A Long-Term Discriminative Single Shot Segmentation Tracker

Benjamin Džubur[1], **Alan Lukežič**[1], **Matej Kristan**[1]

[1] *University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113*
E-mail: bd5830@student.uni-lj.si, {alan.lukezic, matej.kristan}@fri.uni-lj.si

Abstract

State-of-the-art long-term visual object tracking methods are limited to predict target position as an axis-aligned bounding box. Segmentation-based trackers exist, however they do not address long-term disappearances of the target. We propose a long-term discriminative single shot segmentation tracker – D3SLT, which addresses the above shortcomings. The previously developed short-term D3S tracker is upgraded with a global re-detection module, based on an image-wide discriminative correlation filter response and Gaussian motion model. An online learned confidence estimation module is employed for robust estimation target disappearance. Additional backtracking module enables recovery from tracking failures and further improves tracking performance. D3SLT performs close to the state-of-the-art long-term trackers on the bounding box based VOT-LT2021 Challenge, achieving F-score of 0.667, while additionally outputting segmentation masks.

1 Introduction

Visual object tracking is one of the main computer vision problems, in which the task of the tracking algorithm is to continuously localize the target in a video. Target is specified by a single supervised example at the beginning of the video. Visual object tracking challenges can be roughly divided into two main groups: (i) short-term tracking, where target is always visible and (ii) long-term tracking, where target can disappear from the field-of-view or becomes fully occluded. Long-term trackers require a re-detection capability to re-localize the target after target disappearance.

Long-term trackers in general consist of three main components: a short-term tracker, detector and interaction between the first two components. Lukežič et al. [7] use adaptive thresholding of the discriminative correlation filter (DCF) response to detect target absence and employ multiple DCFs, updated at various time scales, in combination with a simple motion model to re-detect the target. On the other hand, many successful long-term trackers from the past two years (LTMU_B [2], CLGS, Megtrack [11]) use MDNet [13] or some similar specialized online verifier for absence detection. For re-detection, the most successful long-term trackers either use a region proposal network (RPN) to regress the target within some region in combination with a simple sliding window search (e.g. LTMU_B) or by using a deep module for global target localization [5]. Interestingly, winner of the VOT-LT2021 challenge mlpLT [3] combines two short-term trackers [1, 15] by running them both simultaneously and then giving the output of both to a verifier [13]. No additional global re-detection module is used.

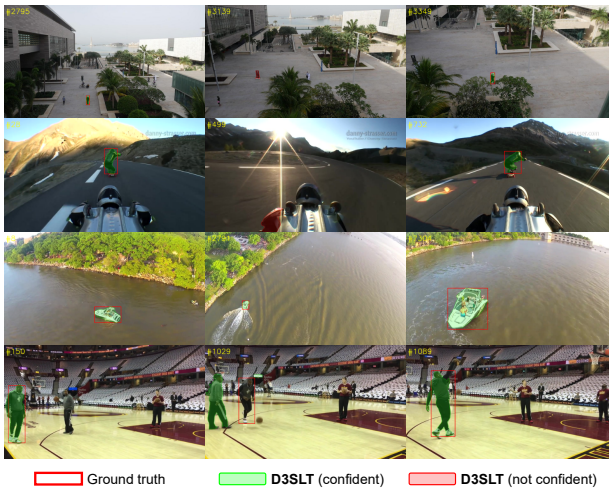

Figure 1: Tracking results of the proposed D3SLT, which successfully detects the absence of the target and re-detects it when it re-appears, which is demonstrated in the first three rows. The tracker successfully recovers after re-detection of the wrong target due to the backtracking module (last row).

State-of-the-art long-term trackers report target position as an axis-aligned bounding box, which is just an approximation of a target position. This approximation is often not very accurate, e.g., when tracking an elongated rotated object, or when target is significantly deforming, since a lot of pixels within the predicted bounding box are actually background. In these situations a segmentation mask is much more accurate prediction. Development of segmentation-based trackers has been recently popularized by the VOT Challenge [11, 12], where ground-truth in a short-term sub-challenge is given as segmentation masks. Wang et al. [14] proposed one of the first segmentation-based short-term trackers by extending a siamese tracking framework with segmentation mask pre-

diction, while Yan et al. [16] presented a segmentation-based framework which can be applied to any bounding-box tracker. Recently, Lukežič et al. proposed a discriminative single-shot segmentation tracker (D3S [8]), which combines a deep discriminative correlation filter (DCF) for robust target estimation and a feature-matching module for per-pixel segmentation. The D3S tracker has demonstrated robust tracking performance due to discriminative formulation as well as high prediction accuracy due to segmentation, while it lacks long-term tracking capabilities. In particular, it fails when target disappears for a longer time period and is not able to recover it after it re-appears in the image.

In this work we present a long-term tracker named D3SLT as a primary contribution, which extends an existing short-term tracker D3S [8] with long-term tracking capabilities. We design a detector based on deep DCF in D3S capable of image-wide target re-detection. A backtracking mechanism is proposed as a secondary contribution, which is used to recover after potential re-detection of a wrong target. The proposed tracker achieves competitive results to the state-of-the-art long-term trackers on VOT-LT2021, while reporting target position as a segmentation mask.

2 Long-term Segmentation Tracker

In this section we describe the proposed long-term discriminative single shot segmentation tracker (D3SLT). The tracker extends the existing short-term tracker D3S [9], described in Section 2.1, with long-term tracking capabilities. A crucial component D3SLT is target re-detection module (Section 2.3), which is activated when target is localized with low confidence (Section 2.2). A potential re-detection of a wrong target is resolved by a backtracking module, described in Section 2.4.

2.1 Short-term tracker

The short-term component is used to track the target in consecutive frames. We use the short-term D3S [9] tracker, which has demonstrated a robust tracking performance as well as a high accuracy of the predicted target segmentation. Target localization in D3S is performed within the search region which is approximately four times larger than target size. High robustness and accuracy are obtained by combining two visual models – one is adaptive and highly discriminative, but geometrically constrained to an Euclidean motion (GEM), while the other is invariant to broad range of transformation (GIM) which provides an approximate segmentation. The outputs of GIM and GEM are combined in the refinement pathway, which produces an accurate segmentation mask of the tracked target. Finally, the scale estimation module is used for robust estimation of the target size. We refer the reader to the original publication [9] for more details about the short-term tracker.

2.2 Confidence estimation module

This module predicts confidence of the current target location prediction. Ideally, high confidence value is outputed when the predicted position highly overlaps with the actual position of the target and low value when target prediction does not cover the target well, e.g., most of the object is occluded, or the object has left the camera view. Confidence score is predicted by classifying the image region cropped around the predicted target position using an online learned classifier. The classifier, called also the verifier, is trained online to distinguish between positive and negative samples using training methodology presented in [6].

At initialization, 100 positive training samples are generated by perturbing a bounding box so that it overlaps with the initial bounding box at least 70%, while 200 negative examples are sampled around the target with overlap less than 50%. During confident tracking, both sets of samples are updated every fifth frame. In positive set, 100 samples are added by randomly perturbing bounding box, so that they overlap with the predicted bounding box by more than 70%. Negative set is updated by adding 100 negative samples collected over the entire image. The re-detection module (Section 2.3) is applied on the image to indentify the most similar regions compared to the tracked target. The top 30 local maxima in the global DCF response are taken as candidate positions. The negative samples are constructed by cropping patches of the same size as the target, centered at the candidate positions. Additional translation, scale and aspect jittering is used when generating the samples. Sets of positive and negative samples are limited to 2000 samples, each. When the limit is reached, the oldest samples are replaced first. After adding new samples to the positive and negative sample sets, the classifier is re-trained.

Low confidence (i.e., lower than a pre-defined threshold τ) indicates that the target is not tracked anymore, thus a re-detection module (Section 2.3) is activated and the verifier as well as the short-term tracker are not updated anymore. Observing K consecutive positive scores deactivates the re-detection module and triggers the backtracking module (Section 2.4). If the backtrack is successful, tracking is resumed with the short-term tracker only.

2.3 Re-detection module

Global (image-wide) target re-detection is performed using a deep discriminative correlation filter (DCF) from the GEM module in D3S, which is correlated over the entire image to obtain a global correlation response $\mathbf{R}^{\mathrm{G}}$. After the target disappears, it is not likely that it will re-appear immediately on the position significantly far away from the last confident target position, denoted as x_c. Thus we introduce a motion prior $\pi(\mathrm{x})$, formulated as a random walk dynamic model, which models the likelihood of target position $\mathrm{x} = [x, y]^T$ by a Gaussian $\pi(\mathrm{x}) = \mathcal{N}(\mathrm{x}_c, \Sigma)$. Note that Σ is a diagonal covariance matrix defined as

$$\Sigma = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}, \sigma_x = \mathrm{x}_w \alpha_s^{\Delta t}, \;\; \sigma_y = \mathrm{x}_h \alpha_s^{\Delta t},$$

where x_w and x_h denote the width and height of the target respectively, α_s is the scale increase parameter and

Δ_t denotes the number of frames since the last confident estimation. Target location estimated by the re-detection module is defined as the position of the maximum in the response $\mathbf{R}^{\mathrm{G}} \odot \pi(\mathrm{x})$, where $\odot$ is the pointwise product.

2.4 Backtracking module

Target re-detection process may fail, leading to tracking of the wrong target. To address these situations we design the backtracking module, which tracks the selected target back in time to resolve possible target switches. Suppose that tracker successfully re-detected the target at frame t. A new instance of the short-term tracker (Section 2.1) is initialized in the frame I_t and is used to track previous frames from the set of past frames $\Gamma = \{I_{t-1}, ..., I_{t-n}\}$, where n is the maximum number of past frames which can be used. The backtracking process can end before reaching n past frames in the following cases: (i) the maximum in GEM correlation response drops under the threshold θ_{GEM}, or (ii) a confident tracking result was obtained before on the frame I_{t-j}. The first case indicates that the tracked target is not visible anymore. We cannot assume whether this is the true target or not, therefore we continue with short-term tracking in frame I_{t+1}. In the second case (ii) or if all n frames have been backtracked, we compute the probability density at the backtracked target position $\pi(\mathrm{x}_{t-j})$ based on the motion model and the last confident target position x_c, as defined in Section 2.3. This density is subjected to a threshold θ_{MM} to decide whether the backtracked target corresponds to the true target (in which case we switch to short-term tracking in I_{t+1}) or a distractor (we continue with global re-detection in I_{t+1}). Finally, the confidence estimation module is updated (Section 2.2) accordingly based on features extracted during backtracking.

3 Experiments

In this section, we present the experimental results of the proposed D3SLT on the VOT-LT2021 dataset. In Section 3.1, we describe implementation details of our tracker. In Section 3.2 we compare our tracker to the state-of-the-art long-term trackers and in Section 3.3 we report the results of the ablation study.

3.1 Implementation details

We use the pre-trained version of the short-term D3S tracker [9] with the ResNet50 [4] backbone. The following parameters were determined manually, based on our preliminary experiments and remained fixed in all experiments. The motion model parameter α_s is set to 1.05, positive and negative samples in the verifier (Section 2.2) are updated every five successfully tracked frames and the classifier is updated using SGD with learning rate 0.0003 for 15 iterations. The decision threshold in the verifier τ is set to 0.5, while K is set to 3. In the backtracking module the maximum number of past frames n is set to 260, the motion model uncertainty threshold θ_{MM} to 0.26 and θ_{GEM} to 0.16.

3.2 Long-term tracking performance

The proposed D3SLT tracker is evaluated on the VOT-LT2021 benchmark [12]. The dataset consists of 50 long video sequences, in which target objects disappear and reappear frequently. Note that axis-aligned bounding boxes are required by the evaluation protocol, thus the predicted segmentation masks are transformed into bounding boxes by generating the smallest axis-aligned bounding box which contains the whole mask. The reported confidence score at each frame is defined as the output of the verifier on the patch, contained in the predicted bounding box. Tracking performance is evaluated using tracking precision (Pr) and tracking recall measures (Re) computed under a set of confidence thresholds. The primary performance measure to rank the trackers is F-score, defined as $F = \frac{2PrRe}{Pr+Re}$. The reader is referred to the original publication [10] for a detailed description of the performance evaluation measures and protocol.

Tracker	Pr	Re	F	Award
mlpLT [3]	**0.741**	**0.729**	**0.735**	① (2021)
STARK_LT	0.721	0.725	0.723	② (2021)
LT_DSE	0.715	0.677	0.695	① (2019, 2020)
LTMU_B [2]	0.701	0.681	0.691	② (2020)
D3SLT	0.669	0.666	0.667	
D3S [9]	0.452	0.465	0.459	
FuCoLoT [7]	0.507	0.346	0.411	

Table 1: State-of-the-art comparison on the VOT-LT2021.

The proposed D3SLT is compared to the state-of-the-art long-term trackers and the results are presented in Table 1. We observe that the state-of-the-art trackers are slightly more robust (i.e., higher recall), however these trackers can not produce a segmentation mask. When compared to the original D3S tracker, our tracker improves F-score by more than 20pp. The results on selected sequences are visualized in Figure 1. While the tracker segments the targets nicely, the predicted axis-aligned bounding boxes which are lined up with the edges of the mask, might not be considered precise when compared to the ground-truth boxes from the dataset. Additionally, the mask sometimes captures noisy background around the edge of the target, further inflating the predicted box. A more intelligent adaptive prediction of the bounding box based on the mask might further improve precision and thus F-score.

The proposed tracker runs at 4.7 FPS, measured on a system with a Ryzen 3700x CPU and a single Nvidia RTX3060Ti GPU. Most of the slow-down compared to the baseline short-term tracker [9], which runs at 9 FPS, is due to the DCF response computation for image-wide target re-detection and frequent backtracking.

3.3 Ablation study

An ablation study was performed to analyze contributions of individual D3SLT components. The variants of the D3SLT without the following components were considered: (i) without the global re-detection module ($\overline{GRE}$); (ii) without the Gaussian motion model in re-detection

($\overline{MM}$); (iii) without the backtracking module ($\overline{BT}$); and (iv) instead of the verifier ($\overline{VER}$) a confidence score based on the correlation response quality was computed, similarly as in [7]. Results of the ablation study are shown in Table 2.

Variant	Pr	Re	F
D3SLT	0.669	0.666	0.667
$\overline{GRE}$	0.654	0.515	0.576
$\overline{MM}$	0.653	0.648	0.651
$\overline{BT}$	0.654	0.642	0.648
$\overline{VER}$	0.649	0.631	0.640

Table 2: Ablation study of D3SLT on the VOT-LT2021 dataset.

Removing the global re-detection module ($\overline{GRE}$) results in over 9pp drop of the F-score, primarily due to the 13pp lower recall. This is expected as the target may disappear through one edge of the frame and re-appear through another. In this case, it is less likely to be re-detected using only the smaller local search region of the short-term tracker.

The motion model provides additional inductive bias in our tracker, reducing the chance of tracking a faraway distractor, when the target is briefly lost due to e.g. short-term occlusion. Removing the motion model ($\overline{MM}$) results in 1.6pp performance drop in F-score comparing to the proposed D3SLT.

When removing the verifier ($\overline{VER}$) and use the correlation response to compute the confidence score, similar as in [7], we observe a drop in F-score of 2.7pp. This indicates that the alternative scoring mechanism is highly correlated with the proposed verifier, though slightly less robust.

When backtracking is disabled ($\overline{BT}$) 1.9pp lower F-score is achieved, compared to the original D3SLT. Clearly, the module helps to ensure that D3SLT does not begin to confidently track a distractor when it loses the true target.

4 Conclusion

We introduced a long-term discriminative single shot segmentation tracker – D3SLT. A recent D3S tracker [9] was used as a short-term component and combined with an image-wide re-detection mechanism based on deep discriminative correlation filter. A separate confidence estimation module was designed for robust estimation of the localization confidence and a backtracking module was developed to recover after potential wrong re-detections.

The tracker achieved respectable performance on the VOT-LT2021 challenge with F-score 0.667. While current state-of-the-art methods perform noticably better for the task of object tracking via bounding boxes, they do not provide accurate segmentation masks of the tracked object. To further improve the D3SLT performance, our future research will focus on bringing robust object-agnostic detectors from the literature to the proposed framework.

Acknowledgement

This work was supported by the ARRS program P2-0214 and project J2-2506.

References

[1] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *ICCV*, 2021.

[2] Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. High-performance long-term tracking with meta-updater. In *CVPR*, 2020.

[3] Matteo Dunnhofer, Kristian Simonato, and Christian Micheloni. Combining complementary trackers for enhanced long-term visual object tracking. *Image and Vision Computing*, 122:104448, 2022.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[5] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Globaltrack: A simple and strong baseline for long-term tracking. In *AAAI*, volume 34, pages 11037–11044, 2020.

[6] Ilchae Jung, Jeany Son, Mooyeol Baek, and Bohyung Han. Real-time mdnet. In *ECCV*, 2018.

[7] Alan Lukežič, Luka Čehovin Zajc, Tomáš Vojíř, Jiří Matas, and Matej Kristan. Fucolot - a fully-correlational long-term tracker. In *ACCV*, 2018.

[8] Alan Lukežič, Jiří Matas, and Matej Kristan. D3s - a discriminative single shot segmentation tracker. In *CVPR*, 2020.

[9] Alan Lukežič, Jiří Matas, and Matej Kristan. A discriminative single-shot segmentation network for visual object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[10] Alan Lukežič, Luka Čehovin Zajc, Tomáš Vojíř, Jiří Matas, and Matej Kristan. Performance evaluation methodology for long-term single-object tracking. *IEEE Transactions on Cybernetics*, 51(12):6305–6318, 2021.

[11] Matej Kristan, et al. The eighth visual object tracking VOT2020 challenge results. In *European Conference on Computer Vision Workshops*, 2020.

[12] Matej Kristan, et al. The ninth visual object tracking VOT2021 challenge results. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 2021.

[13] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *CVPR*, 2016.

[14] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, 2019.

[15] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *ICCV*, 2021.

[16] Bin Yan, Xinyu Zhang, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Alpha-refine: Boosting tracking performance by precise bounding box estimation. In *CVPR*, pages 5289–5298, 2021.