Leveraging Local and Global Cues for Visual Tracking via Parallel Interaction Network

Yaozong Zheng, Bineng Zhong, Qihua Liang, Zhenjun Tang, *Member, IEEE*, Rongrong Ji, *Senior Member, IEEE*, and Xianxian Li

***Abstract*— Despite that both local and context information are crucial for robust tracking, existing CNN-based and transformer-based methods mainly focus on one of these aspects. Consequently, the former fails to exploit rich global context information due to the limited receptive field, while the latter suffers from the deficiencies in constructing the local relationship among neighboring regions. To address this issue, we propose the SiamPIN tracker, based on our Parallel Interaction Network. It consists of two effective modules, namely Global Aggregation Block (GAB) and Local Process Block (LPB). GAB perceives the global context to capture the long-range spatial dependency through a transformer-based architecture. Meanwhile, LPB performs local information extraction using a CNN model to retain the detailed appearance information of the target. These two modules are connected consecutively to compose a Trans-Conv unit block, which transmits the global context information to the local feature extraction procedure, hence enables the interaction of global-local information flow. Several such blocks are cascaded so that our model can learn to aggregate local and context information interactively. The proposed tracker achieves state-of-the-art performance on six benchmark datasets, while maintaining a real time running speed.**

***Index Terms*— Object tracking, siamese network, vision transformer.**

I. Introduction

GIVEN the initial state of the target in the first frame of a sequence, visual object tracking aims to estimate the state of the target object in subsequent frames. It is a highly challenging task in computer vision due to factors such as deformation, interference from similar objects, occlusion, and appearance changes.

In the past few years, previous research has been dominated by convolutional neural networks (CNNs), due to its high efficiency both in computation and local modeling ability. However, existing CNN-based trackers [1], [2], [3], [4], [5], [6] have been greatly challenged by transformer-based trackers [7], [8], [9], [10], which exhibit advanced tracking performance. It is generally believed that its multi-head attention mechanism plays an important role in the reason why tansformer-based trackers can perform so well. For example, compared with convolutions, prior works believe the multi-head attention mechanism is more robust to distractors [11], and can take advantage of long range dependencies (with larger receptive fields) [12] and so on. On the other hand, convolution still maintains its advantages in terms of local processing, which transformer do not have.

Existing works [13], [14], [15] try to exploit take advantages of transformer and CNN. They extensively explored powerful representation learning methods from different perspectives, however, we focus on leveraging different range level cues for visual tracking, which received little attention in existing works. In contrast to using previous convolution fusion modules [2], [5], [16], [17], recent works [7], [8], [9], [10] usually use original transformer structures to fuse global features of both the template and search frames. As a result, they ignore preserving the fine-grained information of the target at the local level. An interesting exploration is to combine CNN and transformer to capture different granularity of information for visual tracking. As shown in Fig.1(a), convolution operation focuses on extracting local features (yellow box) within the limited receptive field, and transformer easy to capture context information from a lager region by the attention mechanism (light blue box). These two modeling paradigms (CNN with local processing, transformer with global computing) can be naturally combined to capture different types of information.

In this work, we propose a Siamese Parallel Interaction Network for tracking, termed as SiamPIN, which aims to combine CNN and transformer to sense the environment between target and background. The proposed method contains two efficient modules named as Global Aggregation Block (GAB) and Local Process Block (LPB). GAB aims to perceive the global context relationships between the target and background via a transformer architecture, while LPB performs local information extraction to retain the fine-grained detail appearance of the target object by a CNN model. Specifically, LPB and

Manuscript received 12 July 2022; revised 14 September 2022 and 29 September 2022; accepted 2 October 2022. Date of publication 10 October 2022; date of current version 5 April 2023. This work was supported in part by the Project of Guangxi Science and Technology under Grant 2022GXNSFDA035079 and Grant GuiKeAD21075030, in part by the National Natural Science Foundation of China under Grant 61972167, in part by the Guangxi "Bagui Scholar" Teams for Innovation and Research Project, in part by the Guangxi Collaborative Innovation Center of Multi-Source Information Integration and Intelligent Processing, in part by the Guangxi Talent Highland Project of Big Data Intelligence and Application, and in part by the Research Fund of Guangxi Key Laboratory of Multi-Source Information Mining and Security under Grant 22-A-03-01. This article was recommended by Associate Editor L. Zheng. *(Corresponding authors: Bineng Zhong; Xianxian Li.)*

Yaozong Zheng is with the Guangxi Key Laboratory of Multi-Source Information Mining and Security, Guangxi Normal University, Guilin 541004, China, and also with the Department of Computer Science and Technology, Huaqiao University, Xiamen 361021, China.

Bineng Zhong, Qihua Liang, Zhenjun Tang, and Xianxian Li are with the Guangxi Key Laboratory of Multi-Source Information Mining and Security, Guangxi Normal University, Guilin 541004, China (e-mail: bnzhong@gxnu.edu.cn; lixx@gxnu.edu.cn).

Rongrong Ji is with the Media Analytics and Computing Laboratory, Department of Artificial Intelligence, School of Informatics, Xiamen University, Xiamen 361005, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2022.3212987.

Digital Object Identifier 10.1109/TCSVT.2022.3212987

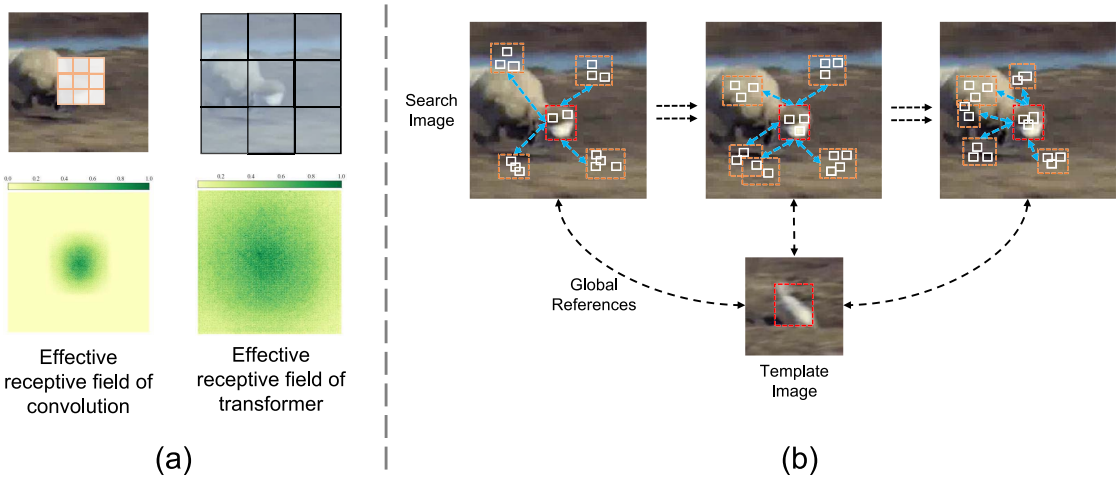


Fig. 1. Left (a): comparison of the effective receptive fields of convolution and transformer. Right (b): the contextual modeling process of target and background. The red box denotes the local receptive field area with target, the yellow box represents the local receptive field area with background, and the white box denotes the discriminative fine-grained features of each local receptive field. The example shows that our method could learn global and local relationships between the target and its backgrounds in an interactive way.

GAB are connected as a Trans-Conv unit block for global-local information flow interaction, that transformers receive local details from CNN model, and CNN absorbs global cues from transformer. From the global perspective, as shown in Fig.1(b), the proposed method can capture global and local relationships by referring to different background regions, which helps the tracker know where to target. It is significant for tracking. Further, we design a new tracking framework to perform accurate tracking which cascades several Trans-Conv unit blocks to better aggregate local and context information. In addition, a decoupled prediction head is applied to learn task-related features and promote the tracking results.

Aside from the good tracking results, our work indicates parallel interaction design of transformer and CNN contributes several new views to the tracking community. Our main contributions can be summarized as follows.

- We propose a novel and neat Siamese parallel interaction network for visual tracking. It is able to exploit rich scene information between the target and its background.
- We design a new tracking framework based on proposed network, which can interactively capture both global and local information to improve discrimination power.
- Our tracker runs in real-time and achieves state-of-the-art results on six benchmarks, especially our method gets a success score (AUC) of 70.2% on the GOT-10k test set.

II. Related Work

A. Siamese Network Based Trackers

In recent years, Siamese-based trackers have drawn great attention from the visual tracking community due to their simple architecture and superior performance. Bertinetto first introduced SiamFC [18], aiming to use a naive correlation layer to localize the target. To solve the scale variation of target, Li et al. [16] introduced region proposal network (RPN) [19] into Siamese network framework, which obtained a more accurate bounding box. SiamRPN++ [17] and SiamDW [20] solved the challenge of network depth, and adopted deep network as feature extractor, which made the performance greatly improved. Fan et al. [21] proposed an occlusion-aware Network to alleviate occlusion challenge. SiamBAN [1], SiamCAR [3], SiamFC++ [22] and Ocean [23] explored anchor-free mechanism into Siamese network framework, discarding the pre-defined anchor setting.

Recently, many trackers based on correlation variants have been proposed. PG-Net [6] adopted a pixel-to-global matching method to suppress the interference of background. SiamGAT [2] designed a graph attention network to embed the information between template nodes and search nodes. ACM [5] introduced an asymmetric convolution, which learns to capture the semantic correlation information of the two branches. AutoMatch [24] applied six matching operators to guarantee stable tracking in different environments. In addition, some further improvements have been made, such as attention mechanism [25], [26], few-shot learning [4], [27], model update mechanism [28], [29], feature alignment [30] and sophisticated network architectures [31].

However, these work can only obtain relationships between local convolution windows in the template and search regions, ignoring global context relationships that are significant to distinguish the target. Meanwhile, it is difficult for these models to obtain larger receptive field for global modeling, due to the locality of the convolution operation, which may hinder the discrimination power of the tracker. Different from these work, our method aims to fully explore the local and global context relationships between the target and background, improving the discrimination to distinguish the target in complex tracking scenarios.

B. Transformer-Based Trackers

The original transformer [12] achieves impressive success in the NLP task with the self-attention mechanism. Recently, most researchers have started to pay attention to self-attention mechanism and migrate it to various vision

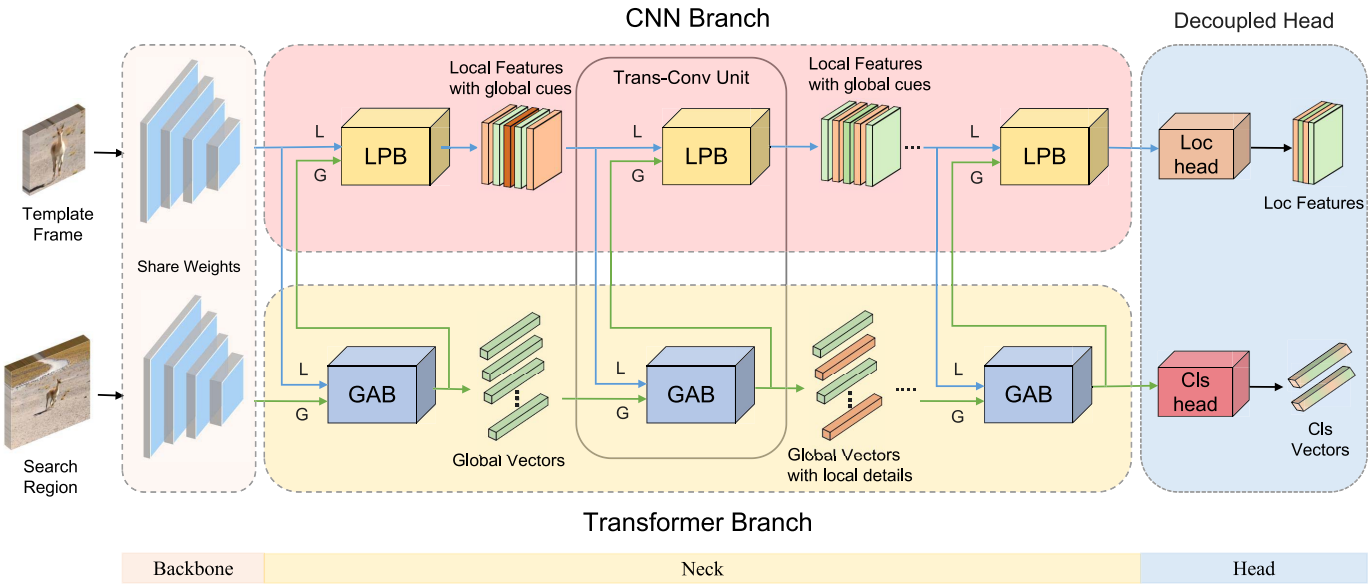


Fig. 2. Overall architecture of our SiamPIN consisting of three components, i.e., the backbone, neck, and head components. In the neck component, the CNN and transformer style structures are incorporated into our SiamPIN via two parallel branches. The grey solid box is a Trans-Conv Unit, including a global aggregation block (GAB) and a local process block (LPB), which fuses two types of features. The G and L denote the inputs of GAB and LPB, respectively. The blue line represents the information flow of CNN branch, and the green line denotes the information flow of transformer branch. In the head component, the decoupled head consists of a decoupled foreground-background classification branch and a decoupled target localization branch.

tasks, such as image recognition [32], object detection [33]. ViT [32] is the first pure transformer-based model in vision tasks, which splits the image into a sequence of patches and provides a attention-style feature extractor. To propose a general attention-based backbone network and alleviate expensive computational burden, several works combining convolution and transformer have been proposed. For example, Conformer [13] and Mobile-Former [14] adopts a concurrent backbone to fuse local features and global representations for enhancing representation learning.

Recently, many studies [7], [8], [9], [10] have introduced transformer into single object tracking filed, which design transformer architecture by the encoder-decoder paradigm and shows remarkable performance improvement. Specifically, TransT [7] introduces a transformer-based feature fusion network with ego-context and cross-feature augment modules for fusing features in both two branches, which prevents the correlation operation from falling into local optimal and losing semantic information. STARK [8] casts object tracking as a direct bounding box prediction problem and designs a transformer-based tracker, which models the global spatio-temporal feature dependencies between target objects and search regions by the self-attention and cross-attention modules. TrDiMP [9] based on the end-to-end DiMP [34] approach, utilizes a transformer architecture to reinforces multiple template features for relationships modeling and propagates tracking cues to the current frame to performs tracking task. Besides, TrTr [10] separates transformer encoder and decoder into two branches to encode the template feature and search feature for tracking.

Nevertheless, these transformer-based models obtain global relationships between feature maps of two branches to localize the strong relevant region as target position in the search region, overlooking capturing the local detailed information in other weakly relevant regions. By complementing these two types of features, it is possible to compensate for their respective shortcomings. Inspired by [14] and [13], in this work, we propose a novel Siamese parallel interaction network containing GAB and LPB modules to generate discriminative features for robust tracking. Benefiting from it, we discard the decoder in transformer architecture, and add two decoupled heads to two branches of the parallel interaction network.

III. Method

A. Overview

As shown in Fig.2, the proposed tracking framework consists of a Siamese feature extraction network, a parallel interaction network, and a decoupled head network. The Siamese feature extraction network utilizes a modified ResNet50 [35] as backbone network, in which the last stage and full connected layers of the standard ResNet50 are removed. The input of the backbone network is a pair of cropped images: a template patch with a size of $3\times128\times128$ and a search patch with a size of $3\times256\times256$. In order to reduce the computational burden, we apply a 1×1 convolution layer and a BatchNorm layer to reduce the output features channel dimension of backbone network.

Different from the previous methods, our parallel interaction network combines the advantages of transformer and CNN, fully exploring global and local information between target and background for visual tracking. First, we flatten template feature and search feature along the spatial dimension to obtain template vectors $f_z \in \mathbb{R}^{c\times H_z W_z}$ and search vectors $f_x \in \mathbb{R}^{c\times H_x W_x}$, as the input of parallel interaction network. Then, as shown in Fig. 2, the template vectors and search vectors

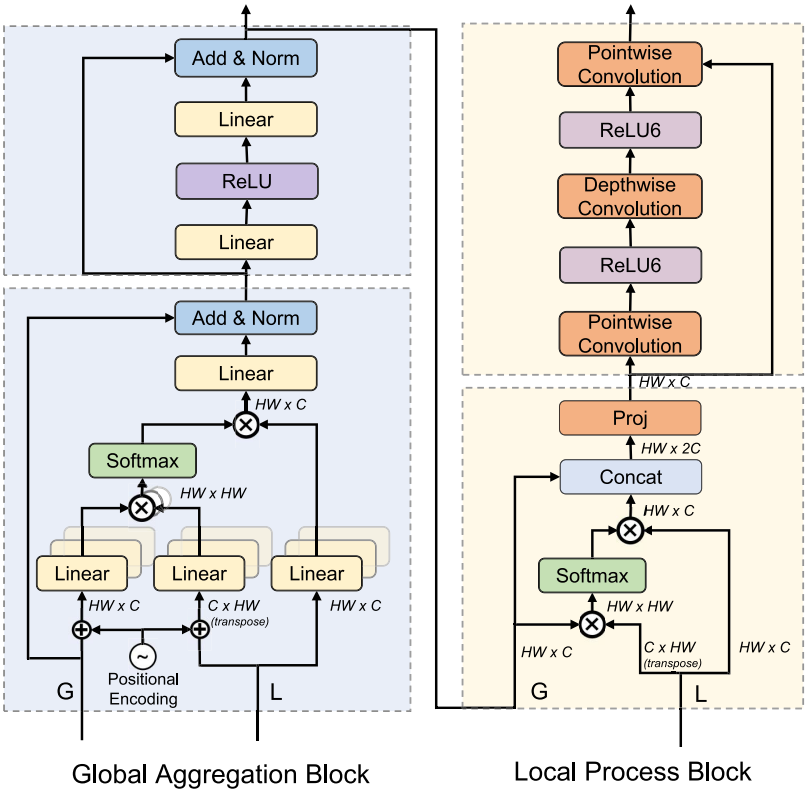


Fig. 3. Architecture of a Trans-Conv Unit. ⊗ denotes matrix multiplication. It consists of two sub-modules, named as Global Aggregation Block (GAB) and Local Process Block (LPB). GAB propagates the spatial detail information of local feature to global vectors. LPB embeds global cues into local features for global perception enhancement.

consecutively feed to GAB, which construct a global context relationships between them by using multi-head attention mechanism. Then, the learned global features are fed back to LPB, extracting local details of the target by leveraging the efficient depthwise convolution and pointwise convolution. The outputs of two parallel branches are fed to two decoupled prediction heads, in which CNN branch performs localization task and transformer branch performs classification.

B. Parallel Interaction Network

Most existing trackers [1], [2], [16], [17], [18], [25] used convolutional neural networks for modeling, which greatly promoted progress in the visual tracking field. Recently, many trackers [7], [8], [9], [10] based on transformer have been proposed, which have explored the modeling under the global receptive field. However, they are unable to take advantage of the local features of CNNs and the global information of transformer at the same time. CNN-based trackers are difficult to capture global context information that benefits for strengthening the discriminability to distinguish the target, while the transformer-based trackers fail to extract local detailed information of weakly relevant regions.

Therefore, we propose a novel parallel interaction network based on transformer and CNN to capture rich global information and fine-grained local details between the target and backgrounds. our model is composed of transformer and CNN branches described in detail next.

1) Transformer Branch: The transformer branch stacks N Global Aggregation Blocks (GAB)[1] in series. As shown in Fig.3, each GAB contains a multi-head attention module and a feed forward network, which can propagate the details of local features to the global vectors. In order to fully explore the long range dependence under the global receptive field, we input the local features $L \in \mathbb{R}^{HW \times C}$ and global vectors $G \in \mathbb{R}^{HW \times C}$ to GAB, and use the scaled dot product function to calculate the similarity matrix between them. The formulation is as follows:

$$\text{Attn}(G, L) = \text{SoftMax}\left(\frac{(G + P_G)(L + P_L)^T}{\sqrt{d}}\right) L, \quad (1)$$

where d is the dimensionality of local features, and P_G, P_L are the position encoding of the global vectors G and local features L.

In transformer architectures, since the 2-dimensional features are flattened into a 1-dimensional sequence of feature vectors before the attention is calculated, spatial position information is lost. To reinforce position information of feature vectors, position encoding are usually added to vectors G and L in each attention layer. However, in the network structure we designed, due to spatial convolution is introduced to embed

[1]Similar to most existing Siamese-based trackers [7], [8], [36], our tracker chooses a fixed crop region size for tracking. However, please note that the goal of our global aggregation blocks is to aggregate context information at a relative large cropped image level, rather than at the original image level. For a more detailed discussion, please see the limitations and potential solutions of a cropping search patch in the Section IV-E.

local spatial position information, we only add a sine position encoding layer in the first attention layer, but do not need it in the subsequent attention layers.

In order to explore more powerful long range dependence capabilities, we divide the G and L feature vectors into multiple sub-feature vectors, and by measuring the similarity with them, using multi-head attention mechanism to propagate relevant information from the local features L. The formula is as follows:

$$\text{MultiHead}(G, L) = \text{Concat}(h_1, \ldots, h_k)W_O, \tag{2}$$

$$h_{k_i} = \text{Attn}(GW_{k_i}^G, LW_{k_i}^L), \tag{3}$$

where $W_O \in \mathbb{R}^{c\times c}$, $W_{k_i}^G \in \mathbb{R}^{c\times d}$, and $W_{k_i}^L \in \mathbb{R}^{c\times d}$ are learnable parameters of projection layers. h_{k_i} indicates the k_i^{th} attention head. k denotes the number of attention heads.

2) CNN Branch: Since convolution operation possesses the inductive bias of locality and translation equivariance, we introduce the CNN model to embed target information to search region, which can effectively extract the local details of target and absorb global cues from transformer architecture. To be specific, the CNN branch is composed of N repeat Local Process Blocks (LPB), and each LPB performs two operations in sequence. First, we use global and local feature vectors feed to LPB, which capture fine-grained appearance information of the target in the local level, and construct local relationships between target and background. The fusion operation of LPB can be summarized as:

$$\text{F}(G, L) = \text{SoftMax}\left(\frac{GL^T}{\sqrt{d}}\right)L, \tag{4}$$

$$\text{F}(G, L)_{i,j} = \frac{exp(\frac{G_i \cdot L_j^T}{\sqrt{d}})}{\sum_{j=1}^{HW} exp(\frac{G_i \cdot L_j^T}{\sqrt{d}})}, \tag{5}$$

where $F(G, L)_{i,j}$ is the similarity matrix calculation formula of each sub-vector of G and L, which measures the degree of importance between G_i and L_j vectors. i, j refers to the index of feature vectors, $G_i \cdot L_j^T$ denotes the dot-product operation of feature vectors.

Then, we concatenate the global vectors and the outputs of the fusion operation along the channel dimension, and use a 1×1 convolution layer and a BatchNorm layer for dimensionality reduction. It is worth noting that apart from this downsampling layer, we have not added other network layers, such as linear layer, in the fusion process. Considering that the linear layer performs a global projection operation on the input vector, at the same time, in order to maintain the simplicity of the fusion operation, we decided not to add other operations.

Further, we adopt efficient depthwise and pointwise convolutions to reinforce local relationships between target and background in the embedded process of LPB. To do so, our method is able to focus on detailed information of the target and filter out the distractors from background. Specifically, in the embedding process, we use a 1×1 up-projection convolution, a 3×3 depthwise spatial convolution, a 1×1 down-projection convolution and a residual structure.

3) Iterative Interaction of Parallel Interaction Network: There are two destinations for the outputs of each GAB in the transformer branch. One output is fed into the next GAB, while the other one is used as the input of the LPB. Similarly, the output of the LPB also has two destinations, one for the next LPB and the other for the GAB in the same Trans-Conv Unit. Owing to the interactive process between GAB and LPB carried out through attention mechanism, the output features of them are aligned in the dimension of space and channel, and the bidirectional interaction can be friendly, without any additional operations.

To obtain more comprehensive global context and fine-grained local information between the target and background, we design a interaction schema, cascade multiple interaction layers by making GAB and LPB as a Trans-Conv unit block, which interactively learning two types of information to enhance the discriminability to distinguish the target. Therefore, the global context information and detailed appearance information are fully explored, helping to filter out the distractors from background and accurately localize target. Although this interaction method is simple in concept, it can effectively cooperate fuse context and local information between the target and background. As shown in Fig. 4, if using LPB or GAB alone to fuse features of both them, it is difficult to obtain rich target information. Considering the convolution kernel is easy to capture the local information of the target by sliding window operation, while transformer is good at capturing global dependence through attention mechanism. Therefore, compared with other two methods, we adopt the interactive interaction schema to more effectively aggregate two types of feature information.

C. Decoupled Head

In this section, we illustrate the design process of decoupled head and the loss functions used in the classification and regression of the training phase. Owing to the parallel structure of the SiamPIN, we discarded the decoder structure of original transformer, making our model more flexible and scalable. Benefiting from it, we add two decoupled heads to two branches of the parallel interaction network, respectively. Meanwhile, since classification and regression are good at focusing on different feature information. The classification task is more sensitive to the salient area features of a target, and the boundary information of the target object is more favorable to the regression task [37], [38].

Therefore, we propose the decoupled head to optimize classification and regression tasks separately, which can better learn features related to the corresponding tasks. The classification and regression branch are a three-layer perception and a three-layer convolution, respectively. The linear layer can globally project input features to learn the relationship between each pixel, which is helpful to distinguish target in complex tracking scenarios, while the convolution layer can retain the spatial information of the target, which is more robust to regress target.

Similar to TransT [7], we simply divide the pixels in the ground truth into positive samples, and define the pixels

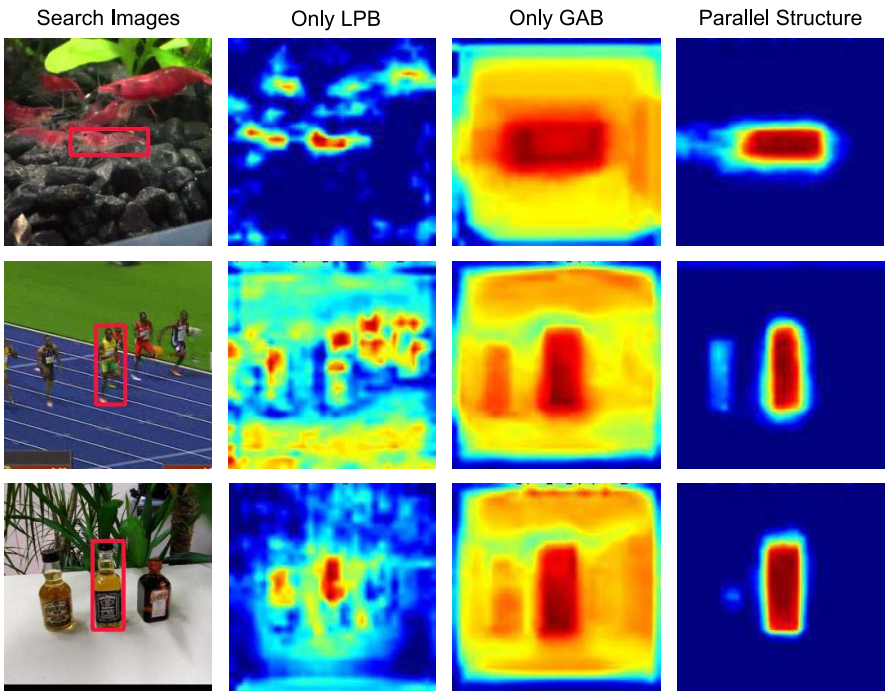


Fig. 4. Feature visualization for three different trackers, including a CNN-based tracker (LPB-based), a pure transformer-based tracker (GAB-based), and our SiamPIN. The 1^{st} column is the search images with ground truth. The 2^{nd} column denotes the feature maps using only LPB (the athlete's head in the second picture is activated). The 3^{rd} column means the feature maps using only GAB, in which most regions are activated, due to transformer's global modeling capability. The 4^{th} column shows the feature maps generated by our parallel interaction network, which are robust to distractors.

outside the ground truth as negative samples. For classification task, we adopt cross-entropy loss (referred as $\mathcal{L}_{cls}$) for calculate classification loss. For the localization task, we use the combination $\mathcal{L}_1$ loss and generalized IoU loss (denoted as $\mathcal{L}_G$) [39]. Our model is trained in an end-to-end fashion and the total loss function is denoted as

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_1 + \lambda_3 \mathcal{L}_G, \tag{6}$$

where λ_1=8, λ_2=5, and λ_3=2 are the regularization parameters to balance the contributions of each loss function in our experiments.

IV. Experiments

A. Implementation Details

We adopt the training splits of the GOT-10K [40], TrackingNet [45], COCO [49], and LaSOT [50] as the training dataset. The input of the network is an image pair, including a template patch with a size of 128×128 pixels and a search patch with a size of 256×256 pixels respectively. These training pairs are sampled from video sequences with the maximum interval of 100 frames. Our model is implemented in Python using Pytorch framework on a server with 4 NVIDIA TITAN X GPUs. At the same time, our tracker is running at a speed of 38 FPS on a GeForce GTX 1080Ti. We use ResNet50 [35] which is pre-trained on ImageNet [51] to initialize backbone network, and the parameters of the BatchNorm layers are frozen. The AdamW [52] is used to optimize the network with initial learning rate of 1×10^{-5} for the backbone, 1×10^{-4} for the rest, and weight decay 1×10^{-4}. Our model is trained for 500 epochs, with $60,000$ image pairs in each epoch and 12 sample pairs in each minibatch. The learning rate drops by a factor of 10 after 400 epochs. As for the number of the parallel interaction network layers, we employ 4 layers in this work.

B. State-of-the-Art Comparison

In this section, we evaluate the proposed model on six challenging tracking benchmarks, and compare with existing state-of-the-art trackers.

1) GOT-10K: [40] The GOT-10K is a large-scale tracking benchmark containing $10k$ video sequences. GOT-10k proposes a protocol, which the trackers only use the *training set* of GOT-10K for training. We follow the protocol to train our model to make a fair comparison, and evaluate our method by using the test set including 180 sequences. We compare it with recent high performance trackers, including STARK [8], TrDiMP [9], TransT [7], SiamRCNN [31], STMTrack [36], SAOT [41],AutoMatch [24], PrDiMP-50 [42], RPT [43], Ocean [23], DiMP50 [34], SiamRPN++ [17], SiamCAR [3], ATOM [44], and the results are shown in Tab.I. Among previous methods, STARK [8] achieves the best performance in average overlap (AO) and success rates at threshold 0.5 and 0.75 ($SR_{0.5}$ and $SR_{0.75}$), respectively. Benefiting from the new paradigm of parallel interaction, our model obtains a new state-of-the-art, gets a AO score of 0.702, $SR_{0.5}$ score of 0.809 and $SR_{0.75}$ of 0.637, respectively. Compared with STARK-ST50, our method outperform 2.2%, 3.2% and 1.4% in AO, $SR_{0.5}$ and $SR_{0.75}$, respectively. These results demonstrate that one benefit of our tracker comes from the parallel interaction network, which is designed to capture global context and local details between target and background.

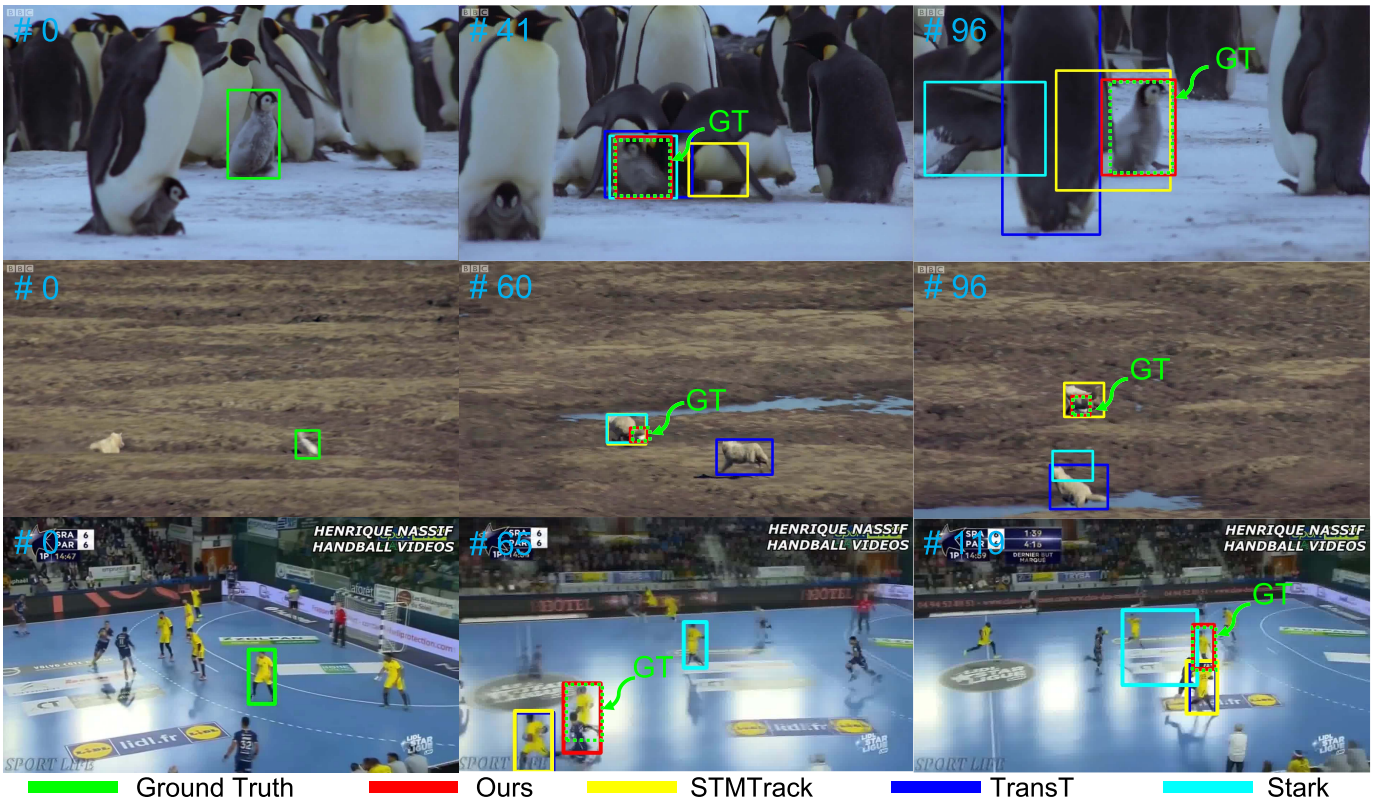


Fig. 5. Qualitative comparison results of our tracker with a CNN based tracker (e.g., STMTrack [36]) and two Transformer based trackers (e.g., TransT [7], Strack [8]) on three challenging sequences from GOT-10k Test set [40]. For example, in the second sequence, when the rabbit (target object) is occluded, the compared trackers fail to localize it. In contrast, benefiting from the proposed method, SiamPIN can accurately localize it. (It is worth noting that there is no open annotation information in the GOT-10k test set. To clearly illustration, we use the green dotted box to indicate the target annotation.)

TABLE I

COMPARISON ON THE GOT-10K TEST SET [40] WITH THE STATE-OF-THE-ART. THE BEST THREE RESULTS ARE HIGHLIGHTED IN RED, BLUE, GREEN, RESPECTIVELY

Tracker	AO↑	$SR_{0.5}$ ↑	$SR_{0.75}$ ↑
Ours	0.702	0.809	0.637
STARK [8]	0.680	0.777	0.623
TrDiMP [9]	0.671	0.777	0.583
TransT [7]	0.671	0.768	0.609
AutoMatch [24]	0.652	0.766	0.543
SiamRCNN [31]	0.649	0.728	0.597
STMTrack [36]	0.642	0.737	0.575
SAOT [41]	0.640	0.749	-
PrDiMP-50 [42]	0.634	0.738	0.543
RPT [43]	0.624	0.730	0.504
Ocean [23]	0.611	0.721	0.473
DiMP50 [34]	0.611	0.717	0.492
SiamCAR [3]	0.569	0.670	0.415
ATOM [44]	0.556	0.634	0.402
SiamRPN++ [17]	0.517	0.616	0.325

TABLE II

COMPARISON ON THE TRACKINGNET TEST SET [45] WITH THE STATE-OF-THE-ART. THE BEST THREE RESULTS ARE HIGHLIGHTED IN RED, BLUE, GREEN, RESPECTIVELY

Tracker	AUC↑	P_{Norm} ↑	P↑
Ours	81.4	85.7	79.0
TransT [7]	81.4	86.7	80.3
SiamRCNN [31]	81.2	85.4	80.0
STARK-S50 [8]	80.3	85.1	-
STMTrack [36]	80.3	85.1	76.7
TrDiMP [9]	78.4	83.3	73.1
PrDiMP-50 [42]	75.8	81.6	70.4
MAML [46]	75.7	82.2	72.5
SiamFC++ [22]	75.4	80.0	70.5
SiamAttn [25]	75.2	81.7	-
KYS [47]	74.0	80.0	68.8
DiMP50 [34]	74.0	80.1	68.7
SiamRPN++ [17]	73.3	80.0	69.4
CGACD [48]	71.1	80.0	69.3
TrTr [10]	71.0	80.3	-
ATOM [44]	70.3	77.1	64.8

Furthermore, as shown in Tab.III, we compare the experimental results between our tracker and two transformer-based trackers (e.g., TransT [7], Stark [8]) in terms of AO score, model parameters, and running speed, respectively. The number of parameters in our tracker is about 20.17 Million (M). Compared with TransT, our tracker achieves

TABLE III

COMPARISON OF PERFORMANCE AND MODEL PARAMETERS ON GOT-10K [40]

	AO	Parameters (M)	Speed (FPS)	Device
TransT [7]	0.671	22.96	33	GTX 1080Ti
Stark [8]	0.680	23.47	30	GTX 1080Ti
Ours	**0.702**	**20.17**	**38**	GTX 1080Ti

3.1% improvement on the AO score with reducing 12.15% model parameters. It can be seen that our tracker achieves the best tracking results on GOT-10K while maintaining a higher inference speed compared with TransT and Stark.

2) TrackingNet [45]: The trackingNet is a large-scale short-term tracking benchmark that provides a containing 511 video sequences in the test set. We compare our tracker with state-of-the-art trackers such as STARK [8], TrTr [10], SiamRCNN [31], STMTrack [36], TrDiMP [9], PrDiMP-50 [42], SiamFC++ [22], MAML [46], SiamAttn [25], KYS [47], DiMP50 [34], SiamRPN++ [17], CGACD [48], ATOM [44]. As shown in Tab.II, our method achieves the good tracking results that outperform most compared trackers. For example, our tracker gets a success score (AUC) of 81.4% and a normalized precision score (P_{Norm}) of 85.7%, outperforming previous state-of-the-art SiamRCNN [31] by 0.2% and 0.3%, respectively. Meanwhile, the performance of our tracker is compatible with the recent transformer-based tracker TransT [7], which is the same performance in terms of AUC score. Furthermore, our tracker also outperforms transformer-based trackers Stark-S50 [8], TrDiMP [9], TrTr [10] by 1.1%, 3% and 10.4% in terms of success score, respectively. These results meet our expectation that rich information of different ranges can be effectively captured in an interactive manner for robust tracking by combining the advantages of global and local features.

3) LaSOT [50]: The LaSOT is a large-scale benchmark with high-quality annotations, which consists of 1400 challenging sequences. Our model is evaluated on the test set with 280 sequences. We compared with recent high-performance trackers, including Stark [8], TransT [7], TrSiam [9], TrTr [10], SiamRCNN [31], STMTrack [36], AutoMatch [24], PrDiMP-50 [42], SiamAttn [25], DiMP50 [34], SiamFC++ [22], KYS [47], SiamBAN [1], SiamRN [27], SiamRPN++ [17], CGACD [48], ATOM [44], the results are shown in Tab.IV. We show two proposed trackers with different input resolution, including SiamPIN-256 and SiamPIN-320. It can be seen that SiamPIN achieves competitive results with an interesting parallel network architecture. Compared with the results from SiamRCNN [31], the proposed tracker with smaller input resolution (SiamPIN-256) gets a normalized precision score (PNorm) of 72.4% which obtains a gain of 0.2%. These results indicate that our learned global-local interaction tracker can provide reliable target location. Compared with TrSiam [9] and TrTr [10] that only uses original transformer for global feature fusion, our SiamPIN-256 outperforms TrSiam and TrTr by 1.7% and 9% in terms of success score, respectively. By increasing the input resolution, SiamPIN-320 further improves the AUC on LaSOT to 65.4% and outperforms the transformer-based tracker TransT [7] by 0.5% and 1.3% in terms of success and normalized precision score, respectively. These results show that the parallel interactive network is helpful to explore the advantages of information interaction with different granularity for visual tracking.

TABLE IV

COMPARISON ON THE LASOT TEST SET [50] WITH THE STATE-OF-THE-ART. THE BEST THREE RESULTS ARE HIGHLIGHTED IN RED, BLUE, GREEN, RESPECTIVELY

Tracker	AUC↑	P_{Norm} ↑	P↑
SiamPIN-320	65.4	75.1	69.3
SiamPIN-256	64.1	72.4	67.7
STARK-ST50 [8]	66.4	76.3	71.2
TransT [7]	64.9	73.8	69.0
SiamRCNN [31]	64.8	72.2	-
TrSiam [9]	62.4	71.2	64.5
STMTrack [36]	60.6	69.3	63.3
PrDiMP-50 [42]	59.8	68.8	60.8
AutoMatch [24]	58.3	-	59.9
DiMP50 [34]	56.9	65.0	56.7
SiamAttn [25]	56.0	64.8	-
KYS [47]	55.4	63.3	-
TrTr [10]	55.1	-	-
SiamFC++ [22]	54.4	62.3	54.7
SiamRN [27]	52.7	63.5	53.1
SiamBAN [1]	51.4	59.8	52.1
CGACD [48]	51.8	62.6	-
ATOM [44]	51.5	57.6	50.5
SiamRPN++ [17]	49.6	56.9	49.1

However, our tracking performance still has a minor gap compared to Stark [8], namely, 1% lower in the results of the AUC score. Based on our observations, one of the key reasons is that LaSOT is a large-scale long-term tracking benchmark. As a short-term tracker, our SiamPIN does not use any long-term strategies, such as online updating or global instance search, which are crucial for long-term tracking. As a result, our SiamPIN is not robust to handle the long-term sequences that contain the target disappears and re-appears challenges, which are still opening issues for tracking. In contrast, Stark design an updating branch to adapt to target appearance changes.

Besides, in order to validate the effectiveness of our tracker in the case of partial occlusion, we visualize success and normalized precision plots of partial occlusion attributes in Fig.6. It is worth noting that since Stark has an update branch, we did not add Stark in attribute evaluation. The SiamPIN-320 obtains 62.9% and 72.1% in terms of success and normalized precision, respectively. It clearly show that our method outperforms TransT and SiamRCNN. These results demonstrate our tracker can capture discriminative features to improve robustness in the case of partial occlusion. However, our tracker still has a lot of room for improving in long-term video sequences.

TABLE V

COMPARISON RESULTS ON THE OTB2015 [53] DATASETS. THE BEST THREE RESULTS ARE HIGHLIGHTED IN RED, BLUE, GREEN, RESPECTIVELY

	SiamRPN++ [17]	ATOM [44]	DiMP50 [34]	SiamBAN [1]	PG-Net [6]	PrDiMP [42]	SiamRN [27]	TrTr [10]	TrSiam [9]	TransT [7]	Ours
AUC	0.696	0.671	0.688	0.696	0.691	0.696	0.701	0.691	0.708	0.694	0.704

TABLE VI

ABLATION STUDIES ON THE VARIANTS OF OUR TRACKER IN GOT-10k BENCHMARK [40]

	Interaction Methods			Prediction Head			Got-10k		
	LPB	GAB	parallel	LPB. Head	GAB. Head	Decoupled Head	AO	$SR_{0.5}$	$SR_{0.75}$
①	✓			✓			0.601	0.723	0.464
②		✓			✓		0.655	0.745	0.582
③			✓	✓			0.69	0.786	0.626
④			✓		✓		0.681	0.774	0.620
⑤			✓			✓	0.702	0.809	0.637

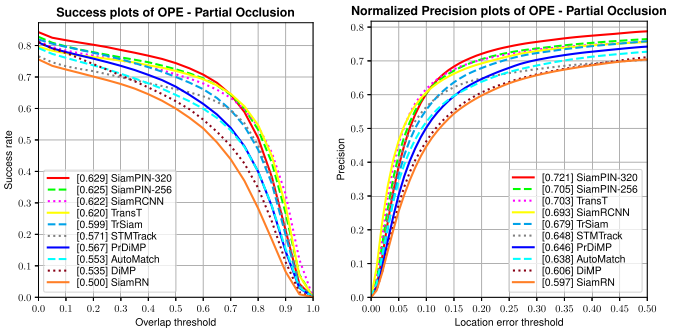


Fig. 6. Success and normalized precision plots of partial occlusion attribute on LaSOT [50].

Although our method does not outperform the latest trackers (which is impossible given the pace of tracking research), it presents a solid exposition of a Siamese tracker that exhibits a state-of-the-art performance. Besides, our work indicates parallel interaction design of transformer and CNN contributes several new views to the tracking community. We hope that our findings can contribute to more advanced design guidelines for the tracking community.

4) OTB2015 [53]: The OTB2015 is a popular tracking benchmark containing 100 well-annotated short-term sequences. The proposed tracker is compared with eight advanced object tracking methods (ATOM [44], SiamRPN++ [17], DiMP-50 [34], SiamRN [27], SiamBAN [1], PG-Net [6], PrDiMP [42], TransT [7], TrTr [10], TrSiam [9]). We report the comparison results in Tab.V. As shown in Tab.V, our method achieves a competitive result with a success score of 0.704. Compared to SiamRN [27] and TransT [7], the success score has been improved by 0.3% and 1%, respectively. Meanwhile, our tracker achieves similar performance to TrSiam [9]. These results are consistent with our expectation, that is, feature learning based on global and local range can effectively capture different context information, thereby improving tracking performance.

5) UAV123 [54]: We evaluate our tracker SiamPIN on UAV123 benchmark, which contains 123 low altitude aerial sequences. The proposed tracker is compared with six advanced object tracking methods (ATOM [44], SiamRPN++ [17], DiMP-50 [34], SiamRN [27], STMTrack [36], TrTr [10], TransT [7], TrSiam [9]). As shown in Tab.VIII, our method obtains 67.9% in terms of AUC score. It outperforms the DiMP [34] and STMTrack [36] by 2.6% and 3.2%, respectively. These results validate that enforcing the parallel interaction network with global-local range feature fusion is effective for robust tracking.

6) VOT2019 [55]: The VOT2019 contains 60 different challenging videos. we adopt the Expected Average Overlap (EAO) to evaluate the proposed tracker and other different trackers. As shown in Tab.VIII, our method achieves an EAO score of 0.300. It outperforms SiamRPN++ [17] on EAO indicator, which ranks third place. These results show that our tracking method is feasible and robust.

C. Ablation Studies

To explore the role of the various components in the proposed tracker, we conduct comprehensive ablation studies on the GOT-10K dataset.

1) Discussion on Local and Global Interaction Methods: Most previous work only use local fusion operation to fuse template and search features, or use global modules to fuse features of both alone. However, the global-local fusion methods are rarely investigated. Thus, to fully exploit the potential of both global and local information, we analyze the impact of three types of interaction methods. The comparison results are shown in Tab.VI. In the ①, ②, ③ and ④ of experiments, we adopt LPB-based, GAB-based and parallel interaction methods to fuse target and search features, respectively. Meanwhile, we also design three type of heads for visual tracking. Specifically, the LPB.Head denotes the classification and regression heads connect to the LPB module, and the GAB.Head represents the classification and regression heads connect to the GAB module. And the experiments are taken with cascaded 4-layer LPB or GAB blocks in all ablation studies in Tab.VI and Tab.VII.

In Tab.VI, the ① compared with ②, it can be found that the use of global aggregation method for learning contextual information of both target and background significantly improves

TABLE VII
ABLATION STUDIES ON THE VARIANTS OF OUR TRACKER IN GOT-10k BENCHMARK [40]

	Interaction Methods		Prediction Head		Got-10k		
	single-branch	dual-branch	LPB. Head	Decoupled Head	AO	$SR_{0.5}$	$SR_{0.75}$
①	✓		✓		0.679	0.777	0.598
②		✓		✓	0.702	0.809	0.637

TABLE VIII
COMPARISON RESULTS ON THE UAV123 [54] AND VOT2019 [55] DATASETS. THE BEST THREE RESULTS ARE HIGHLIGHTED IN RED, BLUE, GREEN, RESPECTIVELY

	ATOM [44]	SiamRPN++ [17]	DiMP-50 [34]	SiamRN [27]	STMTrack [36]	TrTr [10]	TransT [7]	TrSiam [9]	Ours
UAV123 (AUC)	0.642	0.613	0.653	0.643	0.647	0.594	0.691	0.674	0.679
VOT2019 (EAO)	0.292	0.285	-	0.341	-	0.313	-	-	0.300

performance. Compared with ① and ③, we could observe the proposed method is superior to the local fusion schema, which improves 8.9% in AO. ④ compared with ②, achieves improvement with AO, $SR_{0.5}$ and $SR_{0.75}$ gain of 2.6%, 2.9% and 3.8%, respectively. These results verify that, compared with other two methods, our method can learn rich context and local information by using the parallel interaction schema for robust tracking.

2) Discussion on Single-Branch and Dual-Branch Structure: In order to verify the effect of the single-branch and dual-branch structure, we investigate the different feature fusion structures. As shown in Tab.VII, the ① represents the tracker that alternate stack 4-layer GAB modules with LPB modules in a single branch, and the ② denotes the dual-branch tracker with stack 4-layer parallel interaction blocks. Compared with the ①, the ② stably increase the AO, $SR_{0.5}$ and $SR_{0.75}$ by more than 2.3%, 3.2% and 3.9%, respectively. We believe that the dual-branch tracker is able to aggregate global and local information to generate discriminative features, improve the performance of the tracker.

3) Discussion on Decoupled Head: In order to verify the effect of the decoupled head, we investigate the different prediction heads. As shown in Tab.VI, the ③ compared with ④, we found that the model with "LPB.Head" obtains a better performance in the three basic indications. It means that the CNN branch with global clues can better extract the fine-grained appearance information of target, contributing to obtain accurate results in some challenging tracking scenarios. Besides, compared with ③ and ④, the ⑤ combined with parallel interaction network and decoupled head has the best performance, and reaches 0.702, 0.809 and 0.637 in AO, $SR_{0.5}$ and $SR_{0.75}$, respectively. It means that the decoupled head decouples the input features of classification and localization head, which is helpful for the proposed model to learn task-aware features for robust tracking.

Furthermore, as shown in Tab.IX, we explore the impact of most variants of decoupled heads. We evaluate five variants of SiamPIN by modifying the prediction heads and tested their performance on GOT-10k. The first variant of SiamPIN is denoted as SiamPIN-V1, in which the classification branch is on the LPB and the regression branch is on the GAB. The second variant of SiamPIN is denoted as SiamPIN-V2, in which classification and regression branches are on both LPB and GAB. And in the inference phase, we only **use the classification scores and regression results of LPB**. The third variant of SiamPIN is denoted as SiamPIN-V3, in which classification and regression branches are on both LPB and GAB. And in the inference phase, we only **use the classification scores and the regression results of GAB**. The fourth variant of SiamPIN is denoted as SiamPIN-V4, in which classification and regression branches are on both LPB and GAB. And in the inference phase, we **use the classification scores of LPB and the regression results of GAB**. The fifth variant of SiamPIN is denoted as SiamPIN-V5, in which classification and regression branches are on both LPB and GAB. And in the inference phase, we **use the classification scores of GAB and the regression results of LPB**. As shown in Tab.IX, we can see that the best AO score (0.702) can be obtained by setting the classification branch on the GAB and regression branch on the LPB, respectively. These results also illustrated that our decoupled head is effective and reasonable. The decoupled head is helpful for our model to learn the features related to the corresponding task.

4) Discussion on Interaction Layer Number: Finally, we discuss the number of interaction layers for our SiamPIN. In the experiments, we observe that the AO, $SR_{0.5}$ and $SR_{0.75}$ score of our tracker also increased with the increase of the number of interaction layers. Our method achieves good tracking results when using four or six interaction layers. In order to balance speed and performance, we finally chose to use a 4-layer configuration.

5) Discussion on Loss Weights: The values of λ_1, λ_2, and λ_3 are empirically determined. To clearly show how to empirically determine these values, we add ablation studies on the parameters λ_1, λ_2, and λ_3. Our multi-task loss function contains three loss weights λ_1, λ_2, and λ_3, which correspond to the weights of the cross-entropy loss, L1 loss, and generalized IoU loss, respectively. Since there are relatively many parameter combinations, we only change one parameter at a time. Tab.XI shows the experimental results from GOT-10k [40]. It is easy to find that the best AO score (0.702) can be obtained

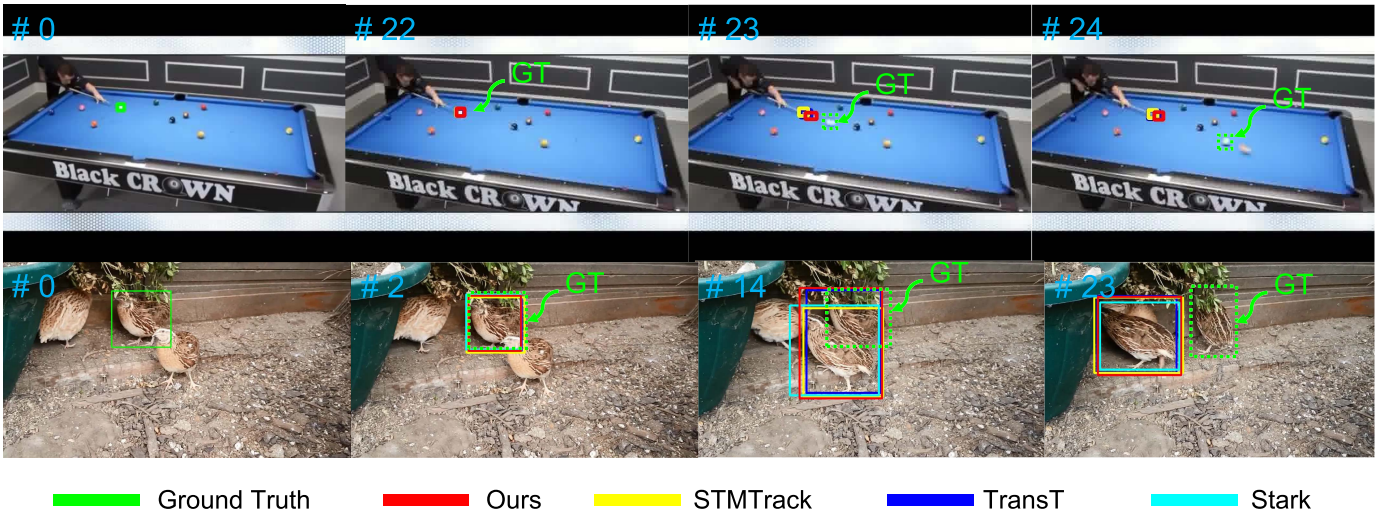


Fig. 7. Failure cases on two challenging sequences (e.g., *124* and *8* sequences) from GOT-10K Test set [40]. It is worth noting that there is no open annotation information in the GOT-10k test set. To clearly illustration, we use the green dotted box to indicate the target annotation.

TABLE IX
PERFORMANCE OF OUR SIAMPIN WITH DIFFERENT TYPES OF PREDICTION HEADS ON GOT-10k [40]

	SiamPIN-V1	SiamPIN-V2	SiamPIN-V3	SiamPIN-V4	SiamPIN-V5	SiamPIN
AO	0.689	0.677	0.673	0.675	0.672	0.702
$SR_{0.5}$	0.788	0.776	0.772	0.774	0.773	0.809
$SR_{0.75}$	0.628	0.620	0.615	0.615	0.618	0.637

by setting them to $\lambda_1 = 8$, $\lambda_2 = 5$, and $\lambda_3 = 2$, respectively. Moreover, these experimental results have verified that our total loss function is not sensitive to the small changes of the weight values and the effectiveness of our method.

D. Visualization Analysis

In order to analyze the effectiveness of our method, we compare the proposed parallel interaction network and other two network structures using only LPB or GAB. As shown in Fig. 4, the feature maps of LPB-based interaction method are goods at activates local regions (e.g. the athlete's head in the second picture is activated). The feature maps of GAB-based method activates almost the entire search regions, due to the global modeling ability of transformer. However, the feature maps generated by our SiamPIN are more robust to distractors, only the target area is activated and the background interference is suppressed. It implies that our method can effectively capture global context relationships and local detailed information of the target, which contribute to filter out distractors from background for object tracking.

In addition, to further clearly show the robustness of our method, we provide some qualitative comparison results in Fig.5. Benefiting from the proposed parallel interaction network, our tracker can effectively alleviate partial occlusion challenge, and achieve better results than the advanced trackers TransT [7] and STMTrack [36].

TABLE X
ABLATION STUDIES ON THE VARIANTS OF OUR TRACKER IN GOT-10k BENCHMARK [40]

	Interaction Layer Number	AO	$SR_{0.5}$	$SR_{0.75}$
①	1	0.531	0.637	0.370
②	2	0.632	0.736	0.544
③	4	0.702	0.809	0.637
④	6	0.690	0.781	0.631

TABLE XI
COMPARISON RESULTS UNDER DIFFERENT WEIGHT PARAMETER VALUES OF OUR MULTI-TASK LOSS FUNCTION ON GOT-10k [40]

λ_1	2	4	6	10	8	8	8	8	8	8	8	8	8
λ_2	5	5	5	5	1	3	7	9	5	5	5	5	5
λ_3	2	2	2	2	2	2	2	2	4	6	8	10	2
AO	0.693	0.683	0.695	0.690	0.684	0.691	0.694	0.890	0.693	0.687	0.689	0.681	0.702

E. Limitation and Future Work

Although our tracker achieves competitive performance compared with the top performance trackers, it performs poorly in the case of fast motion and similar distractors, which are still opening challenges for tracking community. As shown in Fig. 7, we visualize failure cases from two typical sequences (i.e., *124* and *8* sequences) from GOT-10K Test set [40]. Similar to most existing trackers [7], [8], [36], we are short of using larger search area strategy and special

design module [56] to deal with similar distractors. In future work, we will consider adopting global search strategy and candidate association module with motion prediction to model the relationships between the target and background.

V. Conclusion

In this work, we propose a novel Siamese parallel interaction network to capture discriminative features by aggregating local and global context information between the target and background. Furthermore, we design a new tracking framework based on the proposed network. By performing the decoupled prediction head, our tracker can not only learn task-related features, but also improve performance. Extensive experiments show that our tracker obtains the state-of-the-art performance with a real-time running speed. Our work indicates parallel interaction design of transformer and CNN contributes several new views to the tracking community. We hope that our findings can contribute to more advanced design guidelines for the tracking community.

References

[1] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, "Siamese box adaptive network for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6667–6676.

[2] 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.

[3] D. Guo, J. Wang, Y. Cui, Z. Wang, and S. Chen, "SiamCAR: Siamese fully convolutional classification and regression for visual tracking," in *Proc. CVPR*, Jun. 2020, pp. 6268–6276.

[4] J. Zhou et al., "Real-time visual object tracking via few-shot learning," 2021, *arXiv:2103.10130*.

[5] W. Han, X. Dong, F. S. Khan, L. Shao, and J. Shen, "Learning to fuse asymmetric feature maps in Siamese trackers," in *Proc. CVPR*, Jun. 2021, pp. 16570–16580.

[6] B. Liao, C. Wang, Y. Wang, Y. Wang, and J. Yin, "PG-Net: Pixel to global matching network for visual tracking," in *Proc. ECCV*, 2020, pp. 429–444.

[7] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in *Proc. CVPR*, Jun. 2021, pp. 8126–8135.

[8] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, "Learning spatio–temporal transformer for visual tracking," in *Proc. ICCV*, 2021, pp. 10428–10437.

[9] N. Wang, W. Zhou, J. Wang, and H. Li, "Transformer meets tracker: Exploiting temporal context for robust visual tracking," in *CVPR*, Jun. 2021, pp. 1571–1580.

[10] M. Zhao, K. Okada, and M. Inaba, "TrTr: Visual tracking with transformer," 2021, *arXiv:2105.03817*.

[11] S. Paul and P. Chen, "Vision transformers are robust learners," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2071–2081.

[12] A. Vaswani et al., "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[13] Z. Peng et al., "Conformer: Local features coupling global representations for visual recognition," in *Proc. ICCV*, 2021, pp. 357–366.

[14] Y. Chen et al., "Mobile-former: Bridging MobileNet and transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1–9.

[15] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted Windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[16] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with Siamese region proposal network," in *Proc. CVPR*, Jun. 2018, pp. 8971–8980.

[17] 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. CVPR*, Jun. 2019, pp. 4282–4291.

[18] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional Siamese networks for object tracking," in *Proc. ECCV Workshops*, 2016, pp. 850–865.

[19] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015, pp. 91–99.

[20] 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. 4591–4600.

[21] C. Fan, H. Yu, Y. Huang, C. Shan, L. Wang, and C. Li, "SiamON: Siamese occlusion-aware network for visual tracking," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Aug. 6, 2021, doi: 10.1109/TCSVT.2021.3102886.

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

[23] Z. Zhang, H. Peng, J. Fu, B. Li, and W. Hu, "Ocean: Object-aware anchor-free tracking," in *Proc. ECCV*, 2020, pp. 771–787.

[24] 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.

[25] Y. Yu, Y. Xiong, W. Huang, and M. R. Scott, "Deformable Siamese attention networks for visual object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6727–6736.

[26] M. Jiang, Y. Zhao, and J. Kong, "Mutual learning and feature fusion Siamese networks for visual object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3154–3167, Aug. 2021.

[27] S. Cheng et al., "Learning to filter: Siamese relation network for robust tracking," in *Proc. CVPR*, Jun. 2021, pp. 4421–4431.

[28] L. Zhang, A. Gonzalez-Garcia, J. van de Weijer, M. Danelljan, and F. S. Khan, "Learning the model update for Siamese trackers," in *Proc. ICCV*, 2019, pp. 4009–4018.

[29] X. Sun, G. Han, L. Guo, T. Xu, J. Li, and P. Liu, "Updatable siamese tracker with two-stage one-shot learning," 2021, *arXiv:2105.03817*.

[30] J. Fan, H. Song, K. Zhang, K. Yang, and Q. Liu, "Feature alignment and aggregation Siamese networks for fast visual tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1296–1307, Apr. 2021.

[31] P. Voigtlaender, J. Luiten, P. H. S. Torr, and B. Leibe, "Siam R-CNN: visual tracking by re-detection," in *Proc. CVPR*, Jun. 2020, pp. 6577–6587.

[32] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, pp. 1–9.

[33] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. ECCV*, 2020, pp. 213–229.

[34] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proc. ICCV*, 2019, pp. 6181–6190.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.

[36] 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. 13774–13783.

[37] Y. Wu et al., "Rethinking classification and localization for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10183–10192.

[38] G. Song, Y. Liu, and X. Wang, "Revisiting the sibling head in object detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11560–11569.

[39] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. D. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. CVPR*, Jun. 2019, pp. 658–666.

[40] 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.

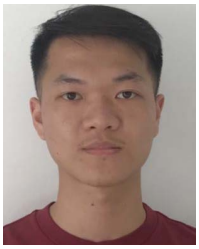
[41] Z. Zhou, W. Pei, X. Li, H. Wang, F. Zheng, and Z. He, "Saliency-associated object tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9846–9855.

[42] 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. 7181–7190.

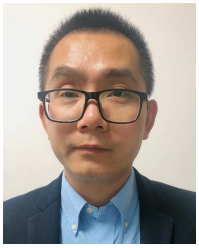
[43] Z. Ma, L. Wang, H. Zhang, W. Lu, and J. Yin, "RPT: Learning point set representation for Siamese visual tracking," in *Proc. ECCV Workshops*, 2020, pp. 653–665.

[44] 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.

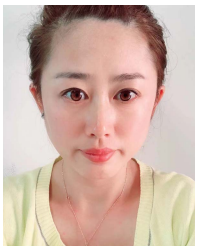
[45] M. Müller, A. Bibi, S. Giancola, S. Al-Subaihi, and B. Ghanem, "Trackingnet: A large-scale dataset and benchmark for object tracking in the wild," in *Proc. ECCV*, 2018, pp. 310–327.
[46] G. Wang, C. Luo, X. Sun, Z. Xiong, and W. Zeng, "Tracking by instance detection: A meta-learning approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6287–6296.
[47] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Know your surroundings: Exploiting scene information for object tracking," in *Proc. ECCV*, 2020, pp. 205–221.
[48] F. Du, P. Liu, W. Zhao, and X. Tang, "Correlation-guided attention for corner detection based visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6835–6844.
[49] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.
[50] 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.
[51] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[52] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019, pp. 1–9.
[53] 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.
[54] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. ECCV*, 2016, pp. 445–461.
[55] M. Kristan, A. Berg, and L. Zheng, "The seventh visual object tracking VOT2019 challenge results," in *Proc. ICCV Workshops*, 2019, pp. 2206–2241.
[56] C. Mayer, M. Danelljan, D. P. Paudel, and L. Van Gool, "Learning target candidate association to keep track of what not to track," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13424–13434.



Yaozong Zheng is currently pursuing the M.S. degree with the School of Computer Science and Technology, Huaqiao University, Xiamen, China. He is also a Visiting Student at Guangxi Normal University, Guilin, China. His research interests include computer vision and machine learning.



Bineng Zhong received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2004, 2006, and 2010, respectively. From 2007 to 2008, he was a Research Fellow with the Institute of Automation and Institute of Computing Technology, Chinese Academy of Science. From September 2017 to September 2018, he was a Visiting Scholar at Northeastern University, Boston, MA, USA. From November 2010 to October 2020, he was a Professor with the School of Computer Science and Technology, Huaqiao University, Xiamen, China. He is currently a Professor with the School of Computer Science and Engineering, Guangxi Normal University, Guilin, China. His current research interests include pattern recognition, machine learning, and computer vision.



Qihua Liang received the B.S. degree in accounting from Xiamen University, Xiamen, China, in 2014. She is currently a Teacher with the School of Computer Science and Engineering, Guangxi Normal University, Guilin, China. Her current research interests include computer vision and pattern recognition.

Zhenjun Tang (Member, IEEE) received the B.S. and M.Eng. degrees from Guangxi Normal University, Guilin, China, in 2003 and 2006, respectively, and the Ph.D. degree from Shanghai University, Shanghai, China, in 2010. He is currently a Professor with the Department of Computer Science, Guangxi Normal University. He has contributed more than 70 international journal articles. His research interests include image processing, video processing, and multimedia security. He is a reviewer of more than 30 SCI journals, such as IEEE journals, Elsevier journals, and Springer journals.

Rongrong Ji (Senior Member, IEEE) is currently a Professor and the Director at the Intelligent Multimedia Technology Laboratory and the Dean Assistant with the School of Information Science and Engineering, Xiamen University, Xiamen, China. His work mainly focuses on innovative technologies for multimedia signal processing, computer vision, and pattern recognition, with over 100 papers published in international journals and conferences. He serves as a program committee member for several Tier-1 international conferences. He also serves as an Associate Editor/the Guest Editor for international journals and magazines such as *Neurocomputing*, *Signal Processing*, *Multimedia Tools and Applications*, *IEEE MultiMedia* magazine, and the *Multimedia Systems*.

Xianxian Li received the Ph.D. degree in computer science and technology from Beihang University, Beijing, China. He is currently a Professor with the School of Computer Science and Engineering, Guangxi Normal University. His research interests include machine learning, data security, blockchain, and distributed systems.