Manipulating Template Pixels for Model Adaptation of Siamese Visual Tracking

Zhenbang Li, Bing Li, Jin Gao, Liang Li, and Weiming Hu

***Abstract*—In this letter, we show that the challenging model adaptation task in visual object tracking can be handled by simply manipulating pixels of the template image in Siamese networks. For a target that is not included in the offline training set, a slight modification of the template image pixels will improve the prediction result of the offline trained Siamese network. The popular adversarial example generation methods can be used to perform template pixel manipulation for model adaptation. Different from current template update methods, which aim to combine the target features from previous frames, we focus on the initial adaptation using target ground-truth in the first frame. Our model adaptation method is pluggable, in the sense that it does not alter the overall architecture of its base tracker. To our knowledge, this work is the first attempt to directly manipulating template pixels for model adaptation in Siamese-based trackers. Extensive experiments on recent benchmarks demonstrate that our method achieves better performance than some other state-of-the-art trackers. Our code is available at https://github.com/lizhenbang56/MTP.**

***Index Terms*—Model adaptation, siamese networks, visual tracking.**

I. Introduction

OBJECT tracking refers to the task of sequentially locating a specified moving object in a video, given only its initial state. Recently, Siamese networks [1], [2] have demonstrated a significant improvement in object tracking performances. Siamese trackers formulate the visual object tracking problem as learning cross-correlation similarities between a target template and a search region. Tracking is then performed by finding the target object from the search image region by computing the highest visual similarity. Despite its recent success, the learned similarity measure of the Siamese network is not necessarily reliable for objects that are not included in the offline training set, leading to poor generalization [3]. Several recent works aim to adapt the model to the current target appearance. For example, TADT [4] identifies the importance of each convolutional filter according to the back-propagated gradients and selects the target-aware features based on activations for representing the targets. However, the feature extractor of TADT is pre-trained on ImageNet [5], not on large-scale visual tracking datasets. This limits the representation ability of its features on the object tracking task. GradNet [6] exploits the discriminative information in gradients and updates the template in the Siamese network through feed-forward and backward operations. However, the extra sub-network increases the computational cost and is prone to overfitting. UpdateNet [7] learns to combine the target features from previous frames. However, it does not use ground truth information to adaptively adjust the template features of the first frame.

In this work, we show that the challenging model adaptation task in visual object tracking can be handled by simply manipulating pixels of the template image in Siamese networks. Given an object tracker, our algorithm modifies template pixels in only a few gradient-descent iterations using the target ground-truth in the first frame. For a target that is not included in the offline training set, we believe that a slight modification of the template image pixels can improve the prediction result of the offline trained Siamese network. We use the adversarial example generation method to achieve this, because it is commonly used to slightly modify the input image, and thereby impose an impact on the prediction result of the network. We depart from the purpose of adversarial sample generating in that the latter is aimed to make the prediction of the network worse, while we hope the prediction of the Siamese network is better. The proposed model adaptation method can be integrated with varieties of Siamese trackers like SiamFC++ [9]. Note that the parameters of the Siamese network remain intact to preserve the generative ability of offline-trained embedding space. We perform comprehensive experiments on 4 tracking benchmarks: VOT2018 [10], TrackingNet [11], GOT-10 k [8], and OTB2015 [12]. Our approach achieves state-of-the-art results while running at over 80 FPS (see Fig. 1).

II. Proposed Algorithm

In this section, we present a new model adaptation approach for Siamese trackers via directly manipulating template pixels.

Manuscript received July 6, 2020; revised September 15, 2020; accepted September 16, 2020. Date of publication September 21, 2020; date of current version October 6, 2020. This work was supported by the National Key R&D Program of China under Grant 2018AAA0102802, Grant 2018AAA0102803, and Grant 2018AAA0102800; in part by the NSFC-General Technology Collaborative Fund for Basic Research under Grant U1636218; in part by the Natural Science Foundation of China under Grant 61751212, Grant 61721004, Grant 61772225, and Grant 61972394; in part by the Key Research Program of Frontier Sciences, CAS, under Grant QYZDJ-SSW-JSC040; and in part by the National Natural Science Foundation of Guangdong under Grant 2018B030311046. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joao Paulo Papa. *(Corresponding author: Jin Gao.)*

Zhenbang Li, Bing Li, and Jin Gao are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhenbang.li@nlpr.ia.ac.cn; bli@nlpr.ia.ac.cn; jin.gao@nlpr.ia.ac.cn).

Liang Li is with the Brain Science Center, Beijing Institute of Basic Medical Sciences, Beijing 100850, China (e-mail: liang.li.brain@aliyun.com).

Weiming Hu is with the CAS Center for Excellence in Brain Science and Intelligence Technology, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: wmhu@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/LSP.2020.3025406