ADVERSARIAL LEARNING-BASED DATA AUGMENTATION FOR ROTATION-ROBUST HUMAN TRACKING

Kexin Chen[#], Xue Zhou[#1], Qidong Zhou[#] and Hongbing Xu[#]

[#]School of Automation Engineering, University of Electronic Science and Technology of China
[1]zhouxue@uestc.edu.cn

ABSTRACT

This paper analyzes the diversity deficiency of positive training samples used to fine-tune CNN-based tracking networks, especially when confronted with large pose changes and out-of-plane rotation challenges. Therefore, we present a novel adversarial learning-based hard positives generation method and embed it into the multi-domain network (MDNet)-based tracking framework. Instead of adopting the dense sampling strategy to generate monotonous positive samples, we cast it as a cross-domain image transformation problem, which is designed to be able to generate hard positive samples with more diversity and some degree of motion blur and pose direction changes. Experimental results on tracking benchmark demonstrate the effectiveness and robustness of our proposed method compared with the state-of-art trackers.

Index Terms— Human tracking, Out-of-plane rotation, GAN-based data augmentation, Hard positives, Multi-domain network

1. INTRODUCTION

Visual single object tracking is one of the fundamental problems in computer vision, the aim of the tracking task is to locate a specific target robustly and accurately in a video sequence[1]. Despite the impressive results current algorithms have achieved, non-rigid object (especially for human) tracking still confronts with several challenging situations, such as large pose variations, out-of-plane rotation, motion blur, etc.

With the rapid development of deep Convolutional Neural Network (CNN)[2], deep CNN has shown its powerful ability for tackling a variety of computer vision problems. Especially CNN-based trackers have already shown state-of-art performance, e.g., fully-convolutional Siamese-based[3] and multi-domain network-based[4], etc. A large number of positive training samples for tracking a specific object may not be available from the first frame. Therefore, the general routine of most CNN-based trackers is to pre-train their networks on a large range of datasets, and fine-tune the networks with the samples collected in the first frame from a Gaussian distribution. This popular dense sampling strategy causes the positive training data lack of diversity, while the negatives are relatively sufficient and various. This kind of class imbalance problem is usually solved by specifically designed loss function[5] or hard negative mining mechanism[4]. Recently adversarial learning-based data augmentation has been widely adopted in the tracking framework to enrich the training dataset, e.g., hard positives augmentation for handling occlusion problem[6] and hard negatives augmentation for long-term tracking[7]. Therefore, inspired by the above methods, in order to make our tracker more robust to pose variations and out-of-plane rotation, we propose a pose-guided hard positive samples augmentation method via generative adversarial network (GAN), and integrate it with the effective MDNet tracking framework. The main contributions of our work can be summarized as follows:

- We introduce a novel hard positive samples augmentation method for tracking articulated human being. We train the GAN model to generate diverse hard positive samples with different pose directions, these pose-guided positives increase the richness of samples while obeying the original target distribution.
- We combine our positive data augmentation mechanism with the MDNet tracking framework, where the generated positives are applied in the online fine-tune process and online updating when tracking failure has been detected.
- The extensive experiments demonstrate the effectiveness and improvement of our method compared to the original MDNet and other state-of-art methods in the public benchmark. Especially when our tracker faces the large pose variations and out-of-plane rotation challenges, the drifting problem can be alleviated effectively.

2. PROPOSED METHOD

2.1. The hard positives generation network

The aim of the pretraining process of most CNN-based trackers is to learn to represent the general and significant deep features, while the online fine-tune process is trying to adjust to identify the target in specific tracking sequence. During the

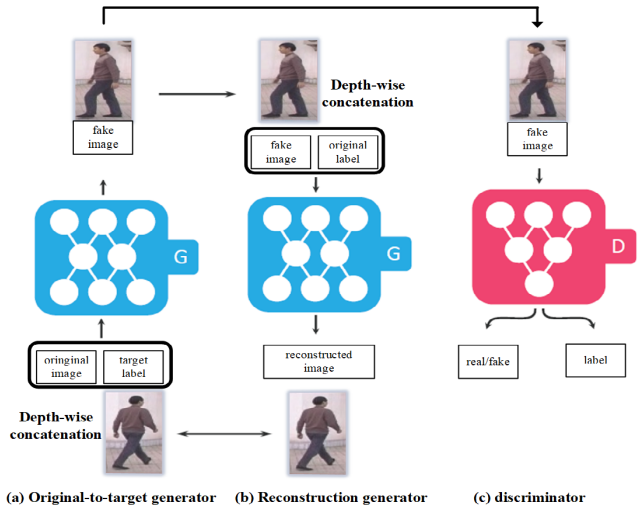

Fig. 1. Structure of hard positives generation networks. The generator network and discriminator network are denoted by G and D, respectively.

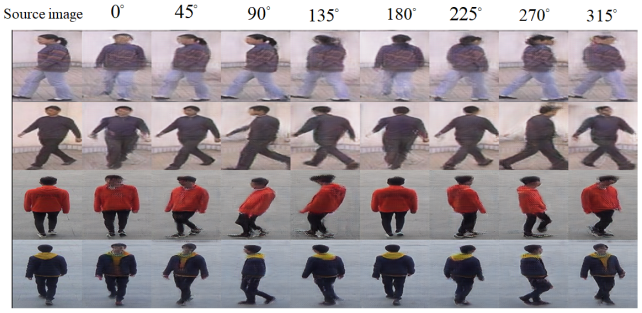

Fig. 2. The illustration of generated images of our StarGAN model. The first column is the original source image. From the second to the ninth column, corresponding to the generated images with eight different pose directions.

fine-tuning process, the amount of positive training samples in the first frame is extremely limited and less diverse compared with the collected negative samples. The tracker fine-tuned with these imbalanced samples is easily subject to the drifting problem, especially when the target undergoes large pose variations and out-of-plane rotation.

To deal with the above problem, we propose to employ generative model network to perform the positives data augmentation. Based on the observed fact of possible variations of a human target, we define a set of the typical poses corresponding to eight directions while a person is walking. We try to train our generative model to transform the given groudtruth image into these eight target domains, that is, the exact same target of different pose directions. The advantages of using these generated images to fine-tune the tracker can be summarized as two aspects. Firstly, compared to the original densely sampled positives which are monotonous and redundant, the generated hard positives samples are more diverse. Secondly, since the generated positives cover the possible variations of the target, the classifier would be able to generalize to recognize the target even with some unseen poses. A lot of impressive generative models have been proposed in recent years, such as variation auto-encoder(VAE)[8], PixelRNN[9] and generative adversarial network(GAN)[10]. In order to maintain the fidelity of generated samples theoretically, which means the distribution of generated images should be the same as the training images' distribution, we choose GAN as our generative model. Note that there are eight target domains in our conception, and basically traditional GAN-based image-to-image translation models have limited ability in handling multiple domains, because these methods are only capable of translating images between each pair of domains. Therefore we choose StarGAN[11] to perform our work because it can successfully learn multi-domain image-to-image translation using only one model. The effectiveness of this model lies in two points: Firstly, the generator(G) takes the depth-wise concatenation of image and the domain label as input. Secondly, the introduction of another reconstruction generator is to form a closed loop in conjunction with the generator network. The objective functions of optimizing generator(G) and discriminator(D) are shown as follows:

$$L_D = -L_{adv} + \lambda_{cls} L^r_{cls} \tag{1}$$

$$L_G = L_{adv} + \lambda_{cls} L^f_{cls} + \lambda_{rec} L_{rec} \tag{2}$$

where L_{adv} means the adversarial loss from the original GAN, L_{rec} denotes the reconstruction loss and L^r_{cls} and L^f_{cls} denote the domain classification loss corresponding to real and fake sample, respectively. For specific information please refer to the original paper[11].

The architecture of our hard positives generation network is shown in Fig. 1. In our method, we define eight walking pose directions, which correspond to eight labels from 0 to 7. In each training epoch, the generator takes the human image and the target poses label as input and generates a fake image of target pose direction. Then the fake image and the original pose label are given to the reconstruction generator, which is trying to reconstruct the image of the original pose direction. The whole generative networks are optimized in a closed loop according to Eqn.(1) and Eqn.(2). For the training process, we choose a human gait dataset which just includes 8 walking pose directions data. When the generative model is done, we input the original human image with its corresponding pose label and the target domain label, and the target image is generated from the reconstruction network. The details will be given in the below subsection. Fig. 2 illustrates some generated images from our StarGAN model in the training stage.

3. INTEGRATED TRACKING METHOD

In order to prove the validity of our proposed hard positives augmentation strategy, we embed it into the effective MDNet

framework. MDNet is an elegant micro neural network designed especially for tracking tasks[4]. It consists of shared layers and branches of domain-specific layers. The shared layers function as generic feature extractor while the domain-specific layers are especially initialized for each sequence to capture the domain-specific information. Similar to most CNN-based trackers, the positive samples used to fine-tune the network are collected around the target obeying Gaussian distribution, which are repeating and redundant.

Therefore, we propose to integrate StarGAN-based data augmentation into MDNet tracking framework. 1) **For offline training**, the StarGAN-based hard positives generator networks are pretrained based on human gait dataset. The multi-domain image-to-image translation can be realized. The MDNet tracker is pretrained as it usually does. 2) **For online tracking**, in the first frame we collect 100 base positive samples around the groundtruth with different scale sizes(from 1 to 1.4), overlap ratios(from 0.8 to 1), and aspect ratios(from 0.8 to 1.2). Each one of the samples is applied to the generation networks to obtain eight hard positives of different poses. As mentioned above, we need the initial pose label of each base sample. To accomplish this, we design a simple pose label estimation method wherein we compare each base sample with the centers of eight poses from training data and assign the label the same as its nearest neighbor's. The similarity is defined based on the L2 distance between two feature vectors. We extract the activation values of the third convolutional layer of MDNet and flatten it into a feature vector, denoted by f. Hence, the initial pose label of each base sample x can be estimated based on:

$$Label(x) = \arg\min_{i} \quad \|f(x^i) - f(x)\|_2 \tag{3}$$

where $f(x^i)|_{i=0}^{7}$ are the feature vectors corresponding to each pose center from StarGAN training dataset respectively. Therefore, during the initial fine-tune for each tracking sequence, we can generate a total of 800 hard positive samples with different pose directions, which enriches the dataset with more diversity to some extent. Fig. 3 shows the comparison between our generated hard positives and densely sampled positives. We can find that the generated samples are much more diverse and provide more information about the target. The tracker would be able to generalize to recognize the target even with some unseen poses, which is effective for alleviating the drifting problem caused by out-of-plane rotation. 3) **For online updating**, we add a failure updating mechanism. As the traditional MDNet does, once the potential failure has been detected, the failure update is conducted by generating 160 positive samples from top 20 convincing tracking boxes. Thus, the positive samples used for updating not only include the original positives sampled around the target but also, more importantly, include the ones generated by StarGAN generator. The proposed integrated tracking framework is shown in Fig. 4.

Fig. 3. Illustration of generated positive samples based on GAN model (a) vs. densely sampled positive samples (b).

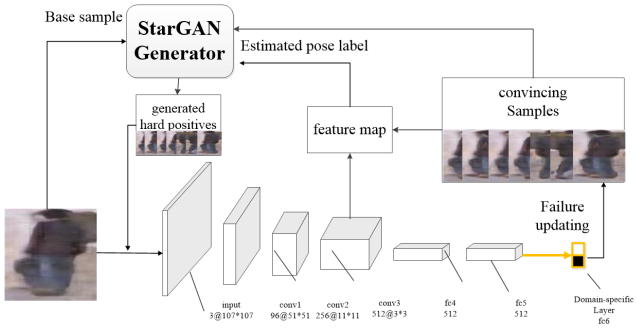

Fig. 4. Framework of the whole tracking process.

4. EXPERIMENT

In this section, we evaluate our method on public OTB benchmark with other state-of-art methods and also introduce comparison studies to further analyze our proposed method.

4.1. Experimental setups

In the pretrain of the GAN, we use the DatasetA of the CASIA database which has human walking clips with different pose directions[12]. Dataset A includes 19139 images of 20 persons, each person has 12 image sequences. For offline MDNet training, we use 58 training sequences collected from VOT2013[13], VOT2014[14] and VOT2015[15] excluding the test videos in OTB100[16]. For online test, we select 33 human tracking video sequences from OTB100 dataset.

The parameter configuration of our experiments is listed as follows. During pretraining the StarGAN network, the two hyper-parameters λ_{cls} and λ_{rec} are set to 1 and 10 in all of our experiments. And the training set image size is adjusted to 128*128, the learning rate of failure updating is set to 10 times larger than the initial one for fast adaption. The batch size is set to be 16 for all experiments. training iterations are set as 300000, other parameters configurations remain the

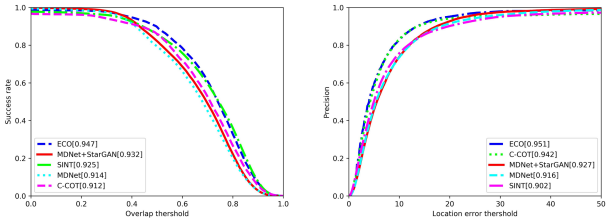

Fig. 5. Quantitative comparison results on OTB100 dataset. The numbers in the square brackets in the legend indicate the representative success plot at threshold 0.4 for success plots, and the representative precision at 20 pixels for precision plot.

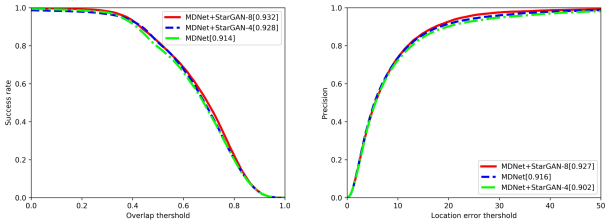

Fig. 6. Comparison results of different divisions of pose directions. MDNet+StarGAN-8 and the MDNet+StarGAN-4 denote the division of 8 and 4 poses respectively.

same with the original StarGan and MDNet.

4.2. Comparison results

In order to demonstrate the effectiveness of our proposed method, we perform a comparison experiment with other state-of-art methods, including SINT[3], ECO[17], C-COT[**?**] and the original MDNet[4]. To quantitatively evaluate these methods, we employ one-pass evaluation (OPE) on two metrics: center location error and bounding box overlap ratio [16]. Fig. 5 illustrates the success plots and precision plots based on the bounding box overlap ratio and center location error, respectively. From the figure, we can see our method, denoted by MDNet+StarGAN performs favorably against other trackers in both measures, and our improved algorithm outperforms the original MDNet. Considering data augmentation, our hard positive generation networks can improve the tracking performance. Fig. 7 presents the superiority of our method qualitatively in three challenging sequences: *Basketball*, *Human2* and *Human4*. Especially the target in *Human2* sequence undergoes continuously out-of-plane rotation problem, it can be seen our method illustrated by red rectangle can locate the human body robustly and accurately.

Earlier we empirically divide the human walking poses into eight directions, corresponding to eight pose labels. To evaluate the influence of the number of pose labels. We conduct another comparison experiment between 8 pose directions and 4 pose directions, i.e., 4 pose directions with rotation degree of 0, 90, 180 and 270. We don't consider a further fine division of pose directions because more direction labels would lead the generator networks to collapse and the boundaries among different pose directions would become obscure. The comparison result is shown in Fig. 6, it can be seen the division of eight directions performs better than the four one.

Fig. 7. Qualitative results of the proposed method on three challenging sequences.

5. CONCLUSION

In this paper, We have introduced a hard positive samples augmentation method for MDNet-based human tracking. We have shown the effectiveness of applying StarGAN model to generate more diverse and less redundant positive samples, which are integrated into online fine tune mechanism of MDNet to alleviate the positives deficiency. Compared with state-of-the-art trackers on public tracking benchmark: OTB. The comparison experimental results have verified our method has significantly boosted the tracking robustness of MDNet when confronting with large pose variations and out-of-plane rotation scenarios.

6. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No.61472063), the Fundamental Research Funds for the Central Universities (No. 2018J062) and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR No.201900014).

7. REFERENCES

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[2] Chong Sun, Huchuan Lu, and Ming-Hsuan Yang, "Learning spatial-aware regressions for visual tracking," in *IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 8962–8970.

[3] Ran Tao, Efstratios Gavves, and Arnold WM Smeulders, "Siamese instance search for tracking," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 1420–1429.

[4] Hyeonseob Nam and Bohyung Han, "Learning multi-domain convolutional neural networks for visual tracking," in *IEEE conference onComputer Vision and Pattern Recognition*, 2015, pp. 4293–4302.

[5] Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, Wangmeng Zuo, Chunhua Shen, Rynson WH Lau, and Ming-Hsuan Yang, "Vital: Visual tracking via adversarial learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8990–8999.

[6] Xiao Wang, Chenglong Li, Bin Luo, and Jin Tang, "Sint++: Robust visual tracking via adversarial positive instance generation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4864–4873.

[7] Kexin Chen, Xue Zhou, Wei Xiang, and Qidong Zhou, "Data augmentation using gan for multi-domain network-based human tracking.," in *IEEE Conference on Visual communication and Image Processing*, 2018.

[8] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[9] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu, "Pixel recurrent neural networks," *arXiv preprint arXiv:1601.06759*, 2016.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[11] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797.

[12] Shuai Zheng, Junge Zhang, Kaiqi Huang, Ran He, and Tieniu Tan, "Robust view transformation model for gait recognition," in *IEEE Conference on Image Processing*, 2011, pp. 2073–2076.

[13] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Luka Cehovin, Gustavo Fernandez, Tomas Vojir, Gustav Hager, Georg Nebehay, and Roman Pflugfelder, "The visual object tracking vot2015 challenge results," in *IEEE conference on computer vision workshops*, 2013, pp. 1–23.

[14] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Luka Cehovin, Gustavo Fernandez, Tomas Vojir, Gustav Hager, Georg Nebehay, and Roman Pflugfelder, "The visual object tracking vot2015 challenge results," in *IEEE conference on computer vision workshops*, 2014, pp. 1–23.

[15] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Luka Cehovin, Gustavo Fernandez, Tomas Vojir, Gustav Hager, Georg Nebehay, and Roman Pflugfelder, "The visual object tracking vot2015 challenge results," in *IEEE conference on computer vision workshops*, 2015, pp. 1–23.

[16] Yi Wu, Jongwoo Lim, and Ming Hsuan Yang, "Online object tracking: A benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2411–2418.

[17] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, Michael Felsberg, et al., "Eco: Efficient convolution operators for tracking.," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6931–6939.