One-Stream Stepwise Decreasing for Visual-Language Tracking

Guangtong Zhang, Bineng Zhong*, Qihua Liang, Zhiyi Mo, Ning Li, Shuxiang Song

***Abstract*—Based on the fixed language descriptions in the initial frames, a visual-language tracker typically adopts a two-stream model structure to align visual and language features at the feature fusion stages. However, this paradigm may degrade the tracking performance due to inaccurate language descriptions and lacking further modal interaction. To address these issues, we propose a one-stream visual-language model called One-stream Stepwise Decreasing for visual-language Tracking (OSDT). Specifically, we first encode the language description using a language encoder. The obtained language features are then combined with visual images and jointly entered into a visual encoder, in which the encoder's self-attention mechanism is utilized to facilitate more interactions between language and visual features. Moreover, to mitigate the problems caused by inaccurate language descriptions, we design a stepwise decreasing multi-modal interaction framework, in which a Feature Filter Module (FFM) is introduced to select language features that are more relevant to visual information to provide semantic guidance for visual feature extraction. Furthermore, without additional feature fusion modules, our one-stream model framework can efficiently use the proposed feature filtering module for feature selection. Consequently, our tracker can achieve a fast speed in the visual-language tracking domain compared to existing state-of-the-art methods. We extensively evaluate our tracker on three benchmarks, i.e., TNL2K, LaSOT, and OTB99, demonstrating competing performance compared to state-of-the-art visual-language tracking methods.**

***Index Terms*—Object Tracking, Visual-Language Tracking.**

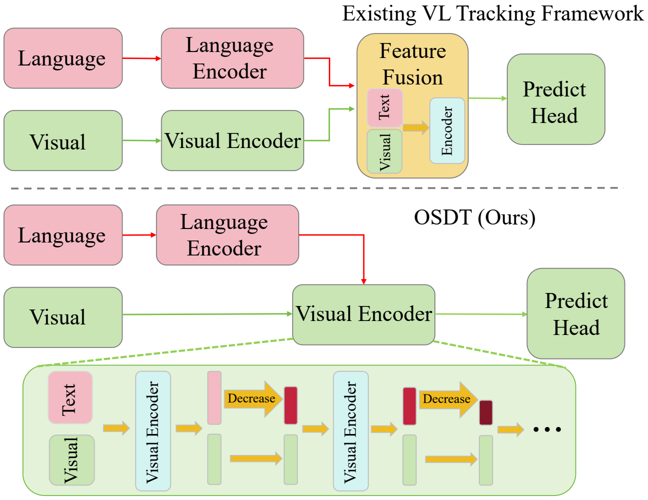


Fig. 1. Comparison between existing VL tracking frameworks and our OSDT model. Conventional VL tracking approaches autonomously extract visual and language features, achieving a basic feature alignment solely at the fusion stage, thus lacking substantial inter-modal interactions. In contrast, our OSDT integrates language features and visual information into the visual encoding layer, facilitating more efficient and multi-layered interactions between modalities.

I. Introduction

Visual tracking with natural language descriptions is an extension of conventional visual tracking methods. Visual language trackers simultaneously utilize visual and language descriptions to localize a target in each frame of a video sequence. Compared to traditional visual tracking methods, augmenting a model with natural language features not only provides additional target descriptive information to a tracker but also allows language descriptions to capture the long-term state features or trends of the targets. Leveraging language descriptions enhances the robustness of a tracker in handling more complex scenes.

In recent years, most visual-language trackers [1], [2], [16] have adopted a two-stream model architecture, where language descriptions and visual information are separately processed through a language encoder and visual encoder. Subsequently, a fusion module combines the language and visual features to obtain the target positions. Despite their success, there is still much to explore in effectively leveraging language descriptions and visual features. One of the main reasons is that this two-stream model structure suppresses the potential for more interactions between the language modality and the visual modality. Furthermore, the two-stream model structure only interacts visual features with language features during the modality fusion stage to jointly determine the target position. This may lead to tracking failures in complex scenes where non-target objects may satisfy the language description or where the language description deviates due to target motion deformation. As illustrated in the example Fig.2, the first sequence may have a language description *The car running in the middle of the road.* However, only the phrase of *The car running* accurately describes the target's current state during its movement. In the second sequence, the language description is *The basketball player in white who is in front of the player in purple.* It can be seen that several targets can satisfy this description in the later part of the sequence.

Inspired by some recent natural language processing work [3], [4] that have demonstrated that language descriptions can guide visual feature extraction, one question is put forward:

Guangtong Zhang, Bineng Zhong, Qihua Liang, Zhiyi Mo , Ning Li, Shuxiang Song are with Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, Guilin, 541004, China, the Guangxi Key Lab of Multi-Source Information Mining & Security, Guangxi Normal University, China.

Zhiyi Mo is currently a Professor in the School of Data Science and Software Engineering, Wuzhou University, Wuzhou 543002, China.

Bineng Zhong is the corresponding author.

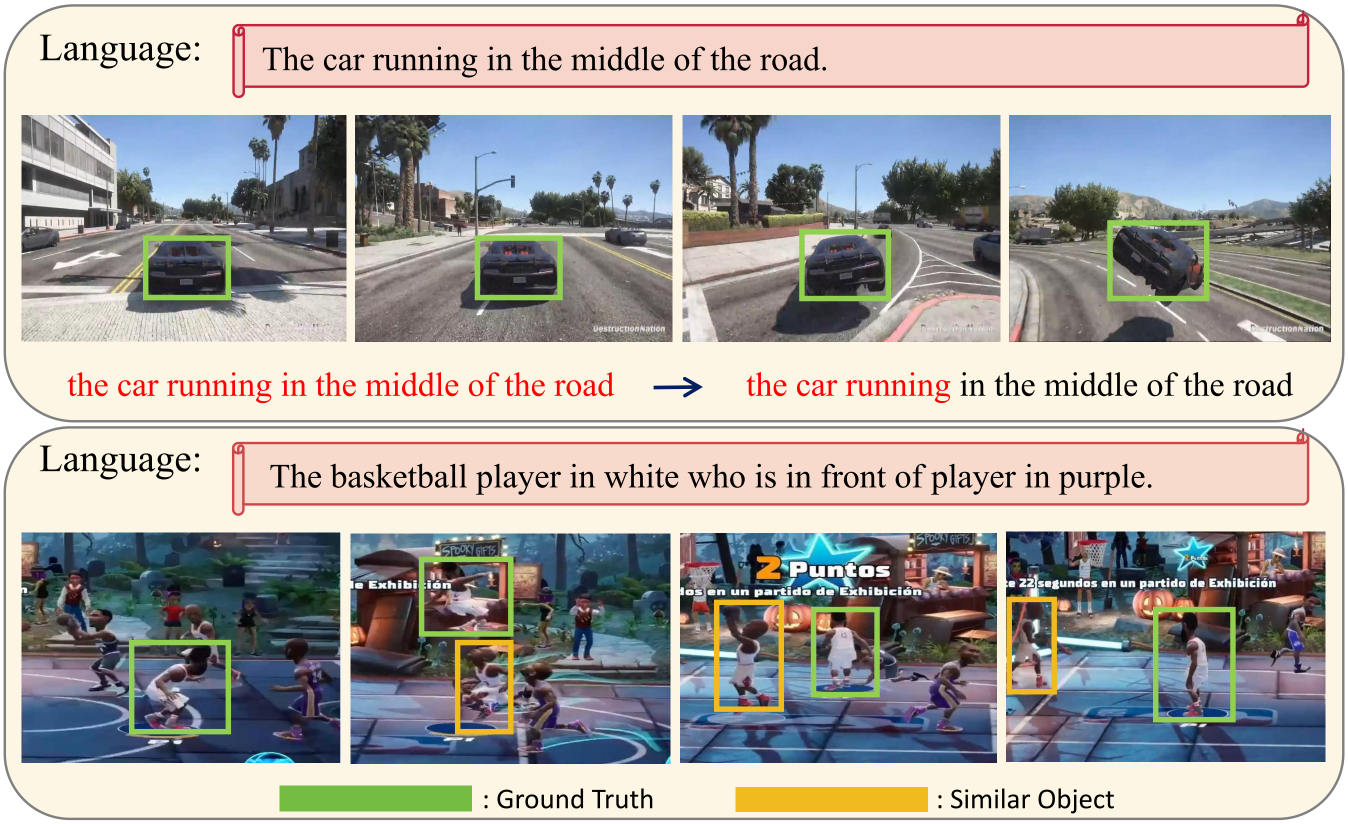

Language:
The car running in the middle of the road.
the car running in the middle of the road → the car running in the middle of the road
Language:
The basketball player in white who is in front of player in purple.
🟩 : Ground Truth
🟨 : Similar Object

Fig. 2. We present two sequences from the TNL2K test set. In the first sequence, the object (car) no longer satisfies the language description *in the middle of the road* after a period of motion; it only fulfills the description *the car running*. In the second sequence, the target within the green bounding box meets the language description, while the object within the yellow bounding box also satisfies the same language description.

Can we develop a simple yet effective one-stream visual language tracker that can more effectively leverage language information and provide more interactions between the visual and language modalities? Our answer is yes. The structure comparison between our proposed one-stream visual-language tracker and previous visual-language trackers is illustrated in Fig.1. In contrast to the previous two-stream visual-language trackers, which utilize separate visual and language feature extractors for each modality and interact between modalities only once in the fusion module, our approach takes a distinct approach. Firstly, we pass the language descriptions through a language feature extractor to obtain language features. Then, the language features and visual representations are jointly processed through visual encoder models and enable multiple interactions between modalities, followed by a head module to obtain the final target position. In the visual encoder modules, we propose a Feature Filter Module (FFM). The FFM applies the visual encoder modules, and subsequently aggregates more relevant language features to guide the next layer of the visual encoder. By employing multiple FFM modules, the language features are stepwise decreasing aggregated until the final encoding layer relies solely on visual features. This stepwise decreasing language feature guidance of visual feature extraction effectively resolves the issue of overcorrection caused by language descriptions. In summary, our main contributions are as follows:

- We propose a simple yet effective visual-language tracker that utilizes a one-stream structure, integrating language features into the visual feature extraction module to facilitate more interactions between the language modality and the visual modality.
- We introduce an effective Feature Filter Module (FFM) that enables the language information more efficient semantic guidance at different stages of visual feature interaction.
- We achieve state-of-the-art algorithm performance on three natural language tracking datasets and demonstrate significantly faster speed compared to existing visual language trackers.

II. Related Work

A. Single Object Visual Tracking.

Ever since the release of SiameseFC [7], the siamese network architecture has gained significant attention in the field of object tracking [8]–[10], [34]–[36]. This architecture involves feeding template frames and search frames into the same backbone network with equal weights. The corresponding features obtained are then fused in a fusion module to obtain the final target position. Based on this structure, numerous excellent trackers have emerged [11], [12]. Alongside the remarkable success of Siamese structure in the field of object tracking, a novel one-stream model structure has emerged. These approaches have initiated discussions on the efficiency of the

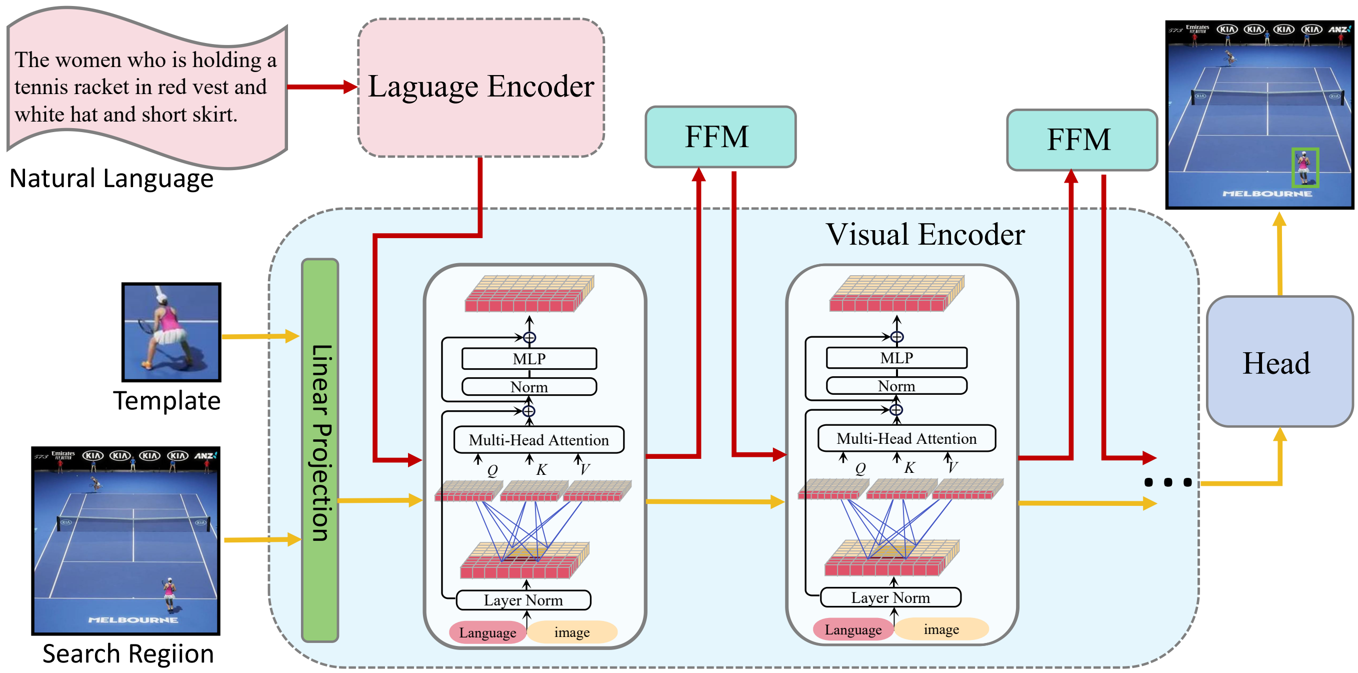


Fig. 3. Overview of the proposed one-stream stepwise decreasing framework. The language feature is first processed by the language encoder to obtain the language representation, which is then combined with the template frame and search frame to enter the visual encoder. The language feature is utilized in a decreasing manner through the feature filter module (FFM) to provide more guidance in the early stage of visual feature extraction. Finally, the output visual feature enters the head module to predict the object's position.

one-stream model structure in object tracking [5], [6]. The one-stream approach can more efficiently utilize template frame features to guide the extraction of search frame features. The template frame features and search frame template features efficiently interact throughout the entire feature extraction process. In contrast, the siamese structure fuses features only after the feature extraction process, making the extraction process relatively independent with fewer interactions.

B. Vision-Language Tracking.

In recent years, the success of Vision Transformers (ViT [13]) has demonstrated the effectiveness of Transformers in the field of computer vision. Multi-modal applications have also garnered increased attention and research, with a particular focus on the integration of language and visual modalities [3]. Early on, Li et al. [14] proposed the TNLS model, which employed LSTM to jointly track using language and visual information. Subsequently, RTTNLD [15] introduced a language-aware tracking approach that utilized language descriptions to provide global tracking proposals for each frame in a sequence, laying a solid foundation for subsequent visual-language tracking. Later, Wang et al. [2] released the new benchmark TNL2K, which focused on natural language and vision tracking and proposed two baseline methods involving natural language and natural language initialization. The creation of TNL2K provided more data and benchmarks for visual-language tracking. Feng et al. [16] proposed the SNLT module, embedded in a Siamese structure model, enabling visual tracking tasks to adapt to visual-language tracking. Subsequently, Li et al. [17] employed a target-specific retrieval module to initialize a local tracker, establishing a new baseline for visual-language tracking. VLT_{TT} [18] introduced an asymmetric model structure that used language to select visual information. JointNLT [1] relied on language descriptions to locate reference objects, creating a joint visual-language grounding and tracking framework. We observed that most existing visual-language trackers are based on two-stream frameworks, where natural language and vision are independently processed to extract features and then fused in a fusion module. Moreover, the utilization of natural language information mostly occurs only once during the fusion module, lacking further interactions. Addressing these challenges, we extend the one-stream model framework from visual tracking to visual language multi-modal tracking to provide more interactions between the visual and language modalities. Additionally, we observed that the early-stage interaction between language and vision is more critical than the late-stage interaction, as the latter may lead to adverse effects. To address this issue, we propose FFM, which can aggregate features more relevant to visual features when interacting between the visual and language modalities.

III. Method

In this section, we primarily introduce the One-Stream Stepwise Decreasing for visual-language Tracking, a one-stream framework for visual language tracking. The proposed tracker consists of four main components: the language encoder, the visual encoder, the feature filter module, and the head module. Additionally, we propose a feature restoration structure within the feature filter module, which will be further analyzed in the

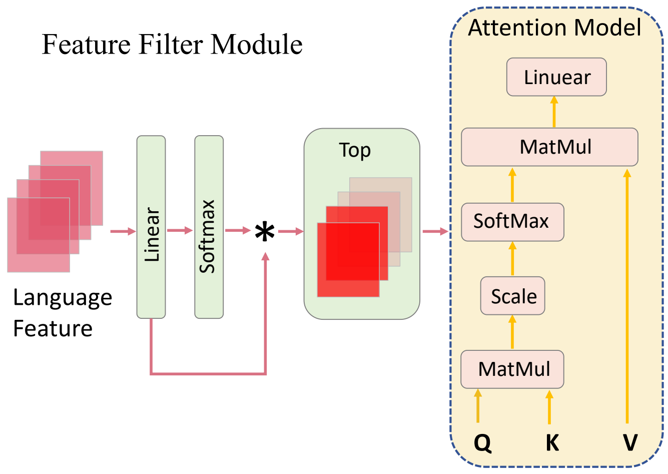


Fig. 4. The proposed feature filter module architecture is capable of selecting language features that are relevant to visual features.

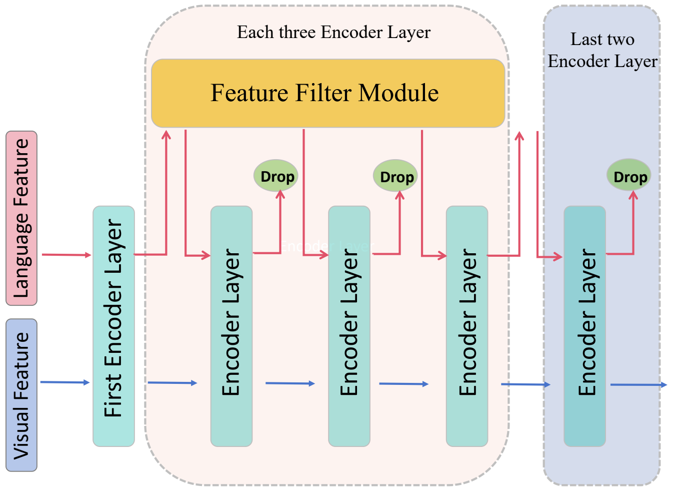


Fig. 5. The proposed feature restoration architecture is capable of addressing the issue of excessive interaction between language features and visual features that may affect their own feature representations.

feature module. The overall model framework is illustrated in Fig.3. Our proposed tracker first processes the natural language description through the language encoder. Subsequently, the obtained language feature is combined with the template frame and search frame to enter the visual encoder. Within the visual encoder, the language feature is adaptively used through a decreasing interaction scheme, entering different feature filter modules at various interaction layers to obtain language features more relevant to the current visual context, which then interacts with visual features. Finally, the output feature map of the search frame enters the head module for predicting the target's position.

A. Language Encoder.

Transformer has been applied and developed earlier in the field of natural language compared to computer vision, and in the natural language domain, there are already many mature language feature extraction models [19], [20] based on Transformer. Therefore, we choose the classic language feature extraction model BERT [19] as the language encoder for our tracker. Given a natural language description as the query Q_l, we first tokenize the language description and inject CLS and SEP tokens, resulting in a token sequence $L = \{\mathrm{CLS}, \boldsymbol{l}_1, \boldsymbol{l}_2, \cdots, \boldsymbol{l}_N, \mathrm{SEP}\}$, where N is the maximum length of the language query. Then, we input the token sequence into our language encoder to obtain language token features $\boldsymbol{T}_q \in \mathbb{R}^{C_q \times N}$, where $C_q = 768$ represents the output embedding dimension. It is worth mentioning that during the entire training process, all parameters of the language encoder are frozen. There are two main reasons for this choice: On the one hand, BERT is a widely used language feature extraction model in natural language processing, and fine-tuning it may lead to the risk of slight or significant forgetting of pre-trained knowledge. On the other hand, freezing BERT reduces GPU memory consumption and training requirements.

B. Vision Encoder.

We propose a visual-language one-stream framework, where the language feature extraction and visual feature extraction are no longer performed independently, but the language features are used to guide visual feature extraction. The constructed one-stream framework establishes a rich and flexible information flow between the template frame, the search frame, and the language description, providing more interactions for visual-language feature extraction. We choose the simple and efficient ViT [13] model as the visual encoder of our OSDT and pre-train it using MAE [23]. The input to the visual encoder includes a pair of images and language information, namely the template frame $z \in \mathbb{R}^{3 \times H_z \times W_z}$, the search frame $x \in \mathbb{R}^{3 \times H_x \times W_x}$, and the language feature extracted through the language encoder $\boldsymbol{T}_q \in \mathbb{R}^{N \times C_q}$. The template frame and search frame are first split and flattened into sequences of patches $z_p \in \mathbb{R}^{N_z \times (3 \cdot P^2)}$ and $x_p \in \mathbb{R}^{N_x \times (3 \cdot P^2)}$, where $P \times P$ is the resolution of each patch and $N_z = H_z W_z / P^2, N_x = H_x W_x / P^2$ is the number of patches in the template frame and search frame, respectively. Subsequently, using a trainable linear projection layer, the patches z_p and x_p are mapped to a D-dimensional latent space, referred to as patch embeddings. The learnable position embeddings are then incorporated into the patch embeddings of the template frame $\boldsymbol{B}_z$ and the search frame $\boldsymbol{B}_x$, generating the final template token embeddings $\boldsymbol{H}_z^0 \in \mathbb{R}^{N_z \times D}$ and search frame token embeddings $\boldsymbol{H}_x^0 \in \mathbb{R}^{N_x \times D}$. This process can be represented as follows:

$$
\begin{aligned}
H_z^0 = & \left[z_p^1 E; z_p^2 E; \cdots; z_p^{N_z} E\right] + B_z, \\
& E \in \mathbb{R}^{(3 \cdot P^2) \times D}, B_z \in \mathbb{R}^{N_z \times D}, \\
H_x^0 = & \left[x_p^1 E; x_p^2 E; \cdots; x_p^{Nx} E\right] + B_x, \\
& B_x \in \mathbb{R}^{N_x \times D},
\end{aligned} \tag{1}
$$

Next, the token sequences $\boldsymbol{H}_z^0$ and $\boldsymbol{H}_x^0$, and the language feature $\boldsymbol{T}_q$ are combined to form $\boldsymbol{H}_{tzx}^0 = \left[\boldsymbol{T}_q; \boldsymbol{H}_z^0; \boldsymbol{H}_x^0\right]$, and the resulting vector $\boldsymbol{H}_{tzx}^{0}$ is input to the transformer encoder layer.

C. Feature Filter Module.

In most existing visual-language trackers [1], [15], the interaction between the language and visual modalities often occurs only once at the feature fusion layer, lacking further integration. In our proposed one-stream language-visual tracking framework, we enable more interactions between the language and visual modalities, allowing the language modality to guide visual feature extraction. However, we raise a question: should all language information always participate in the visual feature interactions? On the one hand, through data observation as shown in Fig.2, we find that natural language descriptions mostly describe the state or trend of the first frame. During the target's movement, deviations from the language descriptions are likely to occur, leading to partial or complete mismatches and thus causing negative effects. On the other hand, constantly involving all language information in the visual feature interactions would increase the computational burden considerably and slow down the model.

To address these issues, we design a stepwise decreasing interaction structure and propose a Feature Filter Module (FFM) to implement this interaction structure. Our proposed FFM is depicted in Fig.4. Specifically, we take the multimodal joint feature $\boldsymbol{H}_{tzx}^{0}$ and output $\boldsymbol{H}_{tzx}^{1}$ after passing through the first-layer attention transformer. We separate the language feature $\boldsymbol{T}_1$ from $\boldsymbol{H}_{tzx}^{1}$ and feed it into the Feature Filter Module (FFM), which consists of a linear layer, a Softmax layer, a top module, and an attention module. The top module selects language features that are more relevant to the visual features, and then the attention module redistributes the feature attention. The entire process can be expressed as follows:

$$\begin{aligned} \boldsymbol{T}_1^t =& Top^n\left(Linear\left(\boldsymbol{T}_1\right) \cdot Softmax\left(Linear\left(\boldsymbol{T}_1\right)\right)\right), \\ \boldsymbol{T}_2 =& \,Attention\left(\boldsymbol{T}_1^t\right), \end{aligned} \tag{2}$$

In the formula Eq.2, we aggregate feature information from n highly responsive channels, where n represents the number of output feature channels. Subsequently, we reassign attention to the highly responsive channels.

Through the Feature Filter Module, we progressively reduce and aggregate the language information that is more relevant to the visual features, guiding the extraction of visual image features. And in the final attention transformer layer, feature extraction relies entirely on visual image depth information. We believe that language features are more effective in guiding visual image feature extraction at an early stage than at a later stage, which we further validate and analyze in the ablation experiments. In these experiments, we found that as language features continuously interact with visual features at deeper levels, they not only affect visual feature extraction but also their own representation due to the interaction with visual features. This issue may arise because most previous visual-language trackers only engage in visual-language interactions at the feature fusion layer without exploring deeper interactions. To address this problem, we embed a Feature Restoration Module in the stepwise decreasing structure, as shown in Fig.5. We apply the feature filter module in a stepwise manner, and when the feature filter module is not required for the filter, we restore the output language feature back to the feature of the previous feature filter module. The effectiveness of the feature filter module is further verified and analyzed in the ablation experiments section.

D. Head and Loss.

We treat the search frame feature tokens obtained through the visual encoder as a two-dimensional spatial feature map and feed it into a fully convolutional network (FCN). The FCN consists of K convolutional layers (Conv), Batch Normalization layers (BN), and ReLU activation layers (ReLU) stacked together. The output of the FCN is considered as the score map $\boldsymbol{M} \in [0,1]^{\frac{H_x}{M} \times \frac{W_x}{M}}$ for object classification and the local offsets $\boldsymbol{D} \in [0,1]^{2 \times \frac{H_x}{M} \times \frac{W_x}{M}}$ to address the discretization errors caused by reduced resolution and normalized bounding box size (width and height) $\boldsymbol{S} \in [0,1)^{2 \times \frac{H_x}{M} \times \frac{W_x}{M}}$. In the object classification score map, the object position is determined as the location with the highest classification score, i.e., $(x_d, y_d) = \arg\max_{(x,y)} \boldsymbol{M}_{xy}$. We obtain the final object bounding box as follows:

$$\begin{aligned} x &= x_d + \boldsymbol{D}\left(0, x_d, y_d\right), \\ y &= y_d + \boldsymbol{D}\left(1, x_d, y_d\right), \\ w &= \boldsymbol{S}\left(0, x_d, y_d\right), \\ h &= \boldsymbol{S}\left(1, x_d, y_d\right), \end{aligned} \tag{3}$$

During model training, we employ both the classification loss and regression loss in the Loss function. To handle focal loss [21] effectively, we use a simple yet efficient weighting strategy. For bounding box regression, we employ ℓ_1 Loss and generalized IoU loss based on the predicted bounding box.

IV. Experiments

A. Implementation Details.

We utilize the ViT [13] model pretrained with MAE [23] as our vision encoder and use the base uncased version of BERT as our language encoder. The visual input to the network is an image pair consisting of a template patch of size 128×128, and a search patch of size 256×256. For the language input, the max length of the language is set to 40, including a CLS and a SEP token. We use the training splits of TNL2k [2], LaSOT [24], OTB99 [26], and RefCOCOg-google [25] multiple training sets for joint training. Our model was implemented in the Pytorch framework on a server with 1 NVIDIA V100 GPU. Our model is trained with 300 epochs, each epoch with 60,000 image pairs and each mini-batch with 64 sample pairs. We also train the model using the AdamW optimizer, set the weight decay to 10^{-4}, the initial learning rate of the backbone to 8×10^{-5}, and other parameters to 8×10^{-4}. After 240 epochs, the learning rate is decreased by a factor of 10. We tested the proposed tracker on an NVIDIA 3090 GPU, and the tracking speed is about 67 FPS.

TABLE I

COMPARISON ON THE TNL2K, LASOT, OTB99 TEST SET WITH THE STATE-OF-THE-ART TRACKER. THE VISION-ONLY TYPE OF METHOD IS EVALUATED BY BOUNDING BOX INITIALIZATION, WHILE THE VISION-LANGUAGE (VL) TYPE OF METHOD IS EVALUATED BY A JOINT BOUNDING BOX AND NATURAL LANGUAGE INITIALIZATION. THE BEST RESULT IS HIGHLIGHTED IN RED.

Type	Method	Published	TNL2k			LaSOT			OTB99			Speed
			AUC	PRE	P_{norm}	AUC	PRE	P_{norm}	AUC	PRE	P_{norm}	
Vision-Only	SiamFC [7]	ECCV2016	29.5	45.0	28.6	33.6	42.0	33.9	-	-	-	86
	SiamBAN [9]	CVPR2020	41.0	48.5	41.7	51.4	59.8	52.1	-	-	-	40
	TransT [12]	CVPR2021	50.7	57.1	51.7	64.9	73.8	69.0	-	-	-	50
	Mixformer [31]	CVPR2022	-	-	-	69.2	78.7	74.7	-	-	-	-
	OSTrack [5]	ECCV2022	54.3	-	-	69.1	78.7	75.2	-	-	-	105
	SeqTrack [32]	CVPR2023	56.4	-	-	71.5	81.1	77.8	-	-	-	15
Vision-Language	VLT_{TT}* [18]	NeurIPS2022	53.1	53.3	-	67.3	72.1	-	76.4	93.1	-	35
	MMTrack [33]*	TCSVT2023	58.6	59.4	75.2	70.0	75.7	82.3	70.5	91.8	-	36
	TNLS [14]	CVPR2017	-	-	-	-	-	-	55.0	72.0	-	-
	DAT [22]	Arxiv2018	-	-	-	27.0	30.0	-	65.0	89.0	-	-
	RTTNLD [15]	WACV2020	25.0	27.0	-	35.0	35.0	43.0	61.0	79.0	73.0	30
	TNL2K-2 [2]	CVPR2021	42.0	42.0	42.0	51.0	55.0	-	68.0	88.0	-	25
	SNLT [16]	CVPR2021	27.6	41.9	-	54.0	57.6	-	66.6	80.4	-	50
	JointNLT [1]	CVPR2023	56.9	58.1	-	60.4	63.6	-	65.3	85.6	-	39
	OSDT	**Ours**	**59.3**	**61.5**	**76.2**	**64.3**	**68.6**	**73.4**	**66.2**	**86.7**	**78.0**	**67**

B. Datasets and Metrics.

We evaluate our method for visual-language tracking on the tracking benchmarks with natural language descriptions and visual images, including TNL2K [2], LaSOT [24], and OTB99 [26]. All these datasets adopt success, precision, and norm to measure tracking performance. It is worth mentioning that we did not include some trackers in the comparison that jointly train on single visual data and visual language data, such as VLT_{TT} [18]. VLT_{TT} utilizes COCO [27], Imagenet-VID [29], Imagenet-DET [29], Youtube-BB [30], GOT-10k [28], LaSOT [24], and TNL2K [2] datasets for training. To highlight the effective utilization of language information by the visual language model, we solely employed the visual-language datasets for training. MMTrack [33] is pre-trained using OSTrack384 [5], introducing a substantial amount of visual object tracking data as prior knowledge (the OSTrack384 model achieves an AUC of 71.1% on LaSOT). And MMTrack's search frame size is 384, whereas most tracking model utilizes a search frame size of 256. Despite the limited training data, our tracker still outperforms on the TNL2K dataset.

C. Comparison with the State-of-the-art Trackers.

To demonstrate the effectiveness of our proposed approach, we evaluated the model on three challenging tracking benchmarks and compared it with the existing state-of-the-art trackers.

TNL2K. The TNL2K dataset is a large-scale dataset containing over one million frames of 2K video sequences, where each sequence is accompanied by natural language descriptions and visual frames. For each sequence, natural language descriptions are provided to describe the target's behavior, typically including information about the target's position, actions, and appearance features. Moreover, each video sequence includes ground-truth bounding box annotations to indicate the target's location and size. The video sequences in the TNL2K dataset cover a diverse range of scenes and complexities, including indoor and outdoor environments, targets of different scales and velocities, as well as various motion and occlusion scenarios. This diversity makes the TNL2K dataset a challenging benchmark for evaluating the robustness and accuracy of visual-language tracking algorithms in complex scenarios. As shown in Tab.I, our results indicate that our tracker achieves the highest AUC (59.3%) and PRE (61.5%) scores on the TNL2K dataset. Compared to JointNLT, our tracker outperforms it on the TNL2K dataset, with a 2.4% improvement in the AUC score and a 3.4% improvement in the PRE score.

LaSOT. The LaSOT dataset contains over 1400 video sequences collected from the internet, encompassing more than 3 million frames of visual images. These sequences include a wide variety of targets with different scales, speeds, motion patterns, and occlusion scenarios, making the dataset challenging in complex scenes. Each video sequence is accompanied by detailed bounding box annotations, indicating the position of the target in each frame. Moreover, To support visual language tracking tasks, the LaSOT dataset provides natural language descriptions of the target behavior, typically including information about the target's position, actions, and appearance features. As shown in Tab.I, our tracker achieves the highest AUC (64.3%) and PRE (68.6%) scores on the LaSOT dataset. Compared to JointNLT, our tracker outperforms it on the LaSOT dataset, with a 3.9% improvement in the AUC score and a 5.0% improvement in the PRE score.

OTB99. The OTB99 dataset consists of 99 video sequences, covering diverse scenarios and complexities. These sequences encompass various everyday objects, such as humans, vehicles, and animals, and exhibit different motion patterns and occlusion scenarios. Each video sequence is accompanied by manually annotated bounding boxes, which indicate the target's position in each frame, and provide natural language descriptions of the target. As shown in Tab.I, our tracker achieves AUC (66.2%) and PRE (86.7%) scores on the OTB99 dataset. Our tracker exhibits slight performance differences compared to the TNL2k tracker on the OTB99 dataset. We attribute this discrepancy to the fact that OTB99 emphasizes the tracker's capability for long-term tracking. However, target descriptions tend to become less accurate in the context of extended tracking. While we have introduced a stepwise decreasing structure to mitigate this issue, our tracker is more susceptible to the influence of language information due to the multiple interactions we provide between language and vision.

D. Ablation Study.

We first validate and analyze the functionalities of each module to demonstrate the effectiveness of our proposed method. In this section, we adopt the same training methodology as the model and conduct evaluations on the TNL2k dataset to demonstrate the effectiveness of the proposed modules.

The Stepwise Decreasing Interaction Structure. To validate the effectiveness of our stepwise decreasing interaction structure, we designed the following experiments. In our OSDT tracker, we utilized the feature filter module to aggregate language features, setting the channel numbers of the aggregated outputs successively to 40→30→15→5→0. Additionally, we also trained the model using the same architecture with constant aggregated output channel numbers set to 40→40→40→40→40. Subsequently, we conducted evaluations on the TNL2K dataset and compared the results of OSDT with and without the stepwise decreasing interaction structure, as shown in Tab.II. Clearly, we observed that without the stepwise decreasing interaction structure, the AUC score (57.2%) decreased by 2.1% and the PRE score (59.1%) decreased by 2.4%, demonstrating the effectiveness of our proposed stepwise decreasing interaction structure. The main reason for its effectiveness lies in the accuracy of the current language description, due to language descriptions being based on the first frame, which might not fully match the moving target's actual appearance or behavior. Moreover, the current natural language descriptions for targets are often simple and imprecise, making them insufficient to serve as absolute features for tracking. Thus, our proposed stepwise decreasing interaction structure addresses these challenges effectively. During the early stage of visual feature extraction, more language features are involved in the interaction process, guiding the extraction of visual features. In the final stage of visual feature extraction, the interaction with language features is reduced, focusing more on the intrinsic deep-level information of the visual.

TABLE II

For the proposed ablation study on the decreasing interaction structure, we conducted experiments by keeping the aggregated channel number constant at 30 and training the model accordingly. Subsequently, we evaluated the performance of the TNL2K dataset. The best result is highlighted in red.

Filter Channel Num	AUC	PRE	P_{norm}
40 → 30 → 15 → 5 → 0 (OSDT)	59.3	61.5	76.2
40 → 40 → 40 → 40 → 30	57.2	59.1	67.3

TABLE III

For the ablation experiment of the proposed feature restoration module, we excluded the feature restoration structure and directly utilized the output features for training. Subsequently, we tested the model on the TNL2k dataset. The best result is highlighted in red.

method	AUC	PRE	P_{norm}
Restoration Feature (OSDT)	59.3	61.5	76.2
Inherit Feature	58.2	60.2	75.1

The Feature Restoration Structure. Although the feature restoration structure is part of our proposed stepwise decreasing structure, we still consider it a crucial component. In our OSDT, when the feature filter module maintains the output channel number without reduction, we do not directly use the language features from the previous layer's output. Instead, we restore the language features back to language-aggregated features. To validate the effectiveness of the feature restoration structure, we removed this structure and directly used the language features from the previous layer. We trained the model with the same training approach and tested it on the TNL2k dataset. The results comparison is shown in Tab.III, and it is evident that the model without the feature restoration module has decreased by 1.1% on the AUC score (58.2%) and decreased by 1.3% on the PRE score (60.2%) on the TNL2k dataset compared to the model with the feature restoration structure. This confirms the effectiveness of the feature restoration structure. We believe that the effectiveness of the feature restoration structure is primarily attributed to the efficient extraction of natural language features by our language encoder, along with the independence between the language modality and the visual modality. In the visual encoder, while the language features continuously interact deeply with the visual features, although the language features can guide the extraction of visual features, they are also influenced by the visual features. This excessive interaction may affect the representation of language features. By utilizing the feature restoration module, we can effectively avoid the impact on the representation of language features during continuous interaction, thereby better guiding the visual encoder in extracting visual features.

Study on the importance of language-annotated data. To validate the effective utilization of linguistic information by our tracker, we trained the tracker without linguistic information using the same training approach, and the comparative results are presented in Tab.IV. The tracker with linguistic information exhibited a decrease of 2.2% in AUC score, 3.4% in PRE score, and 2.6% in P_{norm} score compared to the tracker without linguistic information on the TNL2k dataset. The experimental comparisons clearly demonstrate the tracker's ability to enhance robustness by effectively leveraging linguistic information.

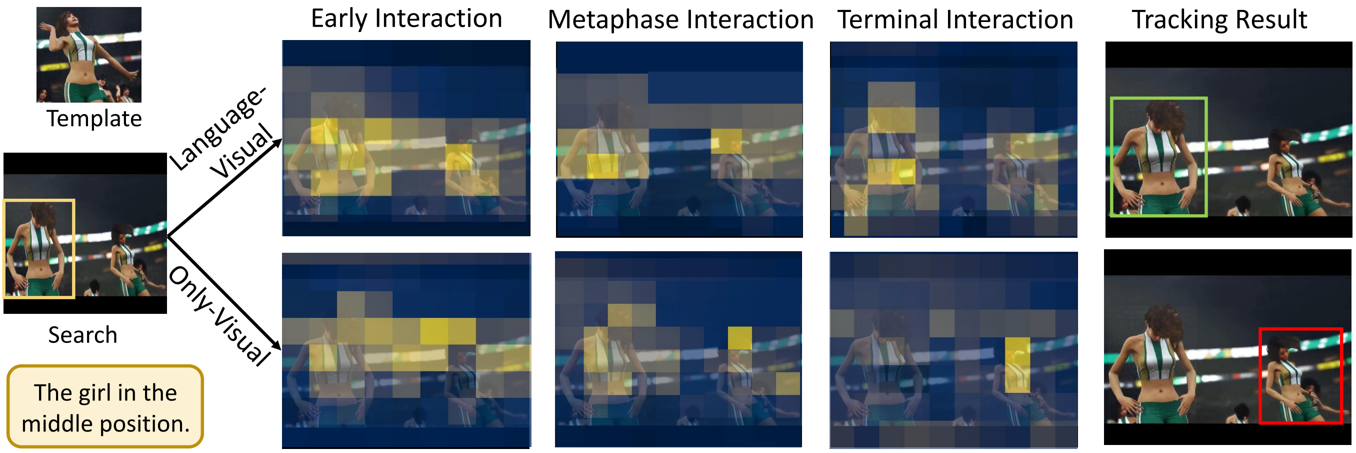


Fig. 6. Attention maps for models utilizing both language and visual information and models relying solely on visual information. Due to limitations in image cropping, the language actually conforms to the description.

TABLE IV
COMPARING THE PERFORMANCE OF USING LANGUAGE-VISUAL INFORMATION WITH USING VISUAL INFORMATION ONLY ON THE LASOT DATASET

method	AUC	PRE	P_{norm}
Visual and Language (OSDT)	64.3	68.6	73.4
Only Visual	62.1	65.2	70.8

E. Analysis and Limitations

Visualization. To visually demonstrate the effectiveness of our tracker, we visualized the attention maps of the tracker, as shown in Fig.6. By comparing the tracker with linguistic information and the one without linguistic information, the early interaction between visual and linguistic modalities assists the tracker in accurately locating the target during the feature extraction stage, thereby enhancing the robustness of the tracker.

Limitations. Through experimental comparisons on the LaSOT dataset, our tracker exhibits a certain performance gap when utilizing linguistic information compared to pure visual tracking. Further analysis was conducted to delve into this phenomenon. LaSOT is a long-term tracking dataset. Language descriptions are based on the initial frame's state, which leads to increasingly inaccurate descriptions as the target moves. And the quantity of data available for visual language tracking is significantly smaller when compared to traditional visual tracking data. The language descriptions in current visual language datasets are based on the initial state of the object, which results in decreasing accuracy of language descriptions as the object moves, leading to a reduction in available information. Therefore, we hope that future visual language datasets will provide language information that describes the ongoing state of the target object.

V. CONCLUSION

In this work, we propose a novel one-stream multi-modal framework for visual-language tracking, which we refer to as One-stream Stepwise Decreasing for visual-language Tracking (OSDT), incorporating a language feature filter module. The core idea is to enable more effective interactions between language and vision through the one-stream model framework. Furthermore, we observe that the early-stage interaction between language and visual features is more effective than the late-stage interaction. Late-stage interaction between modalities can have negative effects on target localization due to the inaccuracy of language information. Therefore, we introduce a stepwise decreasing visual-language feature interaction structure, utilizing the language feature filter module to aggregate more relevant language information and facilitate interactions with visual information at different stages. Our model effectively promotes the learning and interaction of natural language and visual information. Extensive experiments conducted on multiple visual language tracking benchmarks demonstrate the efficiency of our proposed method, achieving state-of-the-art (SOTA) performance in both accuracy and speed.

VI. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No.U23A20383 and U21A20474), the Project of Guangxi Science and Technology (No.2022GXNSFDA035079), the Guangxi ”Bagui Scholar” Teams for Innovation and Research Project, the Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing, the Guangxi Talent Highland Project of Big Data Intelligence and Application, and the Research Project of Guangxi Normal University (No.2022TD002).

REFERENCES

[1] Zhou, Li and Zhou, Zikun and Mao, Kaige and He, Zhenyu, *Joint Visual Grounding and Tracking with Natural Language Specification*. Conference on Computer Vision and Pattern Recognition, CVPR. 2023.

[2] Wang, Xiao and Shu, Xiujun and Zhang, Zhipeng and Jiang, Bo and Wang, Yaowei and Tian, Yonghong and Wu, Feng, *Towards More Flexible and Accurate Object Tracking with Natural Language: Algorithms and Benchmark*. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[3] Radford, Alec and Kim, JongWook and Hallacy, Chris and Ramesh, Aditya and Goh, Gabriel and Agarwal, Sandhini and Sastry, Girish and Askell, Amanda and Mishkin, Pamela and Clark, Jack and Krueger, Gretchen and Sutskever, Ilya, *Learning Transferable Visual Models From Natural Language Supervision*, Cornell University - arXiv,Cornell University - arXiv, 2021.

[4] Lu, Jiasen and Batra, Dhruv and Parikh, Devi and Lee, Stefan, *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*, Neural Information Processing Systems,Neural Information Processing Systems, 2019.

[5] Botao Ye and Hong Chang and Bingpeng Ma and Shiguang Shan, *Joint Feature Learning and Relation Modeling for Tracking: A One-Stream Framework* European Conference on Computer Vision,ECCV, 2022.

[6] Boyu Chen and Peixia Li and Lei Bai and Lei Qiao and Qiuhong Shen and Bo Li and Weihao Gan and Wei Wu and Wanli Ouyang. *Backbone is All Your Need: A Simplified Architecture for Visual Object Tracking*, European Conference on Computer Vision,ECCV, 2022.

[7] Luca Bertinetto and Jack Valmadre and João F. Henriques and Andrea Vedaldi and Philip H. S. Torr, *Fully-Convolutional Siamese Networks for Object Tracking*, European Conference on Computer Vision,ECCV, 2016.

[8] Bo Li and Wei Wu and Qiang Wang and Fangyi Zhang and Junliang Xing and Junjie Yan. *SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks*, Conference on Computer Vision and Pattern Recognition, CVPR, 2019.

[9] Zedu Chen and Bineng Zhong and Guorong Li and Shengping Zhang and Rongrong Ji and Zhenjun Tang and Xianxian Li. *SiamBAN: Target-Aware Tracking With Siamese Box Adaptive Network*,IEEE Transactions on Pattern Analysis and Machine Intelligence,TPAMI, 2022.

[10] Dongyan Guo and Jun Wang and Ying Cui and Zhenhua Wang and Shengyong Chen. *SiamCAR: Siamese Fully Convolutional Classification and Regression for Visual Tracking*, Conference on Computer Vision and Pattern Recognition, CVPR, 2020.

[11] Bin Yan and Houwen Peng and Jianlong Fu and Dong Wang and Huchuan Lu. *Learning Spatio-Temporal Transformer for Visual Tracking*, International Conference on Computer Vision, ICCV, 2021.

[12] Xin Chen and Bin Yan and Jiawen Zhu and Dong Wang and Xiaoyun Yang and Huchuan Lu. *Transformer Tracking*, Conference on Computer Vision and Pattern Recognition, CVPR, 2021.

[13] Alexey Dosovitskiy and Lucas Beyer and Alexander Kolesnikov and Dirk Weissenborn and Xiaohua Zhai and Thomas Unterthiner and Mostafa Dehghani and Matthias Minderer and Georg Heigold and Sylvain Gelly and Jakob Uszkoreit and Neil Houlsby. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, International Conference on Learning Representations,ICLR, 2021.

[14] Li, Zhenyang and Tao, Ran and Gavves, Efstratios and Snoek, Cees G. M. and Smeulders, Arnold W. M. *Tracking by Natural Language Specification*, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[15] Feng, Qi and Ablavsky, Vitaly and Bai, Qinxun and Li, Guorong and Sclaroff, Stan. *Real-time Visual Object Tracking with Natural Language Description*, 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020.

[16] Feng, Qi and Ablavsky, Vitaly and Bai, Qinxun and Sclaroff, Stan. *Siamese Natural Language Tracker: Tracking by Natural Language Descriptions with Siamese Trackers*, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[17] Li, Yihao and Yu, Jun and Cai, Zhongpeng and Pan, Yuwen. *Cross-modal Target Retrieval for Tracking by Natural Language*, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022.

[18] Mingzhe Guo and Zhipeng Zhang and Heng Fan and Liping Jing. *Divert More Attention to Vision-Language Tracking*, NeurIPS, 2022.

[19] Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, Proceedings of the 2019 Conference of the North, 2019.

[20] Liu, Yinhan and Ott, Myle and Goyal, Naman and Du, Jingfei and Joshi, Mandar and Chen, Danqi and Levy, Omer and Lewis, Michael and Zettlemoyer, Luke and Stoyanov, Veselin. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, Cornell University - arXiv,Cornell University - arXiv, 2019.

[21] Lin, Tsung-Yi and Goyal, Priya and Girshick, Ross and He, Kaiming and Dollar, Piotr. *Focal Loss for Dense Object Detection*, 2017 IEEE International Conference on Computer Vision (ICCV), 2017.

[22] Wang, Xiao and Li, Chenglong and Yang, Rui and Zhang, Tianzhu and Tang, Jin and Luo, Bin. *Describe and Attend to Track: Learning Natural Language guided Structural Representation and Visual Attention for Object Tracking*, Cornell University - arXiv,Cornell University - arXiv, 2018.

[23] Kaiming He and Xinlei Chen and Saining Xie and Yanghao Li and Piotr Dollár and Ross Girshick. *Masked Autoencoders Are Scalable Vision Learners*,Conference on Computer Vision and Pattern Recognition, CVPR, 2022.

[24] Heng Fan and Liting Lin and Fan Yang and Peng Chu and Ge Deng and Sijia Yu and Hexin Bai and Yong Xu and Chunyuan Liao and Haibin Ling. *LaSOT: A High-quality Benchmark for Large-scale Single Object Tracking*,Conference on Computer Vision and Pattern Recognition, CVPR, 2018.

[25] Mao, Junhua and Huang, Jonathan and Toshev, Alexander and Camburu, Oana and Yuille, Alan and Murphy, Kevin. *Generation and Comprehension of Unambiguous Object Descriptions*, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[26] Li, Zhenyang and Tao, Ran and Gavves, Efstratios and Snoek, Cees G. M. and Smeulders, Arnold W. M. *Tracking by Natural Language Specification*, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[27] Tsung-Yi Lin and Michael Maire and Serge Belongie and James Hays and Pietro Perona and Deva Ramanan and Piotr Dollár and C. Lawrence Zitnick. *Microsoft COCO: Common Objects in Context*, European Conference on Computer Vision,ECCV, 2014.

[28] Lianghua Huang and Xin Zhao and Kaiqi Huang. *GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[29] Deng, Jia and Dong, Wei and Socher, Richard and Li, Li-Jia and Kai Li and Li Fei-Fei. *ImageNet: A large-scale hierarchical image database*, 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[30] Real, Esteban and Shlens, Jonathon and Mazzocchi, Stefano and Pan, Xin and Vanhoucke, Vincent. *YouTube-BoundingBoxes: A Large High-Precision Human-Annotated Data Set for Object Detection in Video*, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[31] Yutao Cui and Cheng Jiang and Limin Wang and Gangshan Wu. *MixFormer: End-to-End Tracking with Iterative Mixed Attention*, Conference on Computer Vision and Pattern Recognition, CVPR, 2022.

[32] Xin Chen and Houwen Peng and Dong Wang and Huchuan Lu and Han Hu. *SeqTrack: Sequence to Sequence Learning for Visual Object Tracking*, Conference on Computer Vision and Pattern Recognition, CVPR, 2023.

[33] Yaozong Zheng and Bineng Zhong and Qihua Liang and Guorong Li and Rongrong Ji and Xianxian Li. *Towards Unified Token Learning for Vision-Language Tracking*, IEEE Transactions on Circuits and Systems for Video Technology(TCSVT), 2023.

[34] C. Fan, H. Yu, Y. Huang, C. Shan, L. Wang, and C. Li, *Siamon: Siamese occlusion-aware network for visual tracking*, IEEE Transactions on Circuits and Systems for Video Technology, pp. 1–1, 2021.

[35] M. Jiang, Y. Zhao, and J. Kong, *Mutual learning and feature fusion siamese networks for visual object tracking*, IEEE Trans. Circuits Syst. Video Technol, pp. 3154–3167, 2021.

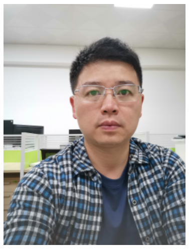
[36] J. Fan, H. Song, K. Zhang, K. Yang, and Q. Liu, *Feature alignment and aggregation siamese networks for fast visual tracking*, IEEE Trans. Circuits Syst. Video Technol, pp. 1296–1307, 2021.

Guangtong Zhang is currently working toward the master's degree at School of Computer Science and Engineering, Guangxi Normal University, Guangxi, 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 is a visiting scholar in Northeastern University, Boston, MA, USA. From November 2010 to October 2020, he is a professor with the School of Computer Science and Technology, Huaqiao University, Xiamen, China. Currently, he is 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 major from the Xiamen University, Xiamen, China, in 2014. Currently, she is 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.

Zhiyi Mo is a professor with the School of Data Science and Software Engineering, Wuzhou University, Wuzhou, China.Currently, He is working toward the Ph.D. degree at School of Computer Science and Engineering, Guangxi Normal University, Guilin, China. His research interests include computer vision, target tracking,and machine learning.



Ning Li received M.Eng. degree from Huazhong University of Science and Technology, Wuhan, China. He is currently studying for his Ph.D. degree in Guangxi Normal University, Guilin, China. His research interests include computer vision, and machine learning.

Shuxiang Song received the Ph.D. degree from the Department of Electronic and Information Engineering, Huazhong University of Science and Technology, Wuhan, China in 2009. He is currently a Full Professor with Guangxi Normal University. His current research interests include intelligent detection, automatic control, and signal and image processing.