SwinTrack: A Simple and Strong Baseline for Transformer Tracking

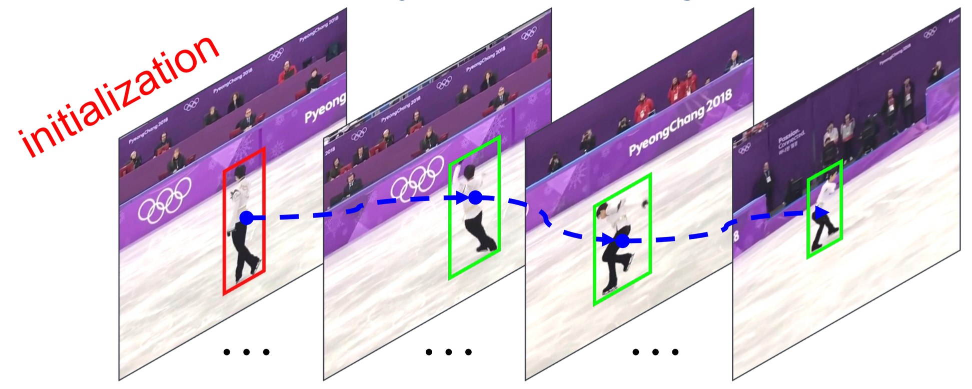
Liting Lin[1,2*] **Heng Fan**[3*] **Zhipeng Zhang**[4] **Yong Xu**[1,2] **Haibing Ling**[5] (*equal contributions)

[1]South China Univ. of Tech. [2]Peng Cheng Lab [3]University of North Texas [4]DiDi Chuxing [5]Stony Brook University

Code/Results

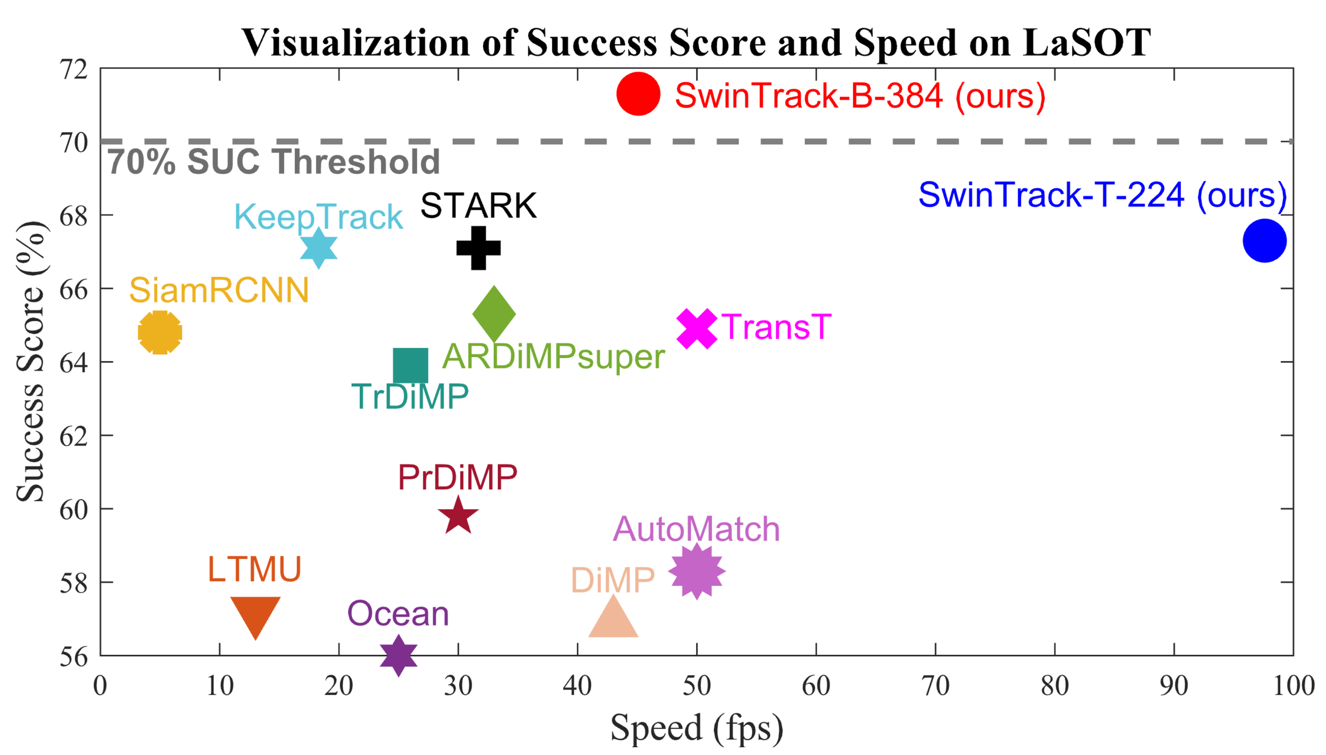
Introduction

- Visual Object Tracking
  - Goal: Continuously localize object of interest in a video



- Motivation
  - Transformer has greatly improved tracking performance
  - Existing approaches usually adopt a hybrid CNN-Transformer architecture, i.e., CNN for feature extraction and Transformer for feature fusion
  - A pure Transformer-based tracking architecture[3], including Transformer-based feature extraction and fusion, is desired
  - Motion information is crucial for temporal visual tracking.
  - Swin Transformer shows SoTA results on various tasks.

Contributions



- A simple but strong baseline, ***SwinTrack***, is proposed with pure Transformer architecture
- We present a simple yet effective motion token in SwinTrack to enhance the robustness
- We conduct empirical studies on different components of SwinTrack, offering guidance for future tracker design
- SwinTrack shows SoTA results on multiple benchmarks, especially setting a new record with **0.713 SUC score** on the challenging LaSOT

SwinTrack Framework

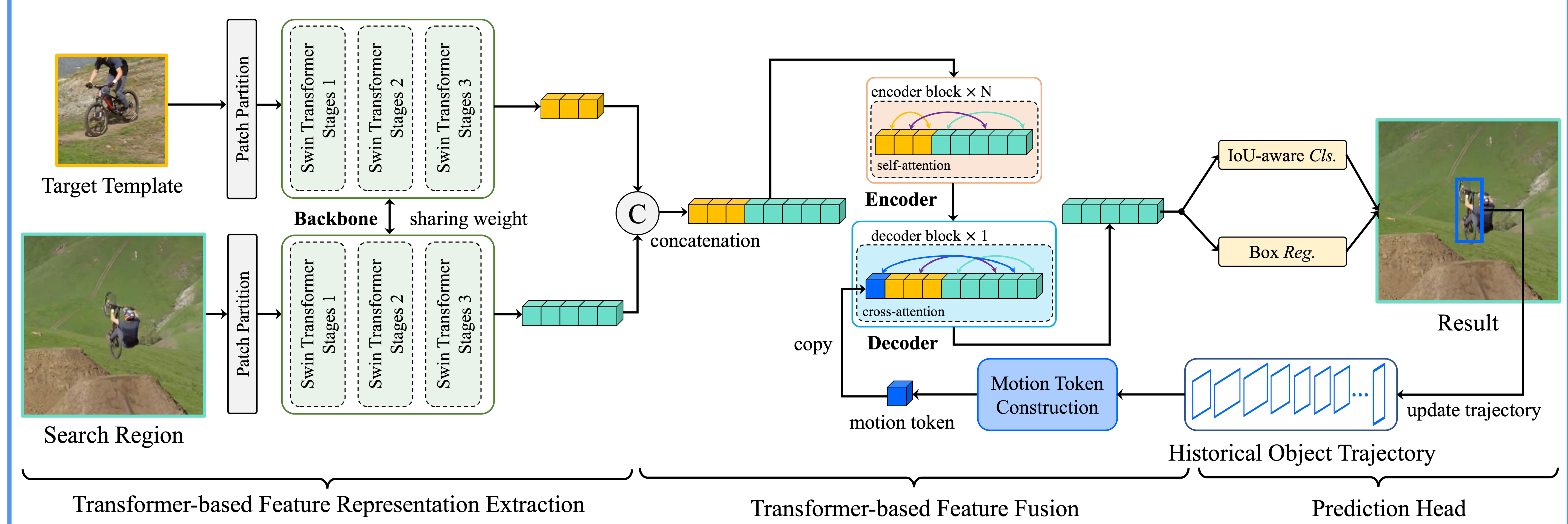


Figure: Architecture of SwinTrack.

- SwinTrack: Siamese Tracking via Vision-Motion Transformer
  - Transformer-based Feature Representation Extraction
    - *Swin Transformer as backbone*
    - *Shared weight to extract template and search tokens*
  - Transformer-based Feature Fusion
    - *Encoder for fusing template and search tokens*
      - Joint vision representation learning

$$\mathrm{f}_z^l, \mathrm{f}_x^l = \mathrm{DeConcat}\left(\mathrm{SelfAttn}\left(\mathrm{Concat}\left(\mathrm{f}_z^{l-1}, \mathrm{f}_x^{l-1}\right)\right)\right)$$

    - *Decoder for fusing vision and motion information*
      - Motion token - Embedding of historical object trajectory

        Concatenation of object past bounding box embedding

$$E_{motion} = Concat\left(E_{s(1)} + E_{s(2)} + \cdots + E_{s(n)}\right)$$

$$s(i) = \max(t - i \times \Delta, 1)$$

      - Vision-motion representation learning

$$\mathrm{f}_{vm} = \mathrm{CrossAttn}\left(\mathrm{Concat}\left(E_{motion}, \mathrm{f}_z^L, \mathrm{f}_x^L\right)\right)$$

    - *Untied positional encoding with multi-dim multi-stream extension*
  - Prediction Head & Loss Function
    - *Response map generation by classification branch*
      - Three-layer perceptron with IoU-aware classification score
    - *Box regression map generation by regression branch*
      - Three-layer perceptron with GIoU loss

Experiments

Table: Comparison with state-of-the-arts.

Tracker	LaSOT SUC	LaSOT P	LaSOT_{ext} SUC	LaSOT_{ext} P	TrackingNet SUC	TrackingNet P	GOT-10k AO	GOT-10k $SR_{0.5}$	GOT-10k $SR_{0.75}$	TNL2k SUC	TNL2k P
C-RPN	45.5	44.3	27.5	32.0	66.9	61.9	-	-	-	-	-
SiamPRN++	49.6	49.1	34.0	39.6	73.3	69.4	51.7	61.6	32.5	41.3	41.2
Ocean	56.0	56.6	-	-	-	-	61.1	72.1	47.3	38.4	37.7
DiMP	56.9	56.7	39.2	45.1	74.0	68.7	61.1	71.7	49.2	44.7	43.4
LTMU	57.2	57.2	41.4	47.3	-	-	-	-	-	48.5	47.3
SiamR-CNN	64.8	-	-	-	81.2	80.0	64.9	72.8	59.7	52.3	52.8
STMTrack	60.6	63.3	-	-	80.3	76.7	64.2	73.7	57.5	-	-
AutoMatch	58.3	59.9	37.6	43.0	76.0	72.6	65.2	76.6	54.3	-	-
TrDiMP	63.9	61.4	-	-	78.4	73.1	67.1	77.7	58.3	-	-
TransT	64.9	69.0	-	-	81.4	80.3	67.1	76.8	60.9	51.0	-
STARK	67.1	-	-	-	82.0	-	68.8	78.1	64.1	-	-
KeepTrack	67.1	70.2	48.2	-	-	-	-	-	-	-	-
SwinTrack-T-224	67.2	70.8	47.6	53.9	81.1	78.4	71.3	81.9	64.5	53.0	53.2
SwinTrack-B-384	71.3	76.5	49.1	55.6	84.0	82.8	72.4	80.5	67.8	55.9	57.1

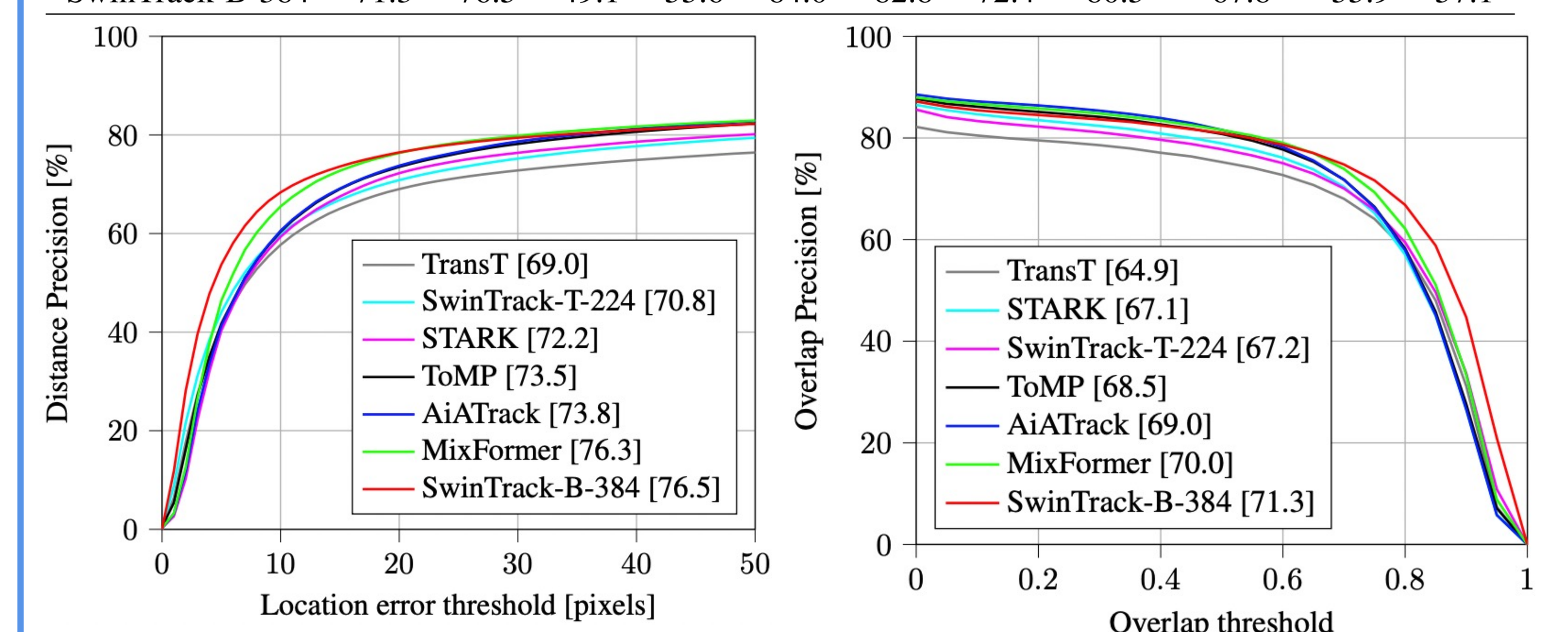


Figure: Comparison with latest Transformer-based trackers on LaSOT.

Table: Ablations with on SwinTrack-T-224 w/o motion token. ❶: baseline; ❷: replacing Transformer backbone w. ResNet-50; ❸: replacing feature fusion w. cross attention-based fusion; ❹: replacing decoder w. a target query-based; ❺: replacing united positional encoding w. absolute sine position encoding; ❻: replacing IoU-aware classification loss w. plain binary cross entropy loss; ❼: removing the Hanning penalty window in inference.

	LaSOT SUC (%)	LaSOT_{ext} SUC (%)	TrackingNet SUC (%)	GOT-10k mAO (%)	Speed *fps*	Params M
❶	66.7	46.9	80.8	70.9	98	22.7
❷	64.2	41.8	79.5	68.2	121	20.0
❸	66.6	45.4	80.2	69.3	72	34.6
❹	66.6	43.2	79.6	69.0	91	25.3
❺	65.7	45.0	80.0	70.0	103	21.6
❻	66.2	46.7	79.4	68.2	98	22.7
❼	65.7	46.0	80.0	69.6	98	22.7

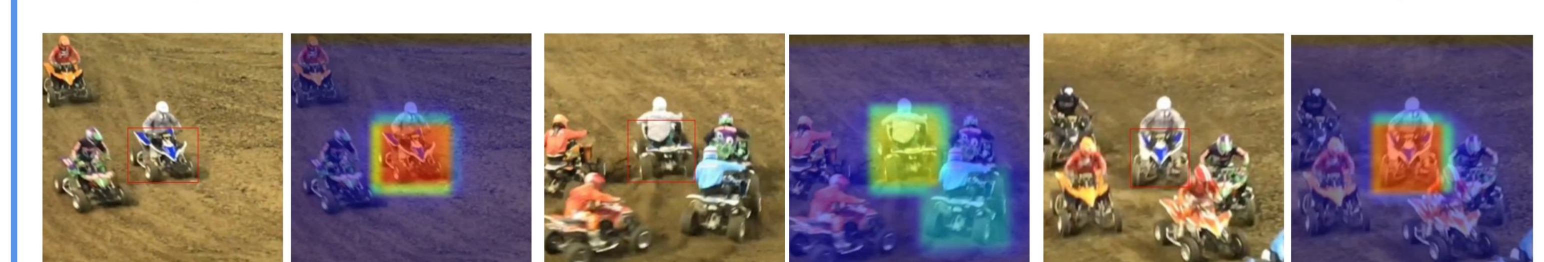


Figure: Visualization of tracking response maps of SwinTrack.

Key References

1. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, *ICCV*, 2021.
2. X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, H. Lu, Transformer tracking, *CVPR*, 2021.
3. A. Dosovitskiy, *et al.*, An Image is worth 16x16 words: Transformers for image recognition at scale, *ICLR*, 2021.
4. A. Vaswani. *et al.*, Attention Is all you need, *NIPS*, 2017.