Mask Guided Pedestrian Tracking from Single Point Initialization

Junan Zhu\textsuperscript{1,2}, Tao Chen\textsuperscript{*} and Jingtai Cao\textsuperscript{1}

\textsuperscript{1}Changchun Institute of Optics, Fine Mechanics, and Physics, Chinese Academy of Sciences, Changchun, Jilin, 130033, China

\textsuperscript{2}University of Chinese Academy of Sciences, Beijing 100049, China

*Corresponding author’s e-mail: chent@ciomp.ac.cn

Abstract. Pedestrian tracking has a wide range of applications, such as intelligent monitoring and unmanned driving. In pedestrian tracking system, object detection algorithms are generally used to provide the initial bounding boxes for object tracking algorithms. If the bounding box provided by detection algorithm is not accurate, it will seriously affect the performance of object tracking. To solve this problem, this paper proposes a mask guided pedestrian tracking algorithm based on Siamese network. Firstly, Mask R-CNN is introduced to acquire exemplar image and its mask. Secondly, a light-weight convolutional neural network (CNN) is applied as the backbone of proposed algorithm. Thirdly, a channel attention module is introduced to our proposed algorithm and integrated with the light-weight CNN. The feature maps of the exemplar image and its mask are adjusted by a channel attention model and fused to enhance the discrimination ability of exemplar branch. Finally, our proposed tracker is trained GOT-10k dataset and evaluated on object tracking benchmark (OTB), experimental results show the excellent performance of proposed algorithm compared with state-of-the-art trackers.

1. Introduction
Pedestrian tracking is a vital research topic in computer vision and has widely applications, such as intelligent surveillance and unmanned driving. A typical computer vision system is shown in figure 1, it consists of image preprocessing, object detection and object tracking. Object tracking is a middle-level task, the main task of object tracking is to obtain the position and movement information of the target in image sequences, which can provide the foundation for further high-level applications, such as posture analysis and behavior understanding. However, there are many challenges in pedestrian tracking, such as occlusions, deformations, illumination variations, rotations, etc. Therefore, pedestrian tracking is always the focus and difficulty in object tracking. Various tracking algorithms are proposed to solve these challenges, which are mainly divided into generating model \cite{1,2} and discriminating model \cite{3-6}. In discriminative model based tracking algorithms, the tracking algorithms based on correlation filtering \cite{3-6} and Siamese network \cite{7-9} are the research hotspot in recent years. The traditional correlation filter based tracking algorithms using hand-crafted features cannot achieve satisfactory tracking performance. With the development of deep learning, convolution features are introduced to the correlation filter based tracking algorithms, these algorithms using deep feature have achieved high tracking accuracy, but the feature extraction process for deep features is computationally complex, which cannot meet the real-time requirements. Recently, the tracking
algorithms based on Siamese network can achieve a good balance between speed and precision, which has attracted great attention of scholars.

In pedestrian tracking system, object detection algorithms are generally used to provide the initial bounding boxes for object tracking algorithms. If the bounding box provided by detection algorithm is not accurate, it will seriously affect the performance of object tracking. To solve this problem, we propose a mask guided pedestrian tracking algorithm based on Siamese network, the main work is as follows:

• A mask guided strategy is proposed to alleviate the bad effect of inaccurate bounding box provided by detection, Mask R-CNN [10] is introduced to offer exemplar image and its mask, a light-weight CNN is applied as the tracker’s backbone, the feature maps of the exemplar image and its mask image are fused to enhance the discrimination ability of exemplar branch.

• Most of the current tracking algorithms use bounding box as input, which cannot meet the needs of tracking algorithms to quickly select targets in some scenarios (e.g., intelligent monitoring). In this paper, Mask R-CNN is used to obtain candidate targets, a single point initialization is proposed to quickly select targets which has advantages over the traditional bounding boxes initialization in tracking methods.

• Experiments are conducted on pedestrian sequences in object tracking benchmark (OTB) [11], experimental results show the excellent performance of proposed algorithm compared with state-of-the-art trackers.

2. Proposed Approach

2.1. Overview

The framework of our proposed method is shown in figure 2, which consists of two parts, the object detection part based on Mask R-CNN and the object tracking part based on the Siamese network. The Siamese network is made up of exemplar and search branches. The exemplar branch contains a main branch and a mask branch; they each contain a light-weight CNN and a channel attention module. However, the search branch only uses a light-weight CNN. Finally, the two branches are joined by a convolution layer, and the feature map of the exemplar image is used as the convolution kernel. The tracking procedure can be formulated as a cross-correlated operation:

$$f(z, x) = \varphi(x) * (\alpha(\varphi(z)) + \lambda \alpha(\varphi(z_m))) + b$$

where $\varphi(*)$ stands for a light-weight CNN, $\alpha(*)$ stands for the channel attention module, $\lambda$ denotes the merge parameter, and $b$ stands for a bias. The framework of our method in figure 2 shows that two inputs are required, namely a point $P_0$ of target area in initial frame and a candidate search region $x$. Then, Mask R-CNN is used to detect the initial frame to acquire the exemplar image $z$ and its mask $z_m$. 

![Figure 1. A typical computer vision system.](image)
Finally, the response map that represents the similarity between the target image and the search image is obtained by cross-correlation calculation. Then, map the location of the maximal value of the response to the original image to acquire the target position in subsequent frames.

![Figure 2. The tracking framework of our proposed algorithm.](image)

2.2. Mask guided initialization

Mask R-CNN [10] proposed by Facebook AI research (FAIR) is a simple and flexible object instance segmentation network. Mask R-CNN is introduced to offer initialization information for our proposed algorithm.

In current object tracking algorithms, most tracking algorithms need to provide a bounding box of the target in the initial frame as input. In practical applications, bounding box is usually obtained by detection algorithm. However, the bounding box provided by the detection algorithm may not be accurate, this will lead to tracking performance degradation. In this paper, we propose a mask guided initialization to solve this problem. An example of mask guided initialization is shown in figure 3. First, Mask R-CNN is used to offer bounding boxes and masks for the first frame. Second, a point $P_0$ of the target area is input to quickly select targets. Third, the exemplar image and its mask can be calculated through the bounding box of the target. Finally, the feature maps of the exemplar image and its mask are adjusted by a channel attention module and fused to enhance the discrimination ability of exemplar branch, thereby mask guided initialization method can alleviate the bad effect of inaccurate bounding box provided by detection.

![Figure 3. An example of the mask guided initialization.](image)
2.3. **Convolutional network architecture**

A light-weight CNN [12] is introduced as the backbone of our proposed algorithm. Details of the CNN are listed in Table 1. It has thirteen convolutional layers and three max pooling layers. The small convolution kernels (3 × 3 filters and 1 × 1 filters) are used in this network, we double the number of channels after every max pooling operation, and place 1 × 1 filters between 3 × 3 convolutions. 1 × 1 filters are used many times compared with AlexNet [13]. 1 × 1 filters can compress the number of channels to reduce memory consumption. 1 × 1 filters can also mix cross-channel information and to increase the nonlinearity improve the generalization capability of this network. We also used batch normalization layers after convolutional layers to accelerate the training and regularize the model.

### Table 1. The architecture of this light-weight CNN.

| Layer | Kernel Size | Out and In Chan. | Stride | For Exemplar | For Search | Chan. |
|-------|-------------|------------------|--------|--------------|------------|-------|
| CONV1-BN | 3 × 3 | 32 × 3 | 1 | 133 × 133 | 261 × 261 | ×32 |
| CONV2-BN | 3 × 3 | 64 × 32 | 1 | 131 × 131 | 259 × 259 | ×64 |
| MP1 | 2 × 2 | | 2 | 65 × 65 | 129 × 129 | ×64 |
| CONV3-BN | 3 × 3 | 128 × 64 | 1 | 63 × 63 | 127 × 127 | ×128 |
| CONV4-BN | 1 × 1 | 64 × 128 | 1 | 63 × 63 | 127 × 127 | ×64 |
| CONV5-BN | 3 × 3 | 128 × 64 | 1 | 61 × 61 | 125 × 125 | ×128 |
| MP2 | 2 × 2 | | 2 | 30 × 30 | 62 × 62 | ×128 |
| CONV6-BN | 3 × 3 | 256 × 128 | 1 | 28 × 28 | 60 × 60 | ×256 |
| CONV7-BN | 1 × 1 | 128 × 256 | 1 | 28 × 28 | 60 × 60 | ×128 |
| CONV8-BN | 3 × 3 | 256 × 128 | 1 | 26 × 26 | 58 × 58 | ×256 |
| MP3 | 2 × 2 | | 2 | 13 × 13 | 29 × 29 | ×256 |
| CONV9-BN | 3 × 3 | 512 × 256 | 1 | 11 × 11 | 27 × 27 | ×512 |
| CONV10-BN | 1 × 1 | 256 × 512 | 1 | 11 × 11 | 27 × 27 | ×256 |
| CONV11-BN | 3 × 3 | 512 × 256 | 1 | 9 × 9 | 25 × 25 | ×512 |
| CONV12-BN | 1 × 1 | 256 × 512 | 1 | 9 × 9 | 25 × 25 | ×256 |
| CONV13 | 3 × 3 | 256 × 256 | 1 | 7 × 7 | 23 × 23 | ×256 |

2.4. **Channel attention module**

Channel attention has been widely used in various computer vision tasks and has been validated as effective for some computer vision tasks. Thus, we introduce the channel attention module in CBAM [14] to improve the tracking performance of our proposed algorithm. The framework of this module is shown in figure 4. Given the input feature map $U \in \mathbb{R}^{W \times H \times C}$, the final output feature map $V \in \mathbb{R}^{W \times H \times C}$ can be calculated by:

$$V = \text{Sigmoid}(\text{MLP}(\text{AvgPool}(U)) + \text{MLP}(\text{MaxPool}(U))) \odot U,$$

where $\odot$ stands for a channel-wise multiplication function.

![Figure 4. The framework of channel attention module in CBAM.](image-url)
2.5. Training the network

GOT-10k [15] is a large-scale dataset for object tracking, which contains more than 10,000 video segments and more than 1.5 million annotated bounding boxes, it covers 563 classes of objects and 87 classes of motion patterns. Therefore, GOT-10k is very suitable for training deep tracking algorithms. It is used as the training dataset to train our Siamese network. The training framework of our proposed algorithm is shown in figure 5, which doesn’t contain mask branch. The training process is similar to that in SiamFC [8], with image pairs as input. The details of the generation of training image pairs can be seen in SiamFC.

![Figure 5. The training framework of our proposed algorithm.](image)

We use logic loss to train our Siamese network. The logic loss function is also similar to that in SiamFC [8] and it can be formulated as:

\[
L(y, v) = \frac{1}{|D|} \sum_{u \in D} l(y[u], v[u]),
\]

where \(v\) stands for the returned score value of the inputted target-candidate image pair and \(y \in \{+1, -1\}\) is the ground truth label. Then, the parameters \(\theta\) of the Siamese network can be trained by minimizing the loss function:

\[
\arg \min_{\theta} \mathbb{E}_{z, x, y} L(y, f(z, x; \theta)),
\]

Stochastic gradient descent (SGD) is used to solve equation (4).

3. Experiments

In this section, we train and evaluate our proposed algorithm on an Intel i7-8700K CPU with a GTX 1080 Ti GPU card using Python and Pytorch. Experiments are conducted on pedestrian sequences in OTB dataset. Implementation details are presented in Section 3.1, and the experiments on OTB dataset are described in Sections 3.2. Ablation studies and further discussion are described in Section 3.4.

3.1. Implementation details

**Training.** We use Kaiming Normal Initialization [16] to initialize the parameters of our Siamese network. The batch size is set to 8. The learning rate exponentially decays from \(10^{-2}\) to \(10^{-5}\). The weight decay is set to 0.0005 and the momentum is set to 0.9.

**Tracking.** To simulate single point initialization, the input of the tracking algorithm is the center position of the annotated bounding box in the initial frame. The sizes of the search and exemplar image are set to \(263 \times 263\) and \(135 \times 135\). The merge parameter \(\lambda\) is set to 0.21. Three fixed scales \(\{0.96; 1; 1.04\}\) are used to better deal with scale variation and the scale factor is set to 0.59 when update the scale using linear interpolation. The cosine window function is used to penalize large displacements, and the window influence is set to 0.27.
3.2. Intermediate results of mask guided initialization

In order to intuitively display the intermediate results of mask guided initialization, the exemplar image and mask image obtained by Mask R-CNN are shown in figure 6. It can be seen from the figure that Mask R-CNN can well complete target detection and segmentation, which provides the foundation for subsequent target tracking.

![Exemplar image and mask image](image)

Figure 6. Exemplar image and mask image.

3.3. Quantitative evaluation

OTB is a widely used tracking performance evaluation benchmark in recent years. There are many pedestrian sequences in OTB dataset. 21 pedestrian tracking sequences are selected from the OTB dataset to evaluate our tracker. Our tracking algorithm is compared with two categories of tracking algorithms, namely, the tracking algorithm based on correlation filters (DSST [6], Staple [5], KCF [4], and SAMF [3]) and the tracking algorithm based on Siamese network (DCFNet [7], SiamFC [8], and CFNet [9]). In OTB dataset, one pass evaluation (OPE) is usually used to evaluate the performance of the tracking algorithm. Coverage rate (OR) and center positioning error (CLE) are used to obtain success and precision plots. The success and precision plots of OPE of the 8 tracking algorithms are shown in figure 7. Our proposed method has an AUC of 62.2% on the 21 pedestrian tracking sequences in the OTB dataset, and the overall tracking performance of our proposed method ranks first among the 8 tracking algorithms.

![Success and precision plots of OPE](image)

Figure 7. Success and precision plots of OPE on pedestrian sequences in OTB dataset.
3.4. Qualitative evaluation
To visually compare the tracking performance of different tracking algorithms, five representative image sequences (Human2, Human5, Human8, Jogging2 and Skating2.2) are selected from the OTB dataset for qualitative experiments. These video sequences contain almost all types of challenges in object tracking. The bounding boxes predicted by different tracking methods are shown in different colors. To better show the results of different methods, we only present bounding boxes of three best tracking algorithms along with our proposed tracking algorithm.

The tracking results are shown in Figure 8. In Human2 the main challenges are scale variation, deformation and occlusion; in Human5, the main challenges are scale variation and occlusion; in Human8, the main challenges are illumination change and scale variation; in Jogging2, the main challenge is occlusion; the main challenges in Skating2.2 are deformation and similar targets. It can be seen from figure 8, our tracker can track targets well and has high tracking performance when dealing with various challenges.

![Figure 8. Qualitative evaluation of the proposed method compared with SiamFC, Staple, and DCFNet on five pedestrian sequences (from top to bottom: Human2, Human5, Human8, Jogging2, and Skating2.2).](image)

3.5. Ablation studies
To verify effectiveness of the proposed mask guided initialization, our method is compared with other algorithms on the pedestrian tracking sequence. The experimental results are shown in figure 9. The Ours (detection) method represents no mask guided initialization, the input of this method is the detection results of Mask R-CNN, and the Ours (ground-truth) method represents no mask guided initialization, and the ground-truth manually annotated in the OTB dataset is used as the input of this tracking algorithm.
3.5.1 Comparison of using ground-truth and Mask R-CNN detection results as input

Ours (detection) and Ours (ground-truth) methods are almost the same, but the initial input bounding boxes of them is different. To compare the effects of different initial input bounding boxes in the first frame, we compare the tracking results of the two tracker. As shown in figure 9, the AUC score of Ours (detection) method has a greater drop compared with Ours (ground-truth). This is because the detection results provided by Mask R-CNN have error compared with the ground-truth. When the initial bounding box is inaccurate, there will be a consistent template drift problem in the following frames, resulting in poor tracking performance. Therefore, this paper proposes a mask guided initialization to alleviate this problem.

3.5.2 The role of mask guided initialization

To validate the effectiveness of mask guided initialization, we compare Ours and Ours (detection) method, the difference between them is that Ours (detection) method does not contain mask guided initialization. As shown in figure 9, the AUC score of Ours improves Ours (detection) method by 10.3%. It can be concluded from the experimental results that using mask guided initialization can alleviate the bad effect of inaccurate bounding box provided by detection and improve tracking performance.

4. Conclusion

In this paper, we propose a mask guided pedestrian tracking algorithm that combines Mask R-CNN and Siamese network. First, Mask R-CNN is introduced to acquire exemplar image and its mask. Second, a light-weight CNN is applied as the backbone of proposed algorithm. Third, a channel attention module is introduced and integrated with the light-weight CNN. The feature maps of the exemplar image and its mask are adjusted by a channel attention model and fused to enhance the discrimination ability of exemplar branch. The proposed tracking algorithm is evaluated on the pedestrian tracking sequence in the OTB dataset. Experimental results show that the mask guided initialization method can alleviate the bad effect of inaccurate bounding box provided by detection, and the pedestrian tracking algorithm with mask guided initialization can achieve good performance.

References

[1] Collins, R.T. (2003) Mean-shift blob tracking through scale space. In Proceedings of computer vision and pattern recognition, pp. 234-240.

[2] Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T. (2002) A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Transactions on signal processing, 50: 174-188.
[3] Li, Y., Zhu, J. (2014) A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration. In Proceedings of European Conference on Computer Vision, pp. 254-265.

[4] Henriques, J.F., Caseiro, R., Martins, P., Batista, J. (2015) High-Speed Tracking with Kernelized Correlation Filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37: 583-596.

[5] Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., Torr, P.H.S. (2016) Staple: Complementary learners for real-time tracking. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1401-1409.

[6] Danelljan, M., Hager, G., Khan, F.S., Felsberg, M. (2017) Discriminative Scale Space Tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39: 1561-1575.

[7] Wang, Q., Gao, J., Xing, J., Zhang, M., Hu, W. (2017) DCFNet: Discriminant correlation filters network for visual tracking. arXiv preprint arXiv:1704.04057.

[8] Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.S. (2016) Fully-Convolutional Siamese Networks for Object Tracking. In Proceedings of European Conference on Computer Vision, Cham; pp. 850-865.

[9] Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A., Torr, P.H.S. (2017) End-to-end representation learning for Correlation Filter based tracking. In Proceedings of International Conference on Computer Vision and Pattern Recognition; pp. 5000-5008.

[10] He, K., Gkioxari, G., Dollár, P., Girshick, R. (2017) Mask R-CNN. In Proceedings of the IEEE international conference on computer vision; pp. 2961-2969.

[11] Wu, Y., Lim, J., Yang, M.H. Object tracking benchmark. (2015) IEEE Transactions on Pattern Analysis and Machine Intelligence, 37: 1834-1848.

[12] Zhu, J., Chen, T., Cao, J. (2019) Siamese Network Using Adaptive Background Superposition Initialization for Real-Time Object Tracking. IEEE Access, 7: 119454-119464.

[13] Krizhevsky, A., Sutskever, I., Hinton, G.E. (2012) ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 2: 1097-1105.

[14] Woo, S., Park, J., Lee, J.Y., Kweon, I.S. (2018) CBAM: Convolutional block attention module. In Proceedings of European Conference on Computer Vision; pp. 3-19.

[15] Huang, L., Zhao, X., Huang, K. (2019) GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild. IEEE Transactions on Pattern Analysis and Machine Intelligence, 100190, 1-1.

[16] He, K., Zhang, X., Ren, S.; Sun, J. (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of Proceedings of the IEEE international conference on computer vision; pp. 1026-1034.