Target Focused Shallow Transformer Framework for Efficient Visual Tracking

Md Maklachur Rahman

Department of Computer Science and Engineering, Texas A&M University, College Station, TX, USA
maklachur@tamu.edu

Abstract

Template learning transformer trackers have achieved significant performance improvement recently due to the long-dependency learning using the self-attention (SA) mechanism. However, the typical SA mechanisms in transformers adopt a less discriminative design approach which is inadequate for focusing on the most important target information during tracking. Therefore, existing trackers are easily distracted by background information and have constraints in handling tracking challenges. The focus of our research is to develop a target-focused discriminative shallow transformer tracking framework that can learn to distinguish the target from the background and enable accurate tracking with fast speed. Extensive experiments will be performed on several popular benchmarks, including OTB100, UAV123, GOT10k, LaSOT, and TrackingNet, to demonstrate the effectiveness of the proposed framework.

Introduction and Motivation

Visual object tracking (VOT) is a fundamental yet challenging task in computer vision, aiming to estimate the future positions of a target given its initial position (Bertinetto et al. 2016). Due to its wide range of applications, from autonomous vehicles, and intelligent surveillance to behavior analysis, VOT has gained much attention in the computer vision community. However, although VOT performance has substantially improved recently, it still faces challenges such as background clutter, occlusion, fast motion, motion blur, deformation, scale variation, and illumination variations.

Deep learning and discriminative correlation filter-based trackers (Nam and Han 2015) were introduced to overcome such challenges, but they failed to maintain another most important aspect of the tracker's evaluation (high tracking speed). As a result, these trackers are not applicable to real-time vision applications. To improve overall tracking performance in accuracy and speed, (Bertinetto et al. 2016) proposed a fully convolutional siamese tracker called SiamFC. Usually, it produces a response map at the end of parallel branches using a simple cross-correlation operation. The response maps use to predict the target location in the subsequent frames. Several follow-up works (Rahman, Fiaz, and Jung 2020; Chen et al. 2021; Fu et al. 2022) proposed based on underlying SiamFC due to its efficient computational power. Although SiamFC-based trackers operate in real-time, they face accuracy challenges due to weak target representation and limited discriminative ability.

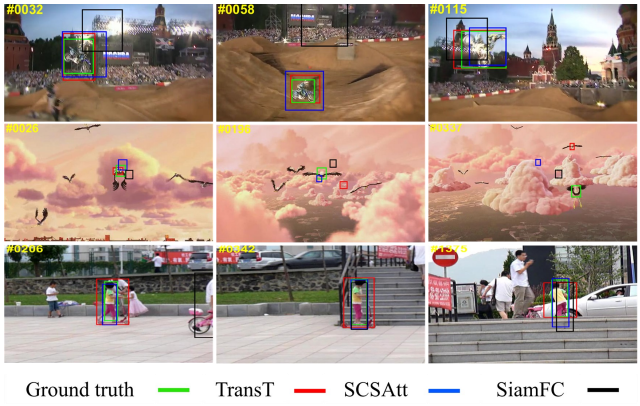

Figure 1: The qualitative comparison shows the weakness of the existing trackers on image sequences (motorRolling, bird1, and girl2) from the OTB100 benchmark.

Recently, integrating vision transformer within the siamese framework achieved significant performance improvement in tracking (Chen et al. 2021). The attention blocks (Rahman, Fiaz, and Jung 2020) are the most important mechanisms in a conventional Transformer tracking architecture. The main limitation of the SA mechanism in the Transformer architecture is its inability to effectively model spatial relationships between positions in the input sequence. SA treats all positions in the sequence equally and does not explicitly encode their relative positions or distances, which can result in the mixing of foreground and background information and create blurred edge regions. This limitation can negatively impact tracking performance, requiring accurate foreground and background information separations. Performing SA in a transformer is also highly computationally expensive, which is equivalent to the squared of input tokens (Chen et al. 2021; Fu et al. 2022). As a result, training and testing for transformer-based frameworks are time-consuming and negatively affect fast-tracking. Figure 1 showcases a qualitative comparison,

highlighting the shortcomings of existing trackers against ground-truth boxes. To overcome such issues, our investigation summarizes as follows:

- We developed a target-focused framework that can learn crucial target features instead of being distracted by non-target information. It improves the network's discriminative power, which boosts tracker accuracy.
- We proposed a shallow Transformer architecture by reducing the Transformer's seriality (increasing network parameters overhead) to speed up tracking objects.
- We conducted extensive experiments to demonstrate the effectiveness of the proposed tracker on two popular and challenging benchmarks: OTB100 and UAV123.

Methodology and Accomplishments

The proposed tracking framework consists of three main components: a siamese-based backbone architecture for feature extraction, a target-focused shallow Transformer that learns to focus important information, and a prediction head for accurate bounding box estimation to locate the target.

The siamese backbone network consists of two fully-convolutional networks that share the same set of parameters (Bertinetto et al. 2016). One network takes the initial frame as the target image, and the other takes the rest as candidate frames. The feature maps extracted by the two networks are concatenated and passed through a non-linear layer to obtain the final feature representation. The encoder of the transformer network takes the feature maps obtained from the backbone network as input and applies self-attention to the feature maps. The decoder takes the output of the encoder and generates a prediction for the target's bounding box after passing through the prediction head. To emphasize crucial target features, we have altered the self-attention mechanism commonly found in vision transformers. Specifically, we implement a Gaussian map-inspired 2D binary mask to pinpoint the target object within feature maps, filtering out non-target features. This reduces redundancy in the fully-connected maps created by the self-attention process, thereby boosting the model's discriminative capacity and effectively enhancing tracker performance.

The prediction head has three branches: classification, regression, and center-ness. Each branch employs a 2D convolutional layer paired with an activation function to produce 2D scores and 4D vectors for bounding boxes. During training, we utilize a multi-task loss, integrating binary cross-entropy for classification, smooth L1 for regression, and center-ness loss to pinpoint the object's center. Existing Transformer trackers often fail to accurately trace the target's center, leading to multiple erroneous bounding boxes. To address these suboptimal bounding box predictions, especially away from the target's center, we have incorporated the center-ness branch alongside the other two branches in our comprehensive end-to-end tracking framework.

We evaluated our tracking framework on two challenging and widely used benchmarks, OTB100 and UAV123. Compared to state-of-the-art (SOTA) trackers such as TransT (Chen et al. 2021) and SeqTrack (Chen et al. 2023), our framework outperformed them, as shown in Table 1.

Trackers	Success Score %		Precision Score %	
	OTB100	UAV123	OTB100	UAV123
Ours	**69.6**	**70.2**	**90.3**	87.2
TransT	69.4	68.1	89.9	87.6
SeqTrack	68.3	68.5	89.1	**89.1**

Table 1: Performance comparison with SOTA trackers.

Specifically, on OTB100, we achieved a success score of 69.6% and precision of 90.3%. For UAV123, the success and precision scores were 70.2% and 87.2%, respectively. It demonstrates our framework's superiority over SOTA trackers, especially in handling challenges such as background clutter and occlusion. Moreover, our framework operates at a fast speed, approximately 34 frames per second (FPS).

Discussion and Future Directions

Based on our investigation, the proposed method represents a significant step towards achieving a balanced tracking framework in speed and accuracy, and we believe that future research efforts can help further improve the overall performance of the tracker. We have identified several areas for improvement, including using adaptive attention mechanisms, exploring different regularization techniques, and investigating the potential of additional contextual information in tracking. Additionally, we plan to investigate the performance of our tracker on more challenging datasets such as GOT10k, LaSOT, and TrackingNet to evaluate its robustness and generalization ability. Our ongoing efforts aim to improve the overall performance of the proposed framework.

Furthermore, shallow Transformer architecture has significant applications in other computer vision tasks, such as object detection, multi-object tracking, and video object segmentation. The ability to learn the most important information while ignoring irrelevant details is crucial for many vision tasks. Our proposed architecture can also be beneficial in these areas. Overall, we believe that our framework represents a significant advancement toward achieving robust and efficient VOT, and future research could further enhance its performance and applicability.

References

Bertinetto, L.; Valmadre, J.; Henriques, J. F.; Vedaldi, A.; and Torr, P. H. 2016. Fully-convolutional siamese networks for object tracking. In *ECCV*, 850–865.

Chen, X.; Peng, H.; Wang, D.; Lu, H.; and Hu, H. 2023. SeqTrack: Sequence to Sequence Learning for Visual Object Tracking. In *CVPR*, 14572–14581.

Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; and Lu, H. 2021. Transformer tracking. In *CVPR*, 8126–8135.

Fu, Z.; Fu, Z.; Liu, Q.; Cai, W.; and Wang, Y. 2022. SparseTT: Visual tracking with sparse transformers. *IJCAI*.

Nam, H.; and Han, B. 2015. Learning Multi-Domain Convolutional Neural Networks for Visual Tracking. *CVPR*.

Rahman, M. M.; Fiaz, M.; and Jung, S. K. 2020. Efficient Visual Tracking With Stacked Channel-Spatial Attention Learning. *IEEE Access*, 8: 100857–100869.