Joint Spatio-Temporal Similarity and Discrimination Learning for Visual Tracking

Yanjie Liang, Haosheng Chen, Qiangqiang Wu, Changqun Xia, and Jia Li, *Senior Member, IEEE*

***Abstract*— Visual tracking is a task of localizing a target unceasingly in a video with an initial target state at the first frame. The limited target information makes this problem an extremely challenging task. Existing tracking methods either perform matching based similarity learning or optimization based discrimination reasoning. However, these two types of tracking methods suffer from the problem of ineffectiveness for distinguishing target objects from background distractors and the problem of insufficiency in maintaining spatio-temporal consistency among successive frames, respectively. In this paper, we design a joint spatio-temporal similarity and discrimination learning (STSDL) framework for accurate and robust tracking. The designed framework is composed of two complementary branches: a similarity learning branch and a discrimination learning branch. The similarity learning branch uses an effective transformer encoder-decoder to gather rich spatio-temporal context information to generate a similarity map. In parallel, the discrimination learning branch exploits an efficient model predictor to train a target model to produce a discriminative map. Finally, the similarity map and the discriminative map are adaptively fused for accurate and robust target localization. Experimental results on six prevalent datasets demonstrate that the proposed STSDL can obtain satisfactory results, while it retains a real-time tracking speed of 50 FPS on a single GPU.**

***Index Terms*— Video object tracking, joint learning, spatio-temporal similarity, spatio-temporal discrimination, adaptive response map fusion.**

I. Introduction

In recent years, video object tracking has become a hot research topic in the field of video analysis, and it has numerous applications, such as human object interaction [1], video surveillance [2], autonomous driving [3] and motion estimation [4]. This task aims to continuously localize a target in a video sequence by initializing the target state with a rectangle/rotated box at the first frame. Despite considerable development being made over the past years, there are still many challenges, such as deformation, occlusion, background clutter, etc. In the community, many researchers have developed various deep learning methods for video object tracking [5], [6]. The recent tracking methods contain matching based similarity learning methods and optimization based discrimination learning methods.

The matching based similarity learning methods (e.g., SiamRPN++ [7], SiamFC++ [8], SiamCAR [9], SiamGAT [10], STMTrack [11]) usually perform direct reasoning from reference frames to a test frame to facilitate visual tracking. Existing similarity learning methods deploy cross-correlation operations [7], [8], [9], a graph attention network [10] or a space-time memory network [11] to produce intermediate feature-level similarity maps between a test frame and reference frames for classification and regression. As the feature-level similarity maps can convey spatio-temporal context information from the reference frames to the test frame, these similarity learning methods can favorably preserve rich spatio-temporal cues with considerable time efficiency. Despite obtaining the favorable performance, the similarity learning methods heavily depend on the generalization of feature matching network learnt offline for accurate tracking, and thus they suffer from limited discrimination and generalization capability (as shown on the first row in Fig. 1). Therefore, it is possible to introduce discrimination learning into the similarity learning methods to improve their capability of distinguishing target objects from background distractors.

The optimization based discrimination learning methods (e.g., ATOM [12], DiMP [13], PrDiMP [14], TrDiMP [15], DCFST [16]) typically use reference frames to train a discriminative target model, and then apply this model to a test frame to facilitate visual tracking. Existing discrimination learning methods employ discriminative correlation filters [12], deep discriminative models [13], [14], [15] or discriminant models [16] to perform online discrimination learning to produce response-level discriminative maps for target localization. The model predictors in the discrimination learning methods have superior distractor discrimination capability. However, these discrimination learning methods only consider reference frames as independent samples, which fails to fully exploit rich

Manuscript received 24 July 2023; revised 28 January 2024; accepted 5 March 2024. Date of publication 13 March 2024; date of current version 12 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62132002, Grant 62202249, and Grant 62102206; in part by the Postdoctoral Science Foundation of China under Grant 2022M721732; and in part by the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (CPSF) under Grant GZC20233362. This article was recommended by Associate Editor J. Shen. *(Corresponding author: Jia Li.)*

Yanjie Liang and Changqun Xia are with the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: liangyj@pcl.ac.cn; xiachq@pcl.ac.cn).

Haosheng Chen is with Chongqing Key Laboratory of Image Cognition, College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: chenhs@cqupt.edu.cn).

Qiangqiang Wu is with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: qiangqwu2-c@my.cityu.edu.hk).

Jia Li is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: jiali@buaa.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2024.3377379.

Digital Object Identifier 10.1109/TCSVT.2024.3377379

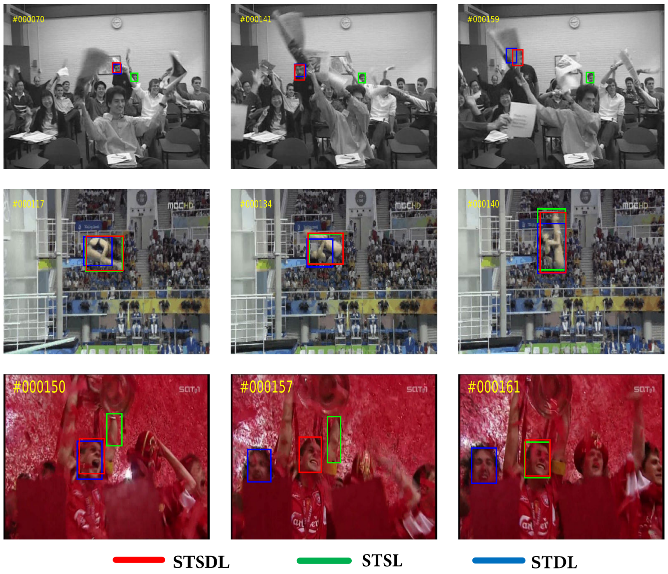


Fig. 1. Comparisons of our STSDL with STSL and STDL on three videos from the OTB2015 dataset. The three videos from top to bottom are *Freeman4*, *Diving* and *Soccer*, respectively.

spatio-temporal context cues. The spatio-temporal context cues contained in successive frames are vital for achieving spatio-temporal consistent tracking results, which has been proven by some similarity learning methods [11] (as illustrated on the second row in Fig. 1). Thus, it is possible to incorporate spatio-temporal similarity learning into the discrimination learning methods to perform accurate target localization across spatio-temporal dimensions.

The aforementioned analysis indicates that the matching based similarity learning and the optimization based discrimination learning are complementary to each other. The former can fully exploit the spatio-temporal context information but it struggles to discriminate target objects from similar distractors. In contrast, the latter is more effective to discriminate similar objects but struggles to maintain spatio-temporal consistency. Therefore, it is reasonable to perform joint spatio-temporal similarity and discrimination learning in a unified framework to explore their potentials. As these two types of learning methods solve the tracking task from different perspectives, there are some challenges for their integration. The matching based similarity learning methods rely on intermediate feature-level similarity maps for classification and regression [7], [8], [11]. In contrast, the optimization based discrimination learning methods produce direct response-level discriminative maps for target localization [13], [14], [15]. It is nontrivial to design a joint learning framework to integrate the feature-level similarity maps and response-level discriminative maps to take full advantage of their complementary properties.

In this work, we propose a novel joint learning framework to model the spatio-temporal similarity and discrimination for visual tracking. The proposed learning framework mainly contains two parallel branches: a similarity learning branch and a discrimination learning branch. The similarity learning branch deploys a transformer encoder-decoder to gather rich spatio-temporal context information to produce a response-level similarity map, while the discrimination learning branch employs a few-shot learner to produce a response-level discriminative map to discriminate target objects from background clutters. Afterwards, the outputs (i.e., the similarity and discriminative maps) from these two parallel branches are adaptively fused to encode both spatio-temporal similarity and discrimination information for target localization.

Fig. 1 compares the proposed joint spatio-temporal similarity and discrimination learning (STSDL) with spatio-temporal similarity learning (STSL) and spatio-temporal discrimination learning (STDL). As shown on the first row, STSL that merely performs similarity learning is distracted by similar objects (see the green boxes). However, the proposed STSDL which performs joint similarity and discrimination learning can effectively discriminate between the target object and similar distractors (see the red boxes). Furthermore, as depicted on the second row, STDL that only conducts discrimination learning cannot enclose the whole diver (see the blue boxes). In contrast, the proposed STSDL can obtain the accurate bounding boxes of the diver (see the red boxes), which benefits from the similarity learning to preserve the spatio-temporal consistency. Moreover, as illustrated on the third row, when the tracking scenario becomes much more challenging, both STSL and STDL cannot achieve satisfactory tracking results (see the green and blue boxes). In contrast, the proposed STSDL that incorporates both similarity learning and discrimination learning can effectively discriminate the target object from background clutters and consistently obtain the accurate bounding boxes of target objects (see the red boxes).

This paper makes four-fold contributions as follows:

- A novel joint spatio-temporal similarity and discrimination learning framework is proposed for visual tracking, which fully exploits the merits of both similarity learning approaches and discrimination learning approaches to enhance the robustness of response map for target localization.
- A lightweight transformer is carefully designed in the similarity learning branch to gather rich spatio-temporal context information, which is beneficial to preserve the spatio-temporal consistency information in a similarity map.
- An efficient few-shot learner is naturally introduced into the discrimination learning branch to discriminate target objects from background distractors, which is effective to retain the spatio-temporal discriminative information in a discriminative map.
- An adaptive response map fusion module is devised to aggregate the complementary response maps from the two branches, which are parallel with each other and share the feature extraction and the loss computation, facilitating the learning framework end-to-end trainable.

We conduct experiments on six prevalent datasets (i.e., GOT10K, TrackingNet, LaSOT, UAV123, OTB2015 and VOT2020). The evaluation results show that our STSDL can obtain the state-of-the-art performance with a real-time tracking speed.

II. Related Work

In this section, we firstly give a brief review of video technologies, and then review the recent video object tracking technologies. We roughly divide the recent video object tracking methods into four categories: tracking by discriminative correlation filters, tracking by discrimination learning, tracking by similarity learning, and tracking by transformers.

A. Video Technologies

In recent years, video technologies have been developed rapidly with a wide range of real-world applications, and nowadays there are various video-related tasks emerging, including person retrieval, person re-ID, video object detection, video object segmentation, video object tracking, etc. For person retrieval, AMR [17] develops an attribute mining and reasoning framework to mine discriminative attribute features and discover their latent relations in a person retrieval system for video surveillance. APN [18] employs deep reinforcement learning to dynamically search for the optimal partition settings for various pedestrian images to construct a robust retrieval system. For person re-ID, U-SSL [19] uses pseudo-pairs to perform self-similarity learning for unsupervised person re-ID, which can be applied to video surveillance systems. The work in [20] deploys generative adversarial networks to produce adversarial examples for generative metric learning, providing a reliable re-ID system in the open world. For video object detection, TCNet [21] presents a triple-cooperative framework to boost the detection performance from the perspectives of target localization, class recognition and relation learning. GGMLCN [22] exploits the global memory using a global memory bank and the local continuity using an object tracker for high-speed high-accuracy video object detection. For video object segmentation, STM [23] develops a space-time memory network, serving as an external memory to store the object features and masks at historical frames, to segment the object at current frame. LOAGLC [24] performs target-aware correspondence learning to obtain temporal coherent object-level features for accurate and robust video object segmentation. For video object tracking, e-TLD [4] presents a tracking-learning-detection framework using a moving event camera for long-term tracking. As an advanced video object tracking method, the proposed STSDL that performs joint spatio-temporal similarity and discrimination learning can also provide potential inspirations for other video technologies, such as person re-ID, video object detection, video object segmentation.

B. Tracking by Discriminative Correlation Filters

Discriminative correlation filters have been developed rapidly in the past decade [25], [26], [27], [28], [29], [30], [31]. DCF [25] is a real-time correlation filter based tracking method, which regresses the circular shifts of a sample to the Gaussian-shaped labels. After that, many tracking methods have been developed by researchers to address the drawbacks of DCF. For instance, MCPF [32] combines the particle filters with correlation filters to deal with scale variations. DeepCFIAP++ [33] incorporates the instance-aware proposals into correlation filters to cope with various complex tracking scenarios. To alleviate spatial effects, several methods [27], [34] introduce spatial regularizations to train correlation filters. To mitigate temporal degradations, some methods [29], [34], [35] introduce temporal regularizations into correlation filter learning. Modern discriminative correlation filters [26], [36], [37], [38] usually extract deep features to represent target objects. HCF [36] incorporates hierarchical responses from multiple convolutional layers for target localization. To reduce the redundancy among high dimensional deep features, ECO [37] learns a factorized matrix on-line to compress deep features. Although performance gains can be obtained by using these advanced techniques, the tracking speed becomes non-real time. In contrast to the aforementioned tracking methods that employ discriminative correlation filters for 2D target localization, OTR [39] performs 3D target reconstruction by learning some view-specific discriminative correlation filters for RGB-D tracking.

C. Tracking by Discrimination Learning

Inherited from discriminative correlation filters, deep discriminative models have receieved considerable attention by researchers in visual tracking community [12], [13], [14], [16], [40], [41]. ATOM [12] employs an online classifier to ensure its discriminability in the presence of distractors. DCFST [16] incorporates the solver of a discriminative target model into neural networks to optimize the feature embedding for robust tracking. DiMP [13] develops a deep regression network to predict a discriminative target model by using both target and background context information. PrDiMP [14] trains a probabilistic regression network with a Kullback-Leibler divergence loss for conditional probability density estimation. Considering DiMP as the baseline tracker, KYS [40] further utilizes complex scene information (e.g., target, background, distractor) for more discriminative tracking. CARE [42] uses a cascaded regression framework with two sequential stages (i.e., convolutional regression stage and ridge regression stage) for discriminative tracking. DET [43] develops an ensemble learning framework to train diverse discriminative models for robust tracking. Although these deep discriminative model based tracking methods can achieve favorable performance, they cannot take full advantage of the rich spatio-temporal context information, which is crucial to achieve spatio-temporal consistent tracking results. In contrast to the existing discrimination learning methods, our STSDL further introduces a lightweight transformer encoder-decoder in the similarity learning branch to preserve spatio-temporal consistency.

D. Tracking by Similarity Learning

Siamese networks [7], [8], [44], [45], [46], [47], [48], [49] have drawn much more attention in the past few years. SiamFC [44] is the first tracking method to match the initial template with the current search region by using a Siamese network. After that, SiamRPN [45] integrates a region proposal network into a Siamese network for accurate tracking. GradNet [50] takes advantage of the discriminative gradient information to update the template in a Siamese network to capture the appearance variations of targets or background clutters over time. LK-SiamFC/LK-SiamRPN [46] introduces a Lucas-Kanade network into SiamFC/SiamRPN for more accurate matching. C-RPN [51] introduces cascaded RPNs from low-level to high-level into a Siamese network to address the challenges of similar background clutters and large scale variations. To facilitate SiamRPN with powerful backbones, some anchor-based Simese networks (e.g., SiamRPN++ [7] and SiamDW [52]) have been developed to alleviate the influence of padding in different manners with considerable performance gains. SiamLTR++ [53] introduces a ranking network into SiamRPN++ to rank proposals for robust tracking. Although these anchor-based Siamese networks can achieve favorable performance, they require to cautiously configure anchor boxes. To avoid the problem, some anchor-free Siamese networks (e.g., SiamFC++ [8], SiamCAR [9], SiamBAN [54], SiamTDN [55] and SiamGAT [10]) have been proposed for direct target classification and bounding box regression. STMTrack [11] develops a space-time memory network to cope with the problem of target appearance variations for anchor-free Siamese tracking. Although the Siamese networks obtain the favorable performance, they cannot effectively discriminate target objects from background distractors. In contrast to the existing similarity learning methods, our STSDL further introduces an efficient few-shot learner into the discrimination learning branch to discriminate target objects from similar distractors.

E. Tracking by Transformers

Recently, transformers have been naturally introduced into visual tracking for their excellent performance in other computer vision tasks. The first type of transformer based tracking methods [15], [56] typically use transformers to predict the discriminative features for tracking. For instance, DTT [56] feeds both reference frame and test frame into a transformer to estimate the target state. In particular, the transformer encoder is responsible for feature encoding, whereas the transformer decoder is employed for feature matching. TrDiMP [15] uses encoded features of reference frame to train a discriminative target model, which is further convolved with decoded features of test frame for target localization. The second type of transformer based tracking methods [57], [58], [59], [60] stack features of both reference frame and test frame with transformers. For instance, TransT [57] uses multiple attention layers to fuse features for target classification and regression. Following the paradigm of DETR [61], STARK [58] adopts a full transformer to mix template and search region features for bounding box prediction. ToMP [59] also employs another full transformer from DETR [61] to predict the parameters of a target classifier and a bounding box regressor. The third type of transformer-based tracking methods [60], [62], [63] typically construct one-stream unified tracking frameworks with transformers. For instance, MixFormer [60] combines feature extraction and feature fusion by using iterative mixed attention modules for end-to-end tracking. OSTrack [62] constructs a transformer that combines feature learning and relation modeling by allowing for bidirectional feature interaction between template and search region.

TABLE I

EXPLANATIONS OF MATHEMATICAL SYMBOLS IN METHODOLOGY

Explanation	Symbol
Backbone features of reference frames	$\mathbf{Z} \in \mathbb{R}^{N \times H \times W \times C}$
Backbone features of test frame	$\mathbf{X} \in \mathbb{R}^{H \times W \times C}$
Target bounding boxes of reference frames	$\mathbf{B}_{ref} \in \mathbb{R}^{N \times 4}$
Target centers of reference frames	$\mathbf{C}_{ref} \in \mathbb{R}^{N \times 2}$
Ground-truth response maps of reference frames	$\mathbf{R}_{ref} \in \mathbb{R}^{N \times H \times W}$
Target bounding box of test frame	$\mathbf{B}_{test} \in \mathbb{R}^{4}$
Ground-truth response map of test frame	$\mathbf{R}_{test} \in \mathbb{R}^{H \times W}$
Predicted similarity map of test frame	$\mathbf{R}_{sim} \in \mathbb{R}^{H \times W}$
Predicted discriminative map of test frame	$\mathbf{R}_{dis} \in \mathbb{R}^{H \times W}$
Predicted response map of test frame	$\mathbf{R}_{fus} \in \mathbb{R}^{H \times W}$
Attention features of reference frames	$\mathbf{A}_{\hat{\mathbf{Z}}} \in \mathbb{R}^{NHW \times C}$
Encoded features of reference frames	$\hat{\mathbf{Z}}_{enc} \in \mathbb{R}^{NHW \times C}$
Attention features of test frame	$\mathbf{A}_{\hat{\mathbf{X}}} \in \mathbb{R}^{HW \times C}$
Encoded features of test frame	$\hat{\mathbf{X}}_{enc} \in \mathbb{R}^{HW \times C}$
Discriminative target model	$\mathbf{W} \in \mathbb{R}^{K \times K \times C}$
Regulation parameter for discriminative loss	λ
Adaptive learning rate for discriminative target model	γ
Gradient of discriminative loss	$\nabla\mathcal{L}$
Jacobian of residuals	J

Although the transformer based tracking methods can achieve excellent performance, they contain many attention layers to calculate similarity matrices between feature maps, thus leading to large memory usage and long training time. This severely impacts the training and inference time. In comparison with these transformer based tracking methods, our STSDL only uses a lightweight transformer encoder-decoder (i.e., a single self-attention layer and a single cross-attention layer) to capture the spatio-temporal context information, which is memory and time efficient. Therefore, it is unfair to compare our STSDL (with small model size) with the transformer-based tracking methods (with large model size).

III. METHODOLOGY

In this section, we firstly give the framework of joint spatio-temporal similarity and discrimination learning (STSDL) in Sec. III-A. Then, we introduce the spatio-temporal similarity learning branch in Sec. III-B and the spatio-temporal discrimination learning branch in Sec. III-C. Afterwards, we illustrate the adaptive response map fusion module in Sec. III-D. Finally, we describe the offline training procedure in Sec. III-E and the online inference process in Sec. III-F. Table I provides a detailed explanation of the mathematical symbols in this section.

A. Overall Framework

The basic idea of our novel framework stems from the observation that similarity learning approaches can preserve rich spatio-temporal cues to achieve *accurate* target localization and discrimination learning approaches are *robust*

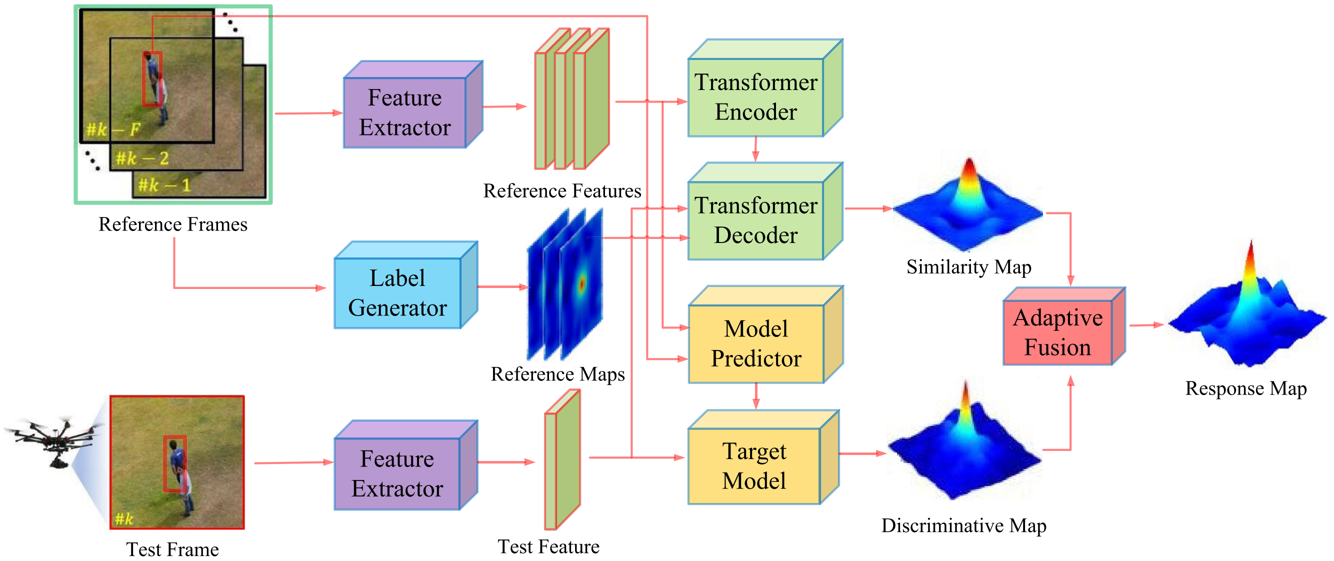


Fig. 2. **Overall framework of our STSDL.** The proposed STSDL incorporates similarity learning and discrimination learning into a unified framework to exploit the complementarity of similarity learning approaches and discrimination learning approaches for accurate and robust tracking. The framework consists of two parallel branches: the spatio-temporal similarity learning (STSL) branch and the spatio-temporal discrimination learning (STDL) branch. The STSL branch employs a transformer encoder-decoder to produce a similarity map, while the STDL branch resorts to a model predictor to generate a discriminative map. Afterwards, the similarity map and discriminative map are merged by using an adaptive fusion module to attain the response map.

to discriminate between target objects and background distractors. In this paper, we propose a joint spatio-temporal similarity and discrimination learning (STSDL) framework to explore the advantages of the above two types of approaches, producing complementary response maps for accurate and robust tracking. Our proposed STSDL mainly contains two complementary branches. The similarity learning branch employs a transformer encoder-decoder to gather rich spatio-temporal information in a similarity map while the discrimination learning branch employs a few-shot learner to encode discriminative target information in a discriminative map. Afterwards, the proposed STSDL devises an adaptive response map fusion module to aggregate the similarity map and discriminative map as a final response map (which preserves both spatio-temporal consistent information and discriminative information) for accurate and robust target localization.

Fig. 2 illustrates the STSDL framework. As shown in the figure, the first frame and preserved past frames are treated as reference frames, the current frame is treated as a test frame. Firstly, the reference frames and the test frame are fed into a shared feature extractor to extract backbone features $\mathbf{Z} \in \mathbb{R}^{N\times H\times W\times C}$ and $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, where H, W and C denote the height, the width and the channel of backbone features, respectively. N is the number of the reference frames. Meanwhile, the reference frames are fed into a label generator to produce Gaussian-shaped reference maps $\mathbf{R}_{ref} \in \mathbb{R}^{N\times H\times W}$. Afterwards, the spatio-temporal similarity learning branch takes both $\mathbf{Z}$ and $\mathbf{X}$ as input, and it uses a transformer encoder-decoder to convert $\mathbf{R}_{ref}$ to a similarity map $\mathbf{R}_{sim} \in \mathbb{R}^{H\times W}$ according to the affinity between $\mathbf{Z}$ and $\mathbf{X}$. In parallel, the spatio-temporal discrimination learning branch also takes both $\mathbf{Z}$ and $\mathbf{X}$ as input. It firstly uses a model predictor to train a discriminative target model (where reference features $\mathbf{Z} \in \mathbb{R}^{N\times H\times W\times C}$ and corresponding annotated bounding boxes $\mathbf{B}_{ref} \in \mathbb{R}^{N\times 4}$ are treated as training pairs) by solving an optimization problem, and then it convolves the discriminative target model with $\mathbf{X}$ to predict a discriminative map $\mathbf{R}_{dis} \in \mathbb{R}^{H\times W}$. Finally, the similarity map $\mathbf{R}_{sim}$ and the discriminative map $\mathbf{R}_{dis}$ from the two parallel branches are adaptively fused into a response map $\mathbf{R}_{fus}$. The three core components of the proposed STSDL consist of a transformer encoder-decoder for spatio-temporal similarity learning, a model predictor for spatio-temporal discrimination learning, and an adaptive fusion module for response map fusion.

Fig. 3 illustrates four examples to show that the final response maps can provide better target localization than the intermediate similarity maps and discriminative maps. As depicted on the first/second row, when the target objects are interfered by background distractors, the similarity maps are confused to localize the targets, whereas the discriminative maps and the response maps can discriminate between target objects and background distractors. As illustrated on the third/fourth row, in the case of significant deformations or large rotations, the discriminative maps suffer from inaccurate target localization. In contrast, the similarity maps and response maps can accurately localize the target centers.

B. Spatio-Temporal Similarity Learning

In this subsection, we introduce the first core component of the proposed STSDL: a transformer encoder-decoder for spatio-temporal similarity learning. As shown in Fig. 2, the spatio-temporal similarity learning branch is responsible for aggregating the rich spatio-temporal consistent target information to produce a similarity response map. In recent years, transformers have been proven to have great potentials to perform information interaction or aggregation [23], [55], [64].

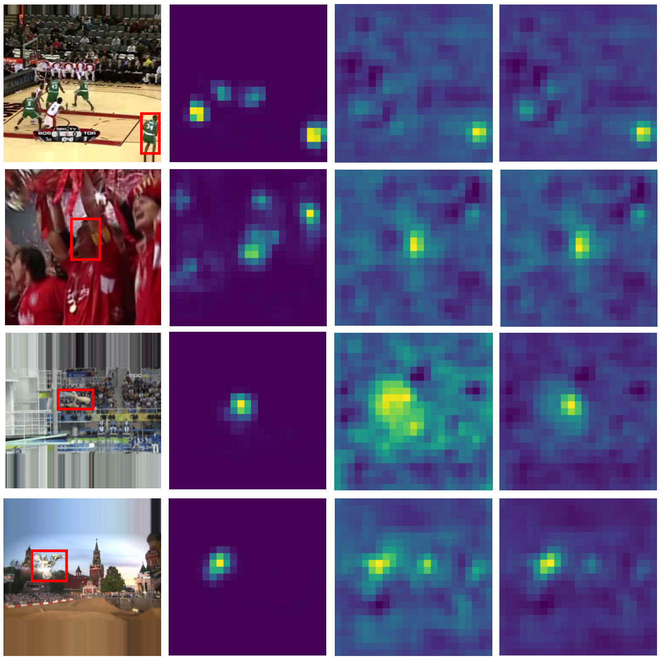

(a) Search Region (b) Similarity Map (c) Discriminative Map (d) Response Map

Fig. 3. Comparisons of similarity maps, discriminative maps and response maps produced from our framework on four examples. The final response maps can perform better target localization than both the intermediate similarity maps and discriminative maps.

Therefore, the spatio-temporal similarity learning branch introduces a lightweight transformer encoder-decoder to perform similarity learning of rich spatio-temporal information.

The key components of the transformer encoder-decoder are attention modules, which have a strong capability to establish long-range dependencies of input features. In our spatio-temporal similarity learning branch, the attention modules are proposed to be more appropriate for visual tracking. Firstly, the key $\mathbf{K} \in \mathbb{R}^{l_k \times d_k}$ and the query $\mathbf{Q} \in \mathbb{R}^{l_q \times d_k}$ are respectively normalized across the channel dimension. Then, the similarity matrix is computed and rescaled with a parameter τ. Finally, the rescaled similarity matrix is normalized with a softmax function to weigh the value $\mathbf{V} \in \mathbb{R}^{l_k \times d_v}$. The above computation can be formulated as follows:

$$Attn(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = SoftMax(\frac{Norm(\mathbf{Q})Norm(\mathbf{K})}{\tau})\mathbf{V}, \quad (1)$$

where $Norm(\cdot)$ refers to the l_2 normalization across the channel dimension, and $SoftMax(\cdot)$ denotes the softmax operator. Fig. 4 illustrates the lightweight transformer, which consists of a transformer encoder and a transformer decoder as follows:

1) Transformer Encoder: The transformer encoder takes the reference feature $\mathbf{Z} \in \mathbb{R}^{N \times H \times W \times C}$ as input, which is further reshaped as $\hat{\mathbf{Z}} \in \mathbb{R}^{NHW \times C}$ for subsequent operations. For the self-attention layer of the transformer encoder, it firstly applies two separate linear functions $\phi(\cdot)$ and $\psi(\cdot)$ to transform the query and key, where the channel dimension can be compressed from C to $C/4$ for efficient matrix multiplication. Then, it computes the attention feature $\mathbf{A}_{\hat{\mathbf{Z}}} \in \mathbb{R}^{NHW \times C}$ according to Eq. (1) as follows:

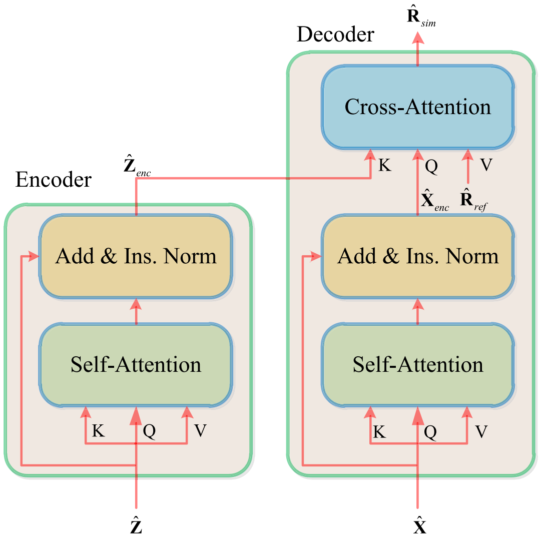


Fig. 4. Overview of our transformer encoder-decoder, which is applied in the spatio-temporal learning branch. It is elaborately designed to propagate spatio-temporal consistent target information from reference frames to a test frame. Specifically, it propagates the labels of reference frames (i.e., reference maps) to produce the label of a test frame (i.e., a similarity map) according to the similarity between reference features and test feature.

$$\mathbf{A}_{\hat{\mathbf{Z}}} = Attn(\phi(\hat{\mathbf{Z}}), \psi(\hat{\mathbf{Z}}), \hat{\mathbf{Z}}). \quad (2)$$

Afterwards, the attention reference feature $\mathbf{A}_{\hat{\mathbf{Z}}}$ is added to the reshaped reference feature $\hat{\mathbf{Z}}$, and the added feature is further fed into the instance normalization layer of the transformer encoder to produce the encoded reference feature $\hat{\mathbf{Z}}_{enc} \in \mathbb{R}^{NHW \times C}$ as follows:

$$\hat{\mathbf{Z}}_{enc} = InsNorm(\mathbf{A}_{\hat{\mathbf{Z}}} + \hat{\mathbf{Z}}), \quad (3)$$

where $InsNorm(\cdot)$ denotes the instance normalization operator. The transformer encoder facilitates the reference feature to be more representative in a reinforcement way, and thus the encoded reference feature is more reasonable to perform feature matching.

2) Transformer Decoder: The transformer decoder is composed of a self-attention module and a cross-attention module. The self-attention module takes the test feature $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ as input, and it processes the test feature $\mathbf{X}$ in a similar way as the transformer encoder, i.e., the attention test feature is firstly computed according to the similarity between the query and the key, and then it is added to the reshaped test feature $\hat{\mathbf{X}}$ for instance normalization:

$$\mathbf{A}_{\hat{\mathbf{X}}} = Attn(\phi(\hat{\mathbf{X}}), \psi(\hat{\mathbf{X}}), \hat{\mathbf{X}}), \quad (4)$$
$$\hat{\mathbf{X}}_{enc} = InsNorm(\mathbf{A}_{\hat{\mathbf{X}}} + \hat{\mathbf{X}}), \quad (5)$$

where $\hat{\mathbf{X}} \in \mathbb{R}^{HW \times C}$, $\mathbf{A}_{\hat{\mathbf{X}}} \in \mathbb{R}^{HW \times C}$ and $\hat{\mathbf{X}}_{enc} \in \mathbb{R}^{HW \times C}$ denote the reshaped test feature, the attention test feature and the encoded test feature, respectively.

The cross-attention module, which is the primary component of our transformer encoder-decoder, propagates the rich spatio-temporal information according to the patch-level correspondence between the test frame and the reference frames.

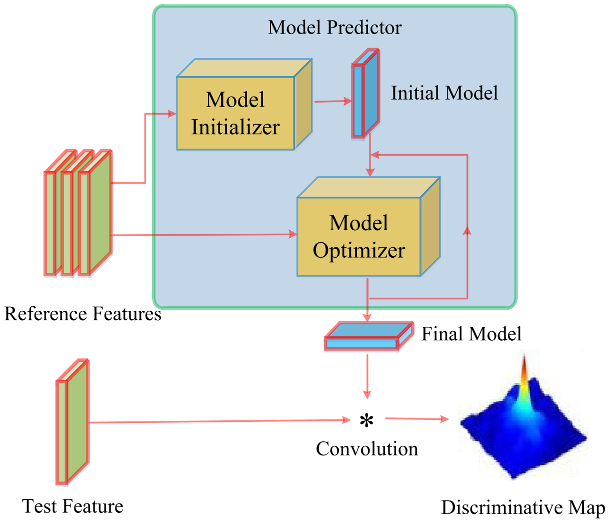


Fig. 5. Overview of the discriminative model predictor, which is applied in the spatio-temporal discrimination learning branch. It is carefully designed to encode spatio-temporal discriminative target information in a target model by using reference features, and then the predicted target model is convolved with the test feature to produce a discriminative map.

The cross-attention module takes the encoded test feature $\hat{\mathbf{X}}_{enc} \in \mathbb{R}^{HW\times C}$ and the encoded reference feature $\hat{\mathbf{Z}}_{enc} \in \mathbb{R}^{NHW\times C}$ as inputs. It firstly uses the same linear function to generate the query $\varphi(\hat{\mathbf{X}}_{enc})$ and key $\varphi(\hat{\mathbf{Z}}_{enc})$, respectively. Then, it transforms the ground-truth maps $\mathbf{R}_{ref} \in \mathbb{R}^{N\times H\times W}$ of reference frames to the similarity map $\mathbf{R}_{sim} \in \mathbb{R}^{H\times W}$ of test frame according to the similarity between the query and the key as follows:

$$\hat{\mathbf{R}}_{sim} = Attn(\varphi(\hat{\mathbf{Z}}_{enc}), \varphi(\hat{\mathbf{X}}_{enc}), \hat{\mathbf{R}}_{ref}), \tag{6}$$

where $\mathbf{R}_{ref} \in \mathbb{R}^{N\times H\times W}$ is reshaped into $\hat{\mathbf{R}}_{ref} \in \mathbb{R}^{NHW}$ for cross-attention, and $\hat{\mathbf{R}}_{sim} \in \mathbb{R}^{HW}$ is reshaped as the similarity response map $\mathbf{R}_{sim} \in \mathbb{R}^{H\times W}$.

C. Spatio-Temporal Discrimination Learning

In this subsection, we describe the second core component of the proposed STSDL: a model predictor for spatio-temporal discrimination learning. As described in Sec. III-B, the spatio-temporal similarity learning branch can aggregate the rich spatio-temporal cues to provide the similarity map. Nevertheless, due to the lack of online adaptation, the similarity learning branch cannot effectively cope with unseen targets, and it is also vulnerable to discriminate between target objects and background distractors. To overcome these limitations, we further introduce a spatio-temporal discrimination branch into the framework to produce a discriminative response map as depicted in Fig. 2. Few-shot learners have been proven to be effective to discriminate between different categories of target objects [13], [14], [16]. Therefore, the spatio-temporal discrimination learning branch employs a few-shot learner (i.e., a model predictor) to perform online discrimination learning.

As depicted in Fig. 5, the model predictor takes the backbone feature $\mathbf{Z} \in \mathbb{R}^{N\times H\times W\times C}$ and the target bounding box $\mathbf{B}_{ref} \in \mathbb{R}^{N\times 4}$ as inputs to optimize the discriminative target model $\mathbf{W} \in \mathbb{R}^{K\times K\times C}$. It initializes the discriminative target model $\mathbf{W}$ with the features residing in the target region. Then, it uses the features residing in both target and background regions to train the discriminative target model $\mathbf{W}$ by using the discriminative loss as follows:

$$\mathcal{L}(\mathbf{W}) = \frac{1}{|S_{train}|}\sum_{(\mathbf{Z}_i,\mathbf{C}_i)\in S_{train}} \|l(\mathbf{Z}_i * \mathbf{W}, \mathbf{C}_i)\|^2 + \lambda\|\mathbf{W}\|^2, \tag{7}$$

where $\mathbf{Z}_i$ denotes the i-th training sample in the training set S_{train}, $\mathbf{C}_i$ represents the corresponding center of target bounding box $\mathbf{B}_i$, λ is the regularization parameter. The discriminative loss $l(s, c)$ is a combination of both regression and hinge losses:

$$l(s, c) = v_c \cdot (m_c s + (1 - m_c)max(0, s) - y_c), \tag{8}$$

where m_c, y_c and v_c denote the target mask, the regression label and the spatial weight, respectively. Note that v_c, m_c and y_c are learnable by using the model predictor.

According to Eq. (7), we employ the steepest gradient descent to train the discriminative target model $\mathbf{W}$ in an iterative manner as follows:

$$\mathbf{W}^{j+1} = \mathbf{W}^j - \gamma \bigtriangledown \mathcal{L}(\mathbf{W}^j). \tag{9}$$

For fast convergence, the model predictor calculates an adaptive learning rate γ as follows:

$$\gamma = \frac{\bigtriangledown\mathcal{L}(\mathbf{W}^j)^T \bigtriangledown \mathcal{L}(\mathbf{W}^j)}{\bigtriangledown\mathcal{L}(\mathbf{W}^j)^T Q^j \bigtriangledown \mathcal{L}(\mathbf{W}^j)}, \tag{10}$$

where $Q^j = (J^j)^T(J^j)$, J^j denotes the Jacobian of the residuals at $\mathbf{W}^j$. As the optimization of $\mathbf{W}^j$ is fully differentiable, thus the model predictor is end-to-end trainable in our network.

After obtaining the optimized target model $\mathbf{W} \in \mathbb{R}^{K\times K\times C}$, it is convolved with the test feature $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ to generate the spatio-temporal discriminative map $\mathbf{R}_{dis} \in \mathbb{R}^{H\times W}$ as follows:

$$\mathbf{R}_{dis} = \mathbf{W} * \mathbf{X}. \tag{11}$$

This online discrimination target model ω has superior capability to discriminate new targets from similar objects. It classifies the test feature into the foreground target and the background region. The spatio-temporal discriminative map of the discrimination branch compensates for the spatio-temporal similarity map of the similarity branch.

D. Adaptive Response Map Fusion

As described in Sec. III-B and Sec. III-C, we can obtain the similarity response map from the similarity learning branch and the discriminative response map from the discrimination learning branch. In this subsection, we introduce the third core component of our STSDL (i.e., an adaptive response map fusion module) to integrate the similarity map and the discriminative map. How to fuse the response maps in an ensemble manner has been developed in some works [36], [65], [66]. The fusion schemes contain maximum value, average peak-to-correlation energy and peak-to-sidelobe ratio, but these handcrafted criteria may not be suitable for the proposed framework.

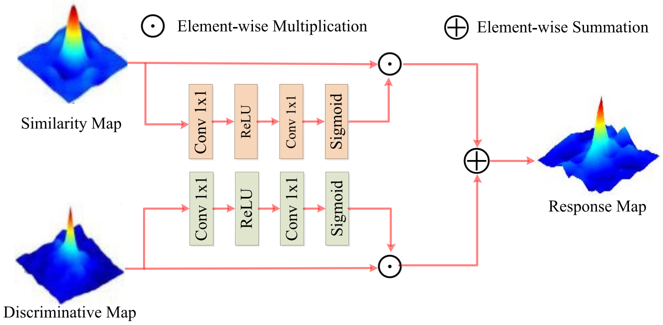


Fig. 6. Overview of the adaptive response map fusion module. Our response map fusion module can adaptively integrate the similarity map and the discriminative map into the final response map.

In this paper, we employ multi-layer perception (MLP) layers to estimate the quality of the predicted similarity map and discriminative map. As illustrated in Fig. 6, we firstly employ a MLP layer to estimate the weight map, and then we conduct element-wise multiplication between the similarity/discriminative map and the corresponding weight map. Finally, we integrate the similarity response map and the discriminative response map in an adaptive manner. The above calculation process can be mathematically formulated as follows:

$$\mathbf{R}_{fus} = \Lambda_{sim}(\mathbf{R}_{sim}) \odot \mathbf{R}_{sim} + \Lambda_{dis}(\mathbf{R}_{dis}) \odot \mathbf{R}_{dis}, \quad (12)$$

where Λ_{sim} and Λ_{dis} respectively denote the multi-layer perception function for similarity map and discriminative map. $\Lambda_{sim}/\ \Lambda_{dis}$ contains two layers, where the first layer is implemented as a 1×1 convolution with a ReLU activation function and the second layer is implemented as a 1×1 convolution with a Sigmoid activation function. $\mathbf{R}_{fus}$ denotes the final response map, which encodes the complementary similarity information and discrimination information for accurate target localization.

E. Offline Training

Our network, which is composed of a feature extractor, a transformer, a few-short learner and an adaptive response map fusion module, is end-to-end trainable by loss minimization. Given a test sample and its corresponding ground-truth labels $(\mathbf{X}, \mathbf{R}_{test}) \in S_{test}$, we construct extra classification losses $\mathcal{L}_{sim}$, $\mathcal{L}_{dis}$ and $\mathcal{L}_{fus}$ based on the response map $\mathbf{R}_{sim}$, $\mathbf{R}_{dis}$ and $\mathbf{R}_{fus}$ as follows:

$$\begin{aligned}\mathcal{L}_{sim} &= \sum_{(\mathbf{X},\mathbf{R}_{test})\in S_{test}} \|r(\mathbf{R}_{sim}, \mathbf{R}_{test})\|^2, \\ \mathcal{L}_{dis} &= \sum_{(\mathbf{X},\mathbf{R}_{test})\in S_{test}} \|r(\mathbf{R}_{dis}, \mathbf{R}_{test})\|^2, \\ \mathcal{L}_{fus} &= \sum_{(\mathbf{X},\mathbf{R}_{test})\in S_{test}} \left\|r(\mathbf{R}_{fus}, \mathbf{R}_{test})\right\|^2, \end{aligned} \quad (13)$$

where $r(\cdot,\cdot)$ denotes the classification loss as defined in DiMP [13].

Algorithm 1 Training Algorithm of Our STSDL

```
input : The test frame $\mathbf{F}_{test} = \{F_k\}$ and the reference
        frames $\mathbf{F}_{ref} = \{F_{k-1}, F_{k-2}, \cdots, F_{k-N}\}$; The
        bounding box of the test frame $\mathbf{B}_{test} = \{B_k\}$
        and the bounding box of the reference frames
        $\mathbf{B}_{ref} = \{B_{k-1}, B_{k-2}, \cdots, B_{k-N}\}$.
output: The final training loss $\mathcal{L}_{final}$.
Extract backbone features of the reference frames $\boldsymbol{Z}$
  and backbone features of the test frame $\boldsymbol{X}$ by using a
  shared feature extractor;
Generate ground-truth labels of the reference frames
  $\boldsymbol{R}_{ref}$ and ground-truth labels of the test frame $\boldsymbol{R}_{test}$
  by using a label generator based on $\boldsymbol{B}_{ref}$ and $\boldsymbol{B}_{test}$;
Feed $\boldsymbol{Z}$, $\boldsymbol{R}_{ref}$ and $\boldsymbol{X}$ into a lightweight transformer for
  spatio-temporal similarity learning to produce a
  similarity map $\boldsymbol{R}_{sim}$ by using Eq. (2)-(6) in Sec. III-B;
Feed $\boldsymbol{Z}$, $\boldsymbol{B}_{ref}$ and $\boldsymbol{X}$ into an efficient few-shot learner
  for spatio-temporal discrimination learning to obtain
  a discriminative map $\boldsymbol{R}_{dis}$ by using Eq. (7)-(11) in
  Sec. III-C;
Feed $\boldsymbol{R}_{sim}$ and $\boldsymbol{R}_{dis}$ into an adaptive response map
  fusion module to produce a final response map $\boldsymbol{R}_{fus}$
  by using Eq. (12) in Sec. III-D;
Employ $\boldsymbol{R}_{sim}$, $\boldsymbol{R}_{dis}$, $\boldsymbol{R}_{fus}$ and $\boldsymbol{R}_{test}$ to compute the
  training loss $\mathcal{L}_{final}$ to train our network by using
  Eq. (13)-(14) in Sec. III-E.
```

To end-to-end train the feature extractor, the transformer, the few-short learner and the adaptive response map fusion module in our joint learning framework, we gather the classification loss of similarity map $\mathcal{L}_{sim}$, the classification loss of discriminative map $\mathcal{L}_{dis}$, the classification loss of fusion map $\mathcal{L}_{fus}$, and the regression loss $\mathcal{L}_{reg}$ together to formulate the final objective function, as follows:

$$\mathcal{L}_{final} = \mathcal{L}_{dis} + \mathcal{L}_{sim} + \mathcal{L}_{fus} + \mu\mathcal{L}_{reg}, \quad (14)$$

where μ is a weighting parameter, $\mathcal{L}_{reg}$ denotes the probabilistic regression loss defined in PrDiMP [14]. The training algorithm of the proposed STSDL is illustrated in Algorithm 1.

F. Online Inference

After network training, the parameters of our network are fixed. In the online inference stage, the proposed STSDL is similar to DiMP [13] and PrDiMP [14]. Given the initial target state (i.e., annotated bounding box), the similarity learning branch uses the transformer encoder-decoder to predict the similarity map. In parallel, the discrimination learning branch uses the model predictor to predict the discriminative map. After prediction, the response map fusion module adaptively gathers the similarity map and discriminative map into the response map. The location corresponding to the maximum value in the response map is the target center. After target localization, an IoU predictor is further applied to predict the target scale. The historical templates are conservatively updated in the template set, which is leveraged by the transformer encoder-decoder and the efficient model predictor for target localization.

IV. Experiments

In this section, we firstly describe implementation details in Sec. IV-A. Secondly, we present datasets and metrics for evaluation in Sec. IV-B. Then, we provide quantitative and qualitative comparisons on six challenging datasets (i.e., GOT10K [67], LaSOT [68], TrackingNet [69], UAV123 [70], OTB2015 [71] and VOT2020 [72]) in Sec. IV-C. Finally, we perform ablation study to demonstrate that the proposed STSDL framework is effective in Sec. IV-D.

A. Implementation Details

Following DiMP [13] and PrDiMP [14], we employ the COCO [73], TrackingNet [69], LaSOT [68] and GOT10K [67] datasets for network training. Specifically, we employ an ADAM optimizer to train our STSDL for 50 epochs, where the batch size is set to 20, the learning rate is initialized to 0.01, and the decay factor is set to 0.2. In Eq. (14), we set the weighting parameter μ to 100. For the MLP layers in our adaptive response map fusion module, the neuron number of the hidden layer is set to 128. The proposed STSDL is implemented using the PyTorch framework in Python, and it runs at a real-time speed of 50 FPS by using ResNet-50 as feature extractor on an NVIDIA GeForce RTX 2080Ti GPU with 12G memory.

B. Datasets and Metrics

In this subsection, we present the datasets and metrics to evaluate the proposed STSDL.

GOT10K [67] is a public tracking database which covers over 560 classes of moving objects in the wild. It offers around 10,000 video segments with a total of over 1.5 million annotated boxes. As the classes in the test set have no overlap with the classes in the train set, the database is appropriate for evaluating the generalization of class-agnostic deep trackers for short-term tracking. There are 180 video segments on the test set, which chooses the average overlap (AO), success rate at an overlap threshold of 0.50 ($SR_{0.5}$) and success rate at an overlap threshold of 0.75 ($SR_{0.75}$) as its evaluation metrics. AO indicates the average overlap between ground-truth and predicted boxes. SR denotes the percentage of successfully tracked frames, where the overlap is no less than a predefined threshold.

TrackingNet [69] is a large-scale tracking database that covers a wide range of object classes. It offers 30000 video segments with around 14 million annotated boxes for both training and evaluation. There are 511 video segments in the test set, which correspond to diverse tracking scenarios. The test set chooses the distance precision at a threshold of 20 pixels (DP@20), normalized distance precision at a threshold of 20 pixels (N.DP@20) and area under curve (AUC) of success plot as its evaluation metrics. DP@20 measures the percentage of successfully tracked frames, where the distance between ground-truth and predicted target centers is no more than 20 pixels. N.DP@20 denotes the percentage of successfully tracked frames, where the distance normalized by the size of ground-truth boxes is no more than 20 pixels. AUC represents the average overlap between ground-truth and predicted boxes, which is equivalent to the AO metric adopted by the GOT10K test set.

LaSOT [68] is a large-scale public tracking database with various challenges stemming from the wild. It provides 1,400 videos with more than 3.5 million frames, where the test set consists of 280 videos. The average video length is over 2500 frames, which far exceeds the average length of videos in other datasets. Therefore, the database is suitable for training and evaluating the deep trackers for long-term tracking. The test set also adopts DP@20, N.DP@20, and AUC as its evaluation metrics.

UAV123 [70] is an aerial public database containing 123 videos, which are collected by an UAV platform. The UAV123 dataset uses DP@20 in precision plot and AUC of success plot as its test metrics.

OTB2015 [71] is a widely-used public database containing 100 videos with various challenging attributes in visual tracking community. We report the precision plot and success plot on the OTB2015 dataset. The success plot shows the percentage of successfully tracked frames, where the overlap between the ground-truth and predicted boxes is no less than a predefined threshold. The precision plot depicts the percentage of successfully tracked frames, where the distance between the ground-truth and predicted centers is no more than a predefined threshold. The OTB2015 dataset adopts DP@20 in precision plot and AUC of success plot as its test metrics.

VOT2020 [72] is a benchmark database consisting of 60 video segments with rotated boxes, and it adopts a reset-based methodology for evaluation. The VOT2020 dataset employs the expected average overlap (EAO), which combines the overlap ratio (accuracy (A)) and the re-initialization times (robustness (R)), as its primary test metric.

C. State-of-the-Art Comparisons

1) Quantitative Results on GOT10K: In Table II, we report the comparison results of the proposed STSDL and eleven competing trackers, including five similarity learning trackers (i.e., SiamRPN++ [7], SiamFC++ [8], SiamCAR [9], SiamGAT [10] and STMTrack [11]) and six discrimination learning trackers (i.e., ATOM [12], DiMP [13], PrDiMP [14], DCFST [16], KYS [40] and DET [43]), on the GOT10K test set to test the performance of these twelve class-agnostic deep trackers for short-term tracking.

As reported in Table II, our STSDL obtains the best tracking results with an AO score of 68.5%, a SR0.5 score of 80.0% and a SR0.75 score of 59.5% on the 180 diverse video segments. Compared with the discrimination learning methods (i.e., ATOM, DiMP, PrDiMP, DCFST, KYS, DET), our STSDL achieves relative performance gains ranging from 6.2%/9.6%/7.4% to 26.2%/48.0%/23.2% in terms of the SR0.5/SR0.75/AO score. This indicates the effectiveness of similarity learning branch in our framework to aggregate rich spatio-temporal cues for performance improvement. In comparison with the similarity learning methods (i.e., SiamRPN++, SiamFC++, SiamCAR, SiamGAT, STMTrack), our STSDL also obtains relative performance gains ranging from 7.7%/3.5%/6.7% to 29.9%/83.1%/32.5% on the

TABLE II

The AO, $SR_{0.5}$ and $SR_{0.75}$ Scores Achieved by Twelve Competing Trackers on the GOT10K Test Set. The First, Second and Third Highest Scores are Marked by **Red**, Blue and *Green*, Respectively

	SiamRPN++	SiamFC++	SiamCAR	SiamGAT	STMTrack	ATOM	DiMP	PrDiMP	DCFST	KYS	DET	STSDL
$SR_{0.5}$ (%) ↑	61.6	69.5	67.0	74.3	73.7	63.4	71.7	73.8	75.3	*75.1*	74.7	**80.0**
$SR_{0.75}$ (%) ↑	32.5	47.9	41.5	48.4	57.5	40.2	49.2	*54.3*	49.8	51.5	50.4	**59.5**
AO (%) ↑	51.7	59.5	56.9	62.7	64.2	55.6	61.1	63.4	*63.8*	63.6	63.0	**68.5**

TABLE III

The DP@20, N.DP@20 and AUC Scores Obtained by Twelve Competing Trackers on the TrackingNet Test Set. The First, Second and Third Highest Scores are Denoted by **Red**, Blue and *Green*, Respectively

	SiamRPN++	SiamFC++	SiamCAR	SiamGAT	STMTrack	ATOM	DiMP	PrDiMP	DCFST	KYS	DET	STSDL
DP@20 (%) ↑	69.4	70.5	68.7	69.8	**76.7**	64.8	68.7	*70.4*	70.0	68.8	70.3	72.8
N.DP@20 (%) ↑	80.0	80.0	80.4	80.7	**85.1**	77.1	80.1	*81.6*	80.9	80.0	81.0	83.0
AUC (%) ↑	73.3	75.4	74.0	75.3	**80.3**	70.3	74.0	*75.8*	75.2	74.0	75.5	78.0

SR0.5/SR0.75/AO metric. This indicates that the discrimination learning branch in our framework is effective to discriminate target objects from background distractors for performance improvement.

2) Quantitative Results on TrackingNet: Table III reports the evaluation results of the proposed STSDL and eleven competing trackers (i.e., SiamRPN++ [7], SiamFC++ [8], SiamCAR [9], SiamGAT [10], STMTrack [11], ATOM [12], DiMP [13], PrDiMP [14], DCFST [16], KYS [40], DET [43]) on the TrackingNet test set to evaluate the performance of the twelve trackers in large-scale diverse tracking scenarios.

As reported in Table III, the proposed STSDL achieves the second best performance on the TrackingNet test set. To be specific, our STSDL achieves a DP@20 score of 72.8%, a N.DP@20 score of 83.0%, and an AUC score of 78.0%. In particular, our STSDL, which integrates the similarity learning and discrimination learning into a unified framework, outperforms most similarity learning approaches (i.e., SiamRPN++, SiamFC++, SiamCAR, SiamGAT) and discrimination learning approaches (i.e., ATOM, DiMP, PrDiMP, DCFST, KYS, DET). This demonstrates that our STSDL is more effective than those approaches to deal with various challenges in large-scale diverse tracking scenarios. The favorable performance of our STSDL can be ascribed to its capability to exploit the complementary merits of these two types of approaches. Our STSDL outperforms the best discrimination learning approach (i.e., PrDiMP) by 3.4%/1.7%/2.9% on the DP@20/N.DP@20/AUC metric, and it is only inferior to the best similarity learning approach (i.e., STMTrack). However, STMTrack employs a space-time memory network to store more reference frames to perform pixel-level feature matching, which requires large memory and computation resources. In contrast, our STSDL uses a lightweight transformer to gather rich spatio-temporal cues, which is memory-efficient and time-efficient.

3) Quantitative Results on LaSOT: To evaluate the deep trackers for long-term tracking, we compare our STSDL with five similarity learning methods (i.e., SiamRPN++ [7], SiamCAR [9], SiamGAT [10], AutoMatch [74], STMTrack [11]), three discrimination learning methods (i.e., ATOM [12], DiMP [13], PrDiMP [14]) and two long-term tracking methods (i.e., GlobalTrack [75], LTMU [76]) on the LaSOT test set. Fig. 7 illustrates the tracking performance of these eleven competitors.

As shown in Fig. 7, our STSDL attains the best tracking performance in comparison with the other ten competing trackers. To be more specific, our STSDL achieves a DP score of 66.2%, a N.DP@20 score of 72.7%, and an AUC score of 64.0% on the LaSOT test set. On one side, the proposed STSDL outperforms the best similarity learning method (i.e., STMTrack) by 4.6%/4.9%/5.6% on the DP@20/N.DP@20/AUC metric. The possible reason is that our STSDL contains the spatio-temporal discrimination learning to effectively discriminate between target objects and background clutters. On the other side, our STSDL is also superior to the best discrimination learning method (i.e., PrDiMP) with considerable performance gains (i.e., 8.2%/4.9%/6.1% on the DP@20/N.DP@20/AUC metric). The superior performance benefits from the spatio-temporal similarity learning branch in our STSDL to preserve the rich spatio-temporal context information. Moreover, our STSDL also significantly outperforms the two representative long-term tracking methods (i.e., GlobalTrack and LTMU). The evaluation results on the LaSOT test set validate that our STSDL is effective to handle diverse long video sequences in the wide.

Quantitative results on fourteen attributes: Besides the overall performance, we also provide the quantitative results of these competing trackers on fourteen challenging attributes. Fig. 8 depicts the success plots of our STSDL and the other ten competitors. As reported in Fig. 8, the best similarity learning method (i.e., STMTrack) achieves the second best performance on the attributes of deformation, rotation, scale variation and illumination variation. This is because that STMTrack employs a space-time memory network to aggregate rich spatio-temporal cues to obtain accurate target bounding boxes in the case of deformation, rotation, scale

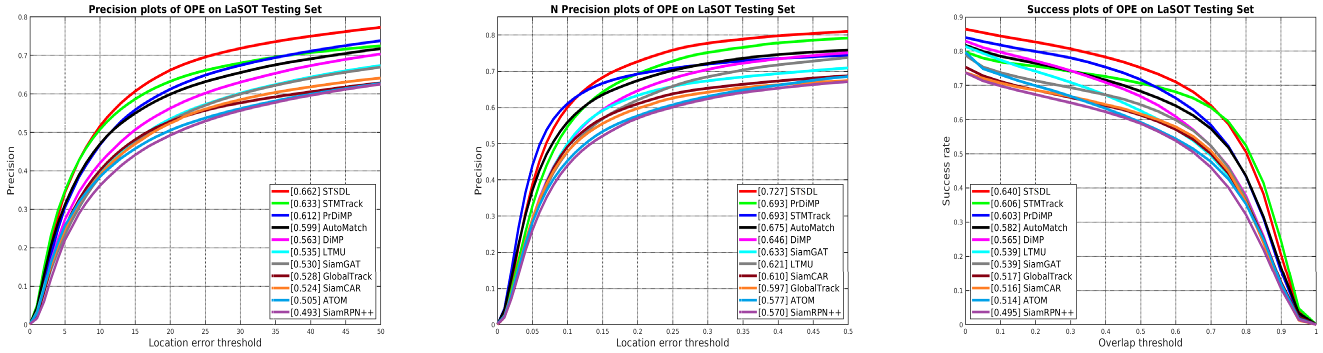


Fig. 7. Precision, normalized precision and success plots of eleven competitors on the LaSOT test set.

TABLE IV

The DP@20 and AUC Scores Obtained by Twelve Competing Trackers on UAV123. The First, Second and Third Highest Scores Are Denoted by **Red**, Blue and Green, Respectively

	SiamRPN++	SiamFC++	SiamCAR	SiamGAT	STMTrack	ATOM	DiMP	PrDiMP	CARE	DET	STSDL
DP@20 (%) ↑	84.0	80.4	83.9	84.3	84.4	82.7	*84.9*	87.2	-	-	**90.0**
AUC (%) ↑	64.2	61.8	64.0	64.6	64.7	61.7	64.2	66.6	64.6	*66.4*	**68.0**

variation and illumination variation. The best discrimination learning method (i.e., PrDiMP) achieves the second best AUC score on the attributes of background clutter, camera motion, fast motion, motion blur, partial occlusion, full occlusion, out-of-view and low resolution. The reason is that PrDiMP uses a few-shot learner to train discriminative target models to search for the discriminative target objects in the case of background clutter, camera motion, fast motion, motion blur, partial occlusion, full occlusion, out-of-view and low resolution. However, both STMTrack and PrDiMP are inferior to the proposed STSDL in terms of the AUC score on the fourteen attributes. For the attribute of viewpoint change, it is necessary to accurately estimate target bounding boxes and robustly discriminate between target objects and surrounding backgrounds. On this attribute, our STSDL outperforms STMTrack and PrDiMP by a large margin (i.e., 15.3% and 14.9%) on the AUC metric. The reason is that our STSDL can benefit from both similarity learning to gather rich spatio-temporal context information and discrimination learning to discriminate target objects from background clutters. The attribute based comparison further shows the robustness of the proposed STSDL for long-term tracking.

4) Quantitative Results on UAV123: To test the effectiveness of our STSDL under aerial tracking scenarios, we compare our STSDL with five similarity learning approaches (i.e., SiamRPN++ [7], SiamFC++ [8], SiamCAR [9], SiamGAT [10] and STMTrack [11]) and five discrimination learning approaches (i.e., ATOM [12], DiMP [13], PrDiMP [14], CARE [42] and DET [43]) on the UAV123 dataset. Table IV reports the evaluation results of our STSDL and the other ten competing trackers. In comparison with the ten competitors, our STSDL obtains the best tracking performance. To be more specific, it obtains 90.0% DP@20 score and 68.0% AUC score on this dataset. Note that SiamRPN++, STMTrack, DiMP, PrDiMP, CARE, DET and our STSDL adopt the same backbone (i.e., ResNet50) for feature extraction, the superior performance of our STSDL demonstrates that joint spatio-temporal similarity and discrimination learning is more effective than both similarity learning and discrimination learning for performance improvement. On this dataset, most discrimination learning approaches achieve better performance than similarity learning approaches. This is because that discrimination learning approaches are more effective than similarity approaches to handle the challenges of viewpoint change, fast motion, camera motion, background clutter, similar object, partial occlusion, full occlusion, out-of-view and low resolution on the 123 aerial video sequences. Nevertheless, our STSDL surpasses the best discrimination learning approach (i.e., PrDiMP) by 3.2%/2.1% on the DP@20/AUC metric.

5) Quantitative Results on OTB2015: In Fig. 9, we compare our STSDL with three similarity learning methods (i.e., SiamRPN++ [7], SiamCAR [9], SiamGAT [10]) and three discrimination learning methods (i.e., ATOM [12], DiMP [13], PrDiMP [14]) on the OTB2015 dataset. As displayed in Fig. 9, our STSDL obtains 93.1% and 71.0% on the DP@20 and AUC metrics, respectively. In comparison with both similarity learning approaches and discrimination learning approaches, our STSDL exhibits better performance while operating at a real-time running speed. On this dataset, most similarity learning approaches obtain better performance than discrimination learning approaches. The reason is that similarity learnig approaches have advantage over discrimination learning approaches to cope with the challenges of deformation, in-plane rotation, out-of-plane rotation, scale variation and illumination variation on the 100 video sequences. However, the best similarity approach (i.e., SiamGAT) is still inferior to our STSDL on the DP@20 metric.

Qualitative results on OTB2015: Besides the above quantitative comparisons, we also provide some qualitative

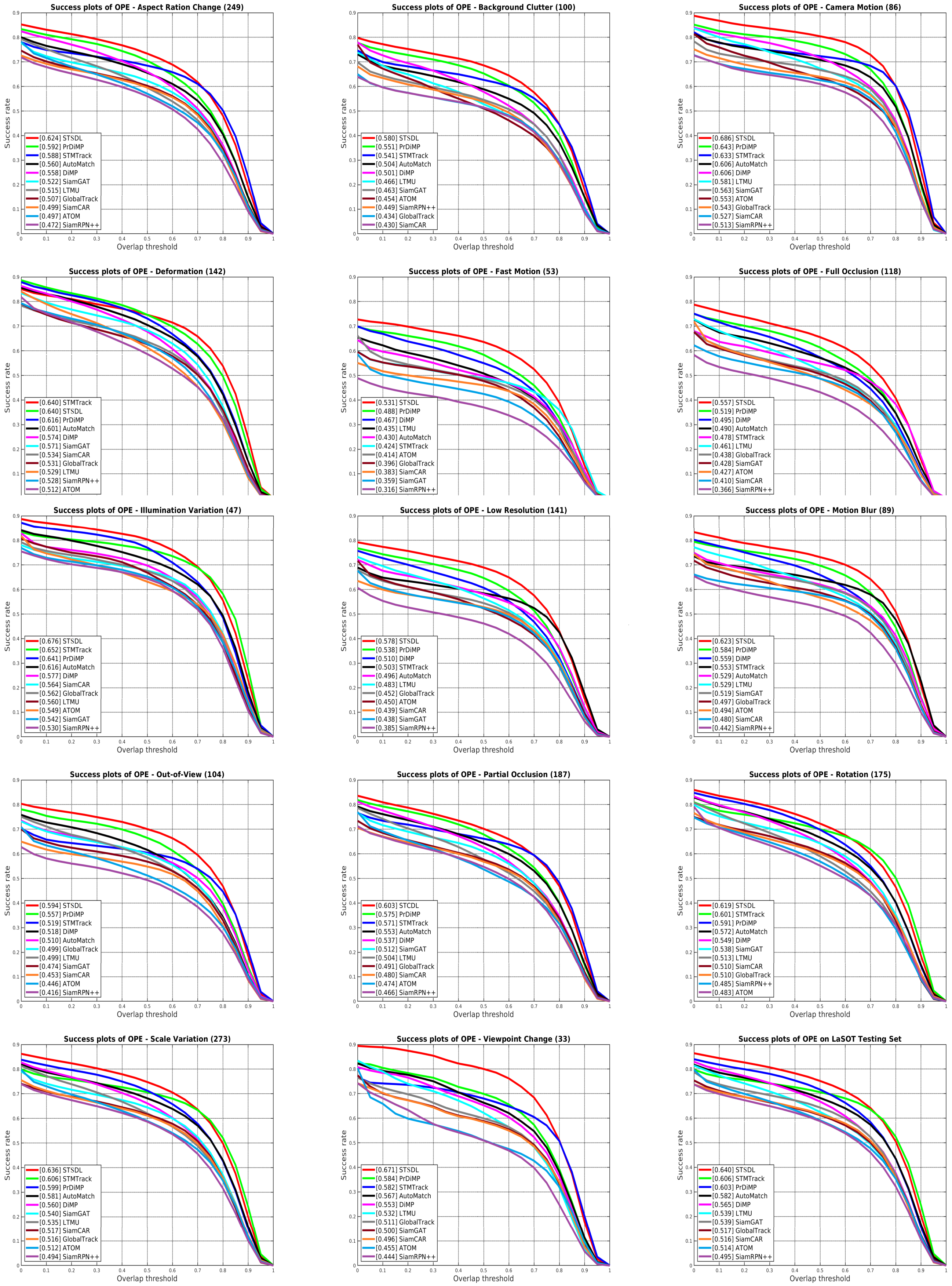


Fig. 8. Attribute based comparison on the LaSOT test set. We illustrate the success plots of eleven competing trackers on fourteen attributes. For clarity, we further illustrate the overall performance of the eleven competitors.

TABLE V

THE EAO, ACCURACY (A) AND ROBUSTNESS (R) SCORES ACHIEVED BY TWELVE COMPETING TRACKERS ON VOT2020. THE FIRST, SECOND AND THIRD HIGHEST SCORES ARE DENOTED BY **RED**, BLUE AND *GREEN*, RESPECTIVELY

	ToMP101	ToMP50	STARK101	STARK50	CSWinTT	TransT	UPDT	SiamCAR	DiMP	ATOM	SiamRPN++	STSDL
EAO ↑	0.309	0.297	0.303	*0.308*	0.304	0.293	0.278	0.273	0.274	0.271	0.244	**0.312**
A ↑	0.453	0.453	**0.481**	0.478	0.480	0.477	0.465	0.449	0.457	0.462	0.443	*0.479*
R ↑	**0.814**	0.789	0.775	0.799	0.787	0.754	0.755	0.732	0.740	0.734	0.672	0.799

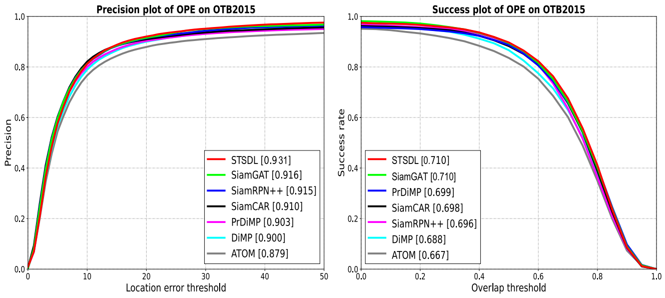


Fig. 9. Precision and success plots of our STSDL and six participants on OTB2015.

comparisons of our STSDL with the best discrimination learning approach (i.e., PrDiMP) and the best similarity learning approach (i.e., SiamGAT) on eight challenging video sequences in Fig. 10. By joint spatio-temporal similarity and discrimination learning, the proposed STSDL can obtain more accurate and robust tracking results than PrDiMP and SiamGAT on these sequences. As shown on the first row in Fig. 10, when the deformed targets are interfered by distractors (i.e., *Skating2-1*, *Skating2-2*), our STSDL can consistently and accurately localize the skaters in the tracking process. In contrast, PrDiMP and SiamGAT will mistakenly drift to the distractors. As illustrated on the second row in Fig. 10, in the cases of large appearance variations and scale variations (i.e., *Board*, *Ironman*), PrDiMP and SiamGAT fail to localize the targets or estimate the target scale. In contrast, our STSDL can successfully and accurately localize the targets over time. As depicted on the third row in Fig. 10, when the targets encounter background clutters (i.e., *Human3*, *Soccer*), our STSDL is able to unceasingly track the targets without failures, whereas PrDiMP and SiamGAT will be disrupted by the background clutters. As observed on the fourth row in Fig. 10, when the target objects and similar distractors are fused together (i.e., *Basketball*, *Girl2*), all the competing trackers cannot achieve satisfactory tracking results. However, when the target objects and similar distractors are separated from each other, our STSDL can successfully recover from failures to obtain more accurate tracking results than PrDiMP and SiamGAT. The qualitative tracking results indicate that our STSDL can accurately and robustly cope with various challenging tracking scenarios.

6) *Quantitative Results on VOT2020:* To evaluate the effectiveness of our STSDL for fine-grained short-term tracking, we compare it with state-of-the-art approaches, including ToMP101 [59], ToMP50 [59], STARK101 [58], STARK50 [58], CSWinTT [77], TransT [57], UPDT [78], SiamCAR [9], DiMP [13], ATOM [12] and SiamRPN++ [7]. Table V reports the evaluation results of our STSDL with eleven state-of-the-art participants on the VOT2020 dataset. Among these competing trackers, the proposed STSDL achieves the highest EAO score (0.312), the second highest R score (0.799), the third highest A score (0.479). In comparison with the discrimination learning approach (i.e., DiMP), our STSDL exhibits a 4.8% higher A score, a 8.0% higher R score and a 13.9% higher EAO score, which indicates the effectiveness of similarity learning branch in our framework for performance improvement. Compared to the similarity learning approach (i.e., SiamCAR), our STSDL boosts the A/R/EAO score by 6.7%/9.2%/14.3%, which validates that the discrimination learning branch in our framework is effective to improve the tracking performance. Moreover, by joint spatio-temporal similarity and discrimination learning, our STSDL outperforms the recent transformer based trackers (e.g., ToMP101, ToMP50, STARK101, STARK50, CSWinTT, TransT) in term of EAO score. The comparison results on the 60 video segments show that our STSDL is also suitable for fine-grained short-term tracking with a reset-based protocol.

D. Ablation Study

We perform ablation study to verify the effectiveness of the proposed STSDL framework. As the GOT10K test set with diverse classes is appropriate to assess the generalization of our STSDL, we choose the GOT10K test set for our ablation studies.

1) *Effectiveness of Spatio-Temporal Similarity Learning and Spatio-Temporal Discrimination Learning:* The proposed framework performs joint spatio-temporal similarity and discrimination learning for visual tracking. The framework consists of two complementary branches: the spatio-temporal similarity learning (STSL) branch and the spatio-temporal discrimination learning (STDL) branch. To test the impacts of the STSL branch and the STDL branch on the performance of our framework, we respectively remove each of the two branches and denote the variants as STSL and STDL. As reported in Table VI, both the STSL branch and the STDL branch are conductive to boost the tracking performance on the GOT10K test set, which verifies the complementarity of similarity learning and discrimination learning. To be more specific, the spatio-temporal similarity learning branch of our framework yields 2.9%/2.4%/4.6% performance gains in terms of AO/$SR_{0.5}$/$SR_{0.75}$ scores. In contrast, the spatio-temporal discrimination learning branch of our framework results in

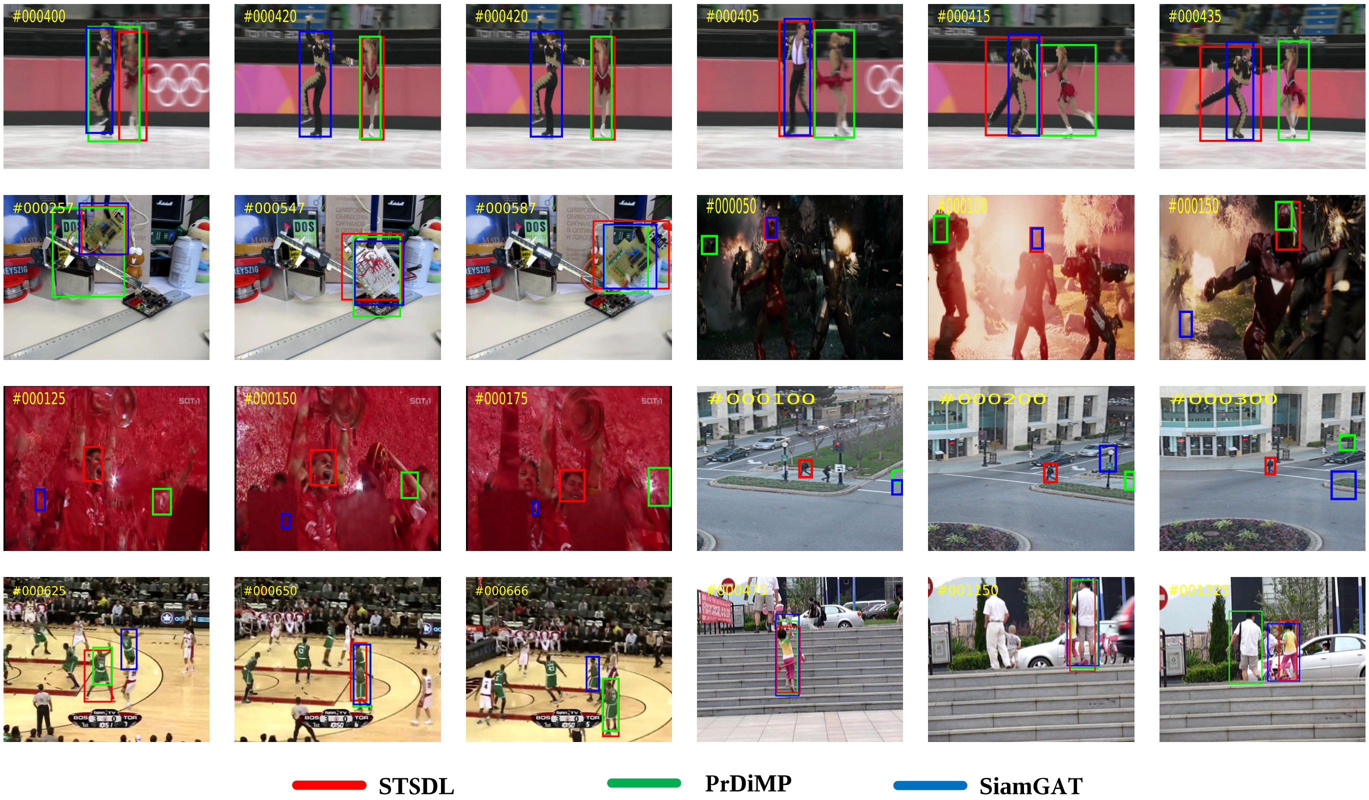


Fig. 10. Comparisons of our STSDL with PrDiMP and SiamGAT on the eight videos from OTB2015. The six videos from left to right and top to bottom are *Skating2-1*, *Skating2-2*, *Board*, *Ironman*, *Soccer*, *Human3*, *Basketball* and *Girl2*, respectively.

TABLE VI

EFFECTIVENESS OF SPATIO-TEMPORAL SIMILARITY LEARNING AND SPATIO-TEMPORAL DISCRIMINATION LEARNING ON THE GOT10K TEST SET. THE HIGHEST SCORES ARE HIGHLIGHTED BY **BOLD**

	STSL branch	STDL branch	AO	$SR_{0.5}$	$SR_{0.75}$
STSL	✓		65.4	76.6	57.1
STDL		✓	66.6	78.1	56.9
STSDL	✓	✓	**68.5**	**80.0**	**59.5**

TABLE VII

EFFECTIVENESS OF MULTIPLE CLASSIFICATION LOSSES ON THE GOT10K TEST SET. THE HIGHEST SCORES ARE HIGHLIGHTED BY **BOLD**

	$\mathcal{L}_{dis}$	$\mathcal{L}_{sim}$	$\mathcal{L}_{fus}$	AO	$SR_{0.5}$	$SR_{0.75}$
baseline	✓			67.2	78.5	57.9
w/o $\mathcal{L}_{fus}$	✓	✓		67.8	79.4	58.2
STSDL	✓	✓	✓	**68.5**	**80.0**	**59.5**

4.7%/4.4%/4.2% performance gains on the AO/$SR_{0.5}$/$SR_{0.75}$ metrics.

2) Effectiveness of Multiple Classification Losses: The proposed framework adopts multiple classification losses for network training. In Table VII, we test the effectiveness of the classification loss of similarity map $\mathcal{L}_{sim}$, the classification loss of discriminative map $\mathcal{L}_{dis}$ and the classification loss of fusion map $\mathcal{L}_{fus}$. In comparison with the tracking method with only $\mathcal{L}_{dis}$ (i.e., "baseline" in Table VII), the tracking method with both $\mathcal{L}_{dis}$ and $\mathcal{L}_{sim}$ (i.e., "w/o $\mathcal{L}_{fus}$" in Table VII) exhibits 0.6% higher AO score, 0.9% higher $SR_{0.5}$ score, 0.3% higher $SR_{0.75}$ score. This indicates that $\mathcal{L}_{sim}$ can result in satisfactory performance gains. From Table VII, compared to the tracking method without $\mathcal{L}_{fus}$ (i.e., "w/o $\mathcal{L}_{fus}$" in Table VII), our STSDL with $\mathcal{L}_{fus}$ achieves 0.7% higher AO score, 0.6% higher $SR_{0.5}$ score, 1.3% higher $SR_{0.75}$ score. This verifies that $\mathcal{L}_{fus}$ is also beneficial to boost the tracking performance.

TABLE VIII

EFFECTIVENESS OF ADAPTIVE FUSION MODULE ON THE GOT10K TEST SET. THE HIGHEST SCORES ARE HIGHLIGHTED BY **BOLD**

	Adaptive fusion	AO	$SR_{0.5}$	$SR_{0.75}$
w/o $\mathcal{M}_{fus}$		66.2	77.5	57.6
STSDL	✓	**68.5**	**80.0**	**59.5**

3) Effectiveness of Adaptive Fusion Module: The proposed framework adopts an adaptive fusion module for response map fusion. In Table VIII, we test the effectiveness of our adaptive fusion module by removing it from our framework (i.e., we directly fuse the similarity map and the discriminative map by element-wise summation). As shown in Table VIII, the proposed STSDL using our adaptive fusion module outperforms the tracking method without using our adaptive fusion module (i.e., "w/o $\mathcal{M}_{fus}$" in Table VIII) by 3.5%/3.2%/3.3% in terms of the AO/$SR_{0.5}$/$SR_{0.75}$ scores on the GOT10K test set.

V. Conclusion

In this work, we develop a new framework to perform joint spatio-temporal similarity and discrimination learning for visual tracking. Our framework consists of two complementary branches: a spatio-temporal similarity learning branch and a spatio-temporal discrimination learning branch. The similarity learning branch is responsible for rich spatio-temporal information propagation by using a transformer encoder-decoder, predicting the similarity response map. In contrast, the discrimination learning branch is eligible for discriminative target model prediction by using a model predictor, producing the discriminative response map. Moreover, these two complementary response maps are adaptively fused for accurate target localization. Quantitative and qualitative tracking results on six prevalent datasets show that our tracking method can attain favorable performance in comparison with both similarity learning methods and discrimination learning methods, and it is with a real-time tracking speed of 50 FPS on a single GPU. The proposed STSDL can accurately and robustly extract a small portion of valuable object-level information from massive video data for video analysis, which has its great potentials in real-word applications (e.g., video surveillance, autonomous driving, intelligent transportation).

References

[1] A. Bandini and J. Zariffa, "Analysis of the hands in egocentric vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6846–6866, Jun. 2023.

[2] S. A. Ahmed, D. P. Dogra, S. Kar, and P. P. Roy, "Trajectory-based surveillance analysis: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 7, pp. 1985–1997, Jul. 2019.

[3] M. Wang et al., "An online multiobject tracking network for autonomous driving in areas facing epidemic," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 25191–25200, Dec. 2022.

[4] B. Ramesh et al., "E-TLD: Event-based framework for dynamic object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3996–4006, Oct. 2021.

[5] P. Li, D. Wang, L. Wang, and H. Lu, "Deep visual tracking: Review and experimental comparison," *Pattern Recognit.*, vol. 76, pp. 323–338, Apr. 2018.

[6] S. M. Marvasti-Zadeh, L. Cheng, H. Ghanei-Yakhdan, and S. Kasaei, "Deep learning for visual tracking: A comprehensive survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 3943–3968, May 2022.

[7] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of Siamese visual tracking with very deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4282–4291.

[8] Y. Xu, Z. Wang, Z. Li, Y. Yuan, and G. Yu, "SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 12549–12556.

[9] Y. Cui et al., "Joint classification and regression for visual tracking with fully convolutional Siamese networks," *Int. J. Comput. Vis.*, vol. 130, pp. 550–566, Jan. 2022.

[10] D. Guo, Y. Shao, Y. Cui, Z. Wang, L. Zhang, and C. Shen, "Graph attention tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9543–9552.

[11] Z. Fu, Q. Liu, Z. Fu, and Y. Wang, "STMTrack: Template-free visual tracking with space-time memory networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13769–13778.

[12] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: Accurate tracking by overlap maximization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4660–4669.

[13] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6182–6191.

[14] M. Danelljan, L. Van Gool, and R. Timofte, "Probabilistic regression for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7183–7192.

[15] N. Wang, W. Zhou, J. Wang, and H. Li, "Transformer meets tracker: Exploiting temporal context for robust visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1571–1580.

[16] L. Zheng, M. Tang, Y. Chen, J. Wang, and H. Lu, "Learning feature embeddings for discriminant model based tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 759–775.

[17] Y. Shi, Z. Wei, H. Ling, Z. Wang, J. Shen, and P. Li, "Person retrieval in surveillance videos via deep attribute mining and reasoning," *IEEE Trans. Multimedia*, vol. 23, pp. 4376–4387, 2021.

[18] Y. Shi et al., "Adaptive and robust partition learning for person retrieval with policy gradient," *IEEE Trans. Multimedia*, vol. 23, pp. 3264–3277, 2021.

[19] L. Wu et al., "Pseudo-pair based self-similarity learning for unsupervised person re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 4803–4816, 2022.

[20] D. Liu et al., "Generative metric learning for adversarially robust open-world person re-identification," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 19, no. 1, 2023, Art. no. 20.

[21] Q. Qi, T. Hou, Y. Yan, Y. Lu, and H. Wang, "TCNet: A novel triple-cooperative network for video object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 3649–3662, Aug. 2023.

[22] L. Han and Z. Yin, "Global memory and local continuity for video object detection," *IEEE Trans. Multimedia*, vol. 25, pp. 3681–3693, 2023.

[23] S. W. Oh, J. Lee, N. Xu, and S. J. Kim, "Space-time memory networks for video object segmentation with user guidance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 442–455, Jan. 2022.

[24] J. Fan, B. Liu, K. Zhang, and Q. Liu, "Semi-supervised video object segmentation via learning object-aware global-local correspondence," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8153–8164, Dec. 2022.

[25] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[26] Y. Liang, Q. Wu, Y. Liu, Y. Yan, and H. Wang, "Correlation filter tracking with shepherded instance-aware proposals," in *Proc. 26th ACM Int. Conf. Multimedia (MM)*, Oct. 2018, pp. 420–428.

[27] K. Dai, D. Wang, H. Lu, C. Sun, and J. Li, "Visual tracking via adaptive spatially-regularized correlation filters," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4670–4679.

[28] L. Xiong, Y. Liang, Y. Yan, and H. Wang, "Correlation filter tracking with adaptive proposal selection for accurate scale estimation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 1816–1821.

[29] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, "Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5596–5609, Nov. 2019.

[30] K. Nai, Z. Li, and H. Wang, "Learning channel-aware correlation filters for robust object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7843–7857, Nov. 2022.

[31] Y. Liu, Y. Liang, Q. Wu, L. Zhang, and H. Wang, "A new framework for multiple deep correlation filters based object tracking," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 1670–1674.

[32] T. Zhang, C. Xu, and M.-H. Yang, "Learning multi-task correlation particle filters for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 365–378, Feb. 2019.

[33] Y. Liang, Q. Wu, Y. Liu, Y. Yan, and H. Wang, "Deep correlation filter tracking with shepherded instance-aware proposals," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 11408–11421, Aug. 2022.

[34] Y. Liang, Y. Liu, Y. Yan, L. Zhang, and H. Wang, "Robust visual tracking via spatio-temporal adaptive and channel selective correlation filters," *Pattern Recognit.*, vol. 112, Apr. 2021, Art. no. 107738.

[35] F. Li, C. Tian, W. Zuo, L. Zhang, and M. Yang, "Learning spatial-temporal regularized correlation filters for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4904–4913.

[36] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Robust visual tracking via hierarchical convolutional features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2709–2723, Nov. 2019.

[37] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, “ECO: Efficient convolution operators for tracking,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6931–6939.

[38] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, “Joint group feature selection and discriminative filter learning for robust visual object tracking,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7950–7960.

[39] U. Kart, A. Lukežic, M. Kristan, J. Kämäräinen, and J. Matas, “Object tracking by reconstruction with view-specific discriminative correlation filters,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1339–1348.

[40] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, “Know your surroundings: Exploiting scene information for object tracking,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 205–221.

[41] Z. Zhou, X. Li, N. Fan, H. Wang, and Z. He, “Target-aware state estimation for visual tracking,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2908–2920, May 2022.

[42] N. Wang, W. Zhou, Q. Tian, and H. Li, “Cascaded regression tracking: Towards online hard distractor discrimination,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1580–1592, Apr. 2021.

[43] N. Wang, W. Zhou, and H. Li, “Learning diverse models for end-to-end ensemble tracking,” *IEEE Trans. Image Process.*, vol. 30, pp. 2220–2231, 2021.

[44] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, “Fully-convolutional Siamese networks for object tracking,” in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016, pp. 850–865.

[45] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, “High performance visual tracking with Siamese region proposal network,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8971–8980.

[46] S. Yao, X. Han, H. Zhang, X. Wang, and X. Cao, “Learning deep lucas-kanade Siamese network for visual tracking,” *IEEE Trans. Image Process.*, vol. 30, pp. 4814–4827, 2021.

[47] W. Zhou, L. Wen, L. Zhang, D. Du, T. Luo, and Y. Wu, “SiamCAN: Real-time visual tracking based on Siamese center-aware network,” *IEEE Trans. Image Process.*, vol. 30, pp. 3597–3609, 2021.

[48] C. Zhuang, Y. Liang, Y. Yan, Y. Lu, and H. Wang, “Bounding box distribution learning and center point calibration for robust visual tracking,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 4718–4722.

[49] H. Zhang, L. Cheng, T. Zhang, Y. Wang, W. J. Zhang, and J. Zhang, “Target-distractor aware deep tracking with discriminative enhancement learning loss,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 6267–6278, Sep. 2022.

[50] P. Li, B. Chen, W. Ouyang, D. Wang, X. Yang, and H. Lu, “GradNet: Gradient-guided network for visual object tracking,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6161–6170.

[51] H. Fan and H. Ling, “Siamese cascaded region proposal networks for real-time visual tracking,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7944–7953.

[52] Z. Zhang and H. Peng, “Deeper and wider Siamese networks for real-time visual tracking,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4586–4595.

[53] F. Tang and Q. Ling, “Learning to rank proposals for Siamese visual tracking,” *IEEE Trans. Image Process.*, vol. 30, pp. 8785–8796, 2021.

[54] Z. Chen et al., “SiamBAN: Target-aware tracking with Siamese box adaptive network,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 54, no. 4, pp. 5158–5173, Apr. 2023.

[55] Y. Liang, P. Zhao, Y. Hao, and H. Wang, “Siamese template diffusion networks for robust visual tracking,” in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6.

[56] B. Yu et al., “High-performance discriminative tracking with transformers,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9836–9845.

[57] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, “Transformer tracking,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8122–8131.

[58] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, “Learning spatio-temporal transformer for visual tracking,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10428–10437.

[59] C. Mayer et al., “Transforming model prediction for tracking,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8731–8740.

[60] Y. Cui, C. Jiang, L. Wang, and G. Wu, “MixFormer: End-to-end tracking with iterative mixed attention,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13608–13618.

[61] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 213–229.

[62] B. Ye, H. Chang, B. Ma, S. Shan, and X. Chen, “Joint feature learning and relation modeling for tracking: A one-stream framework,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 341–357.

[63] X. Hu et al., “Transformer tracking via frequency fusion,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 2, pp. 1020–1031, Feb. 2024.

[64] L. Lin et al., “Dual semantic fusion network for video object detection,” in *Proc. 28th ACM Int. Conf. Multimedia (MM)*, Oct. 2020, pp. 1855–1863.

[65] N. Wang, W. Zhou, Q. Tian, R. Hong, M. Wang, and H. Li, “Multi-cue correlation filters for robust visual tracking,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4844–4853.

[66] A. Lukezic, T. Vojír, L. C. Zajc, J. Matas, and M. Kristan, “Discriminative correlation filter with channel and spatial reliability,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nov. 2017, pp. 4847–4856.

[67] L. Huang, X. Zhao, and K. Huang, “GOT-10k: A large high-diversity benchmark for generic object tracking in the wild,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1562–1577, May 2021.

[68] H. Fan et al., “LaSOT: A high-quality benchmark for large-scale single object tracking,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5374–5383.

[69] M. Müller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem, “TrackingNet: A large-scale dataset and benchmark for object tracking in the wild,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 310–327.

[70] M. Mueller, N. G. Smith, and B. Ghanem, “A benchmark and simulator for UAV tracking,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 445–461.

[71] Y. Wu, J. Lim, and M.-H. Yang, “Object tracking benchmark,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.

[72] M. Kristan et al., “The eighth visual object tracking VOT2020 challenge results,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 547–601.

[73] T. Lin et al., “Microsoft COCO: Common objects in context,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[74] Z. Zhang, Y. Liu, X. Wang, B. Li, and W. Hu, “Learn to match: Automatic matching network design for visual tracking,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13319–13328.

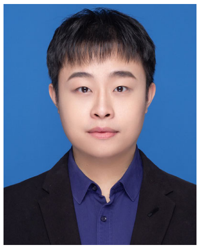
[75] L. Huang, X. Zhao, and K. Huang, “Globaltrack: A simple and strong baseline for long-term tracking,” in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 11037–11044.

[76] K. Dai, Y. Zhang, D. Wang, J. Li, H. Lu, and X. Yang, “High-performance long-term tracking with meta-updater,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6297–6306.

[77] Z. Song, J. Yu, Y.-P. P. Chen, and W. Yang, “Transformer tracking with cyclic shifting window attention,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8791–8800.

[78] B. Goutam, J. Joakim, and D. Martin, “Unveiling the power of deep tracking,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 493–509.

Yanjie Liang received the Ph.D. degree in computer science from the School of Informatics, Xiamen University, Xiamen, China, in 2021. He is currently an Assistant Research Fellow with the Peng Cheng Laboratory. He has published several papers in IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, *Pattern Recognition*, and ACM MM. His current research interests include computer vision, machine learning, and visual tracking.



Haosheng Chen received the B.E. and M.E. degrees in computer science and technology and software engineering from Zhengzhou University, Zhengzhou, China, in 2014 and 2017, respectively, and the Ph.D. degree in computer science and technology from Xiamen University, Xiamen, China, in 2021. He is currently with the College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include pattern recognition, video analysis, and bio-inspired vision.

Qiangqiang Wu is currently pursuing the Ph.D. degree with the Department of Computer Science, City University of Hong Kong, Hong Kong. He has published several papers in IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Industrial Electronics, CVPR, ICCV, AAAI, and ACM MM. His current research interests include computer vision, deep learning, adversarial learning, and visual tracking.

Changqun Xia received the Ph.D. degree from the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China, in 2019. He is currently an Associate Professor with the Peng Cheng Laboratory. He has published several papers in IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, CVPR, ICCV, AAAI, and ACM MM. His research interests include computer vision and image understanding.

Jia Li (Senior Member, IEEE) received the B.E. degree from Tsinghua University in 2005 and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2011. He is currently a Full Professor with the School of Computer Science and Engineering, Beihang University, Beijing, China. Before he joined Beihang University in June 2014, he used to conduct research at Nanyang Technological University, Peking University, and Shanda Innovations. He has been supported by the Research Funds for Excellent Young Researchers from the National Natural Science Foundation of China since 2019. He is the author or coauthor of more than 100 technical articles in refereed journals and conferences, such as IEEE Transactions on Pattern Analysis and Machine Intelligence, *International Journal of Computer Vision*, IEEE Transactions on Image Processing, CVPR, and ICCV. His research interests include computer vision and multimedia big data, especially the understanding and generation of visual content. He is a Fellow of IET, a Distinguished Member of CCF, and a Senior Member of ACM and CIE. He was also selected into the Beijing Nova Program in 2017 and ever received the Second-Grade Science Award of the Chinese Institute of Electronics in 2018, two Excellent Doctoral Thesis Award from the Chinese Academy of Sciences in 2012 and the Beijing Municipal Education Commission in 2012, and the First-Grade Science-Technology Progress Award from the Ministry of Education, China, in 2010.