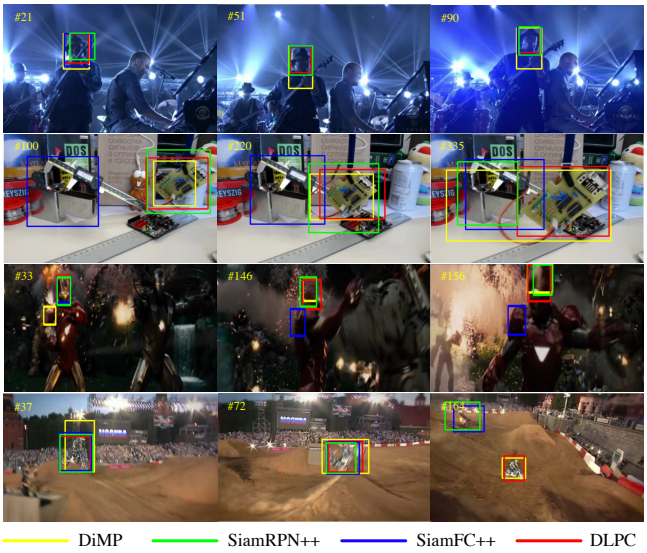

Fig. 3. Comparisons of DLPC with DiMP [16], SiamRPN++ [3] and SiamFC++ [4] on four challenging videos from OTB100 [17]. From top to bottom, the videos are respectively Shaking, Board, Ironman and MotorRolling.

4. CONCLUSION

In this paper, we propose a robust online tracking method DLPC, which mainly contains a bounding box distribution learning (BDL) module and a center point calibration (CPC) module. Specifically, the proposed BDL module learns the target bounding box distribution, which improves the localization performance of our tracker. Moreover, the proposed CPC module effectively utilizes the uncertainty predicted in the bounding box distribution to calibrate the classification score, which ensures accurate center point estimation. Extensive experiments on four challenging datasets (i.e., OTB100, VOT2019, LaSOT and TrackingNet) show that the proposed DLPC can significantly boost the tracking performance in comparison with several state-of-the-art real-time trackers while running at a fast speed of 60 fps.

5. REFERENCES

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr, "Fully-convolutional siamese networks for object tracking," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 850–865.

[2] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu, "High performance visual tracking with siamese region proposal network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8971–8980.

[3] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan, "Siamrpn++: Evolution of siamese visual tracking with very deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4282–4291.

[4] Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu, "Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 12549–12556.

[5] Heng Fan, Lu Xu, and Jinhai Xiang, "Complementary siamese networks for robust visual tracking," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 2247–2251.

[6] Ying Hu, Hanyu Xuan, Jian Yang, and Yan Yan, "Channel attention based generative network for robust visual tracking," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 4082–4086.

[7] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang, "Bounding box regression with uncertainty for accurate object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2888–2897.

[8] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee, "Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 502–511.

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 740–755.

[11] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling, "Lasot: A high-quality benchmark for large-scale single object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5374–5383.

[12] Lianghua Huang, Xin Zhao, and Kaiqi Huang, "Got-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1562–1577, 2021.

[13] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem, "Trackingnet: A large-scale dataset and benchmark for object tracking in the wild," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 300–317.

[14] Hyeonseob Nam and Bohyung Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4293–4302.

[15] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg, "Atom: Accurate tracking by overlap maximization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4660–4669.

[16] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte, "Learning discriminative model prediction for tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6182–6191.

[17] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang, "Object tracking benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, 2015.

[18] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Luka ˇCehovin Zajc, Ondrej Drbohlav, Alan Lukezic, Amanda Berg, et al., "The seventh visual object tracking vot2019 challenge results," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.