Improving Multiple Object Tracking with Single Object Tracking - Supplementary Material -

Linyu Zheng[1,2], Ming Tang[1], Yingying Chen[1,2,3], Guibo Zhu[1,2], Jinqiao Wang[1,2,3], Hanqing Lu[1,2]
[1] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] ObjectEye Inc., Beijing, China
{linyu.zheng,tangm,yingying.chen,gbzhu,jqwang,luhq}@nlpr.ia.ac.cn

Abstract

In this supplementary material, we first show the offline training pipeline of the proposed SOTMOT. Then, we provide some extra discussions of SOTMOT.

1. Offline Training Pipeline

We illustrate our whole offline training pipeline in Figure 1 for better understanding, which corresponds to the **Offline Training** of Section 3.2 and **Training Dataset** of Section 5.1 in the original paper.

2. Extra Discussions

2.1. Selection of SOT Model

There has been significant progress in deep convolutional neural networks (CNNs) based trackers in recent years. From a technical standpoint, existing state-of-the-art CNNs-based trackers mainly fall into two categories: Siamese-based [1] and discriminative model training-based [2]. The former treat tracking as a problem of similarity learning and has achieved state-of-the-art performance on many challenging benchmarks. However, these trackers are less robust to the heavy background clutters, especially to the interference of similar objects. In our SOTMOT, the main task of SOT model is to distinguish between different instances of the same category of objects. Therefore, Siamese-based SOT model is not the best choice for SOTMOT. Different from Siamese-based trackers, discriminative model training-based ones train discriminative model online, thus are robust to the interference of similar object. We select DCFST [3] because it is a state-of-the-art, flexible, and efficient discriminative model training-based tracker.

2.2. ID Switch of SOTMOT

It is seen from the Table 2 of the original paper that the ID switches (IDSW) of SOTMOT on MOT16 and MOT17 are remarkably higher than those of FairMOT and FairMOTv2, whereas lower than those of FairMOT and FairMOTv2 on MOT20. We think the reasons for this phenomenon are as follows.

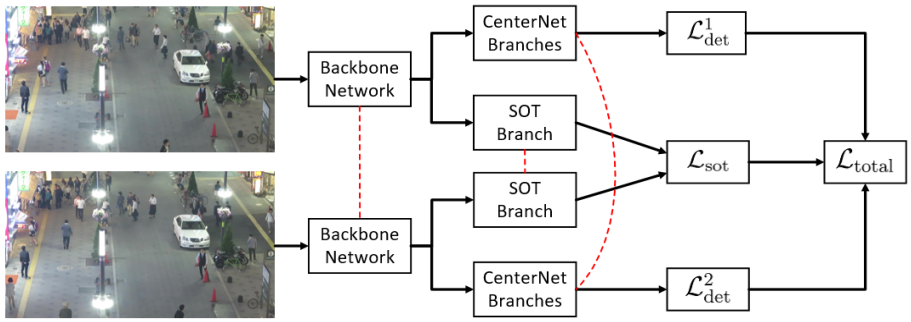
In SOTMOT, the SOT model of a new target (trajectory) is trained with the target samples in the fixed-size neighbourhood of the target. Therefore, if merely a few training samples can be obtained in the initial stage, the generalization ability of the SOT model will be weak until enough samples are collected in updating the SOT model (Eq. 6), leading to more IDSW. It is known that the density of targets on MOT20 is much larger than that on MOT16 and MOT17, resulting in enough training samples in the initial stage. Therefore, the IDSW of SOTMOT is obviously smaller than that of FairMOT on MOT20. The issue of insufficient training samples can be alleviated with adapted-size neighbourhood in the future version of our SOTMOT.

References

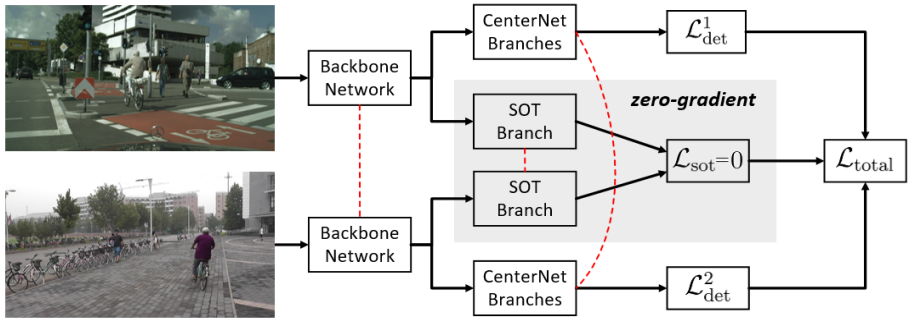
[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016. 1

[2] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):583–596, 2014. 1

[3] Linyu Zheng, Ming Tang, Yingying Chen, Jinqiao Wang, and Hanqing Lu. Learning feature embeddings for discriminant model based tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020. 1



(a) From a video snippet within the nearest 100 frames.



(b) From still images.

Sharing Parameters

Figure 1: The whole offline training framework for the input pair of images sampled (a) from a video snippet within the nearest 100 frames and (b) from still images.