Transformer Meets Tracker: Exploiting Temporal Context for Robust Visual Tracking

Seungmin Lee (profile2697@gmail.com; 2020-20866),
Dept. of Electrical and Computer Engineering, Seoul National University

1. Introduction and Motivation

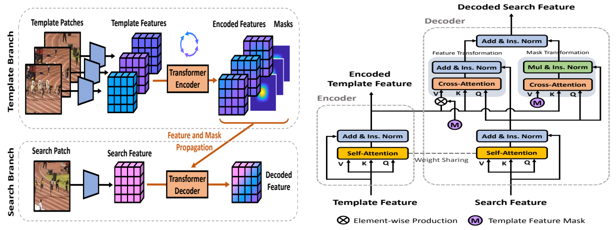
To successfully tracking objects in videos, we need to capture valuable temporal information existing across frames. Unfortunately, most existing methods are based on per-frame detection, limiting the ability to model temporal relationships between frames. Moreover, those methods assume that tracking targets move smoothly while targets in real-world videos move noisily and are blurred or occluded. This paper claims that a siamese-like tracking pipeline combined with a transformer can effectively mitigate the problems by modeling temporal relationships.

2. Methods

The proposed pipeline consists of two branches: The template branch and the search branch. The template branch takes multiple target templates and extracts features and corresponding masks used to detect the target in the later pipeline such as DCF [2, 3] or Siamese [1]. In this branch, the transformer encoder captures the temporal relationships between each template and, thus, reinforces desirable features and suppresses noise across templates. On the other hand, the search branch generates decoded features that highlight the potential target areas using search frames and the features encoded by the template branch (Fig. 1).

2.1. The Template Branch: Transformer Encoder

At timestamp 0, the encoder transformer takes the concatenation of templates created by data augmentations. The encoder transformer has similar architecture to the original one [4] (Fig. 1). After the first timestamp, search frames that show high accuracy are used as pseudo templates. The process models the temporal relationship. The transformed output feature has the same shape as the concatenated templated fed as the input. The pipeline propagates the transformed feature to the search branch's decoder transformer.



(a) The Proposed Siamese-like Pipeline (b) The Encoder-Decoder Transformer used in the branches

Figure 1. (a) The Proposed Siamese-like Pipeline, (b) The Encoder-Decoder Transformer used in the branches.

2.2. The Search Branch: Transformer Decoder

The decoder takes the search frame features and the extracted template features. The decoder first transforms the search features using the self-attention that is shared with the encoder. Then, the decoder inputs the transformed search features and the template features to cross-attention to generate decoded features. The cross-attention introduces cross-frame relationships to the decoded features.

2.3. DCF or Siamese Pipeline

Given the extracted template features and decoded response map, we generate predictions using the existing DCF [2, 3] and Siamese [1], which basically convolving the decoded features using the template features and masks.

2.4. Results

The proposed method beats the existing state-of-the-art techniques with a comfortable margin. Furthermore, the provided qualitative results show that the proposed method refines the response maps. Finally, the ablation experiments show that the transformer helps to improve tracking performance.

3. Personal Note

The paper is challenging to understand because the description skips or omits some vital information such as inputs and outputs or the shape of ground truths. Moreover, I believe that section 3 is helpful to only those already familiar with this task. Nevertheless, the proposed method, which tries to capture the temporal information, seems compelling and insightful.

References

[1] L. Bertinetto, et al. Fully-convolutional siamese networks for object tracking. In *ECCV Workshops*, 2016. 1

[2] G. Bhat, et al. Learning discriminative model prediction for tracking. In *ICCV*, 2019. 1

[3] M. Danelljan, et al. Atom: Accurate tracking by overlap maximization. In *CVPR*, 2019. 1

[4] A. Vaswani, et al. Attention is all you need. In *NeurIPS*, 2017. 1