

Unifying Visual and Vision-Language Tracking via Contrastive Learning

Yinchao Ma¹, Yuyang Tang¹, Wenfei Yang¹, Tianzhu Zhang^{1*}, Jinpeng Zhang², Mengxue Kang²

¹Deep Space Exploration Laboratory/School of Information Science and Technology, University of Science and Technology of China

²Intelligent Science Technology Academy of CASIC

{imyc,yuyangtang,yangwf}@mail.ustc.edu.cn, tzzhang@ustc.edu.cn

Abstract

Single object tracking aims to locate the target object in a video sequence according to the state specified by different modal references, including the initial bounding box (BBOX), natural language (NL), or both (NL+BBOX). Due to the gap between different modalities, most existing trackers are designed for single or partial of these reference settings and overspecialize on the specific modality. Differently, we present a unified tracker called UVLTrack, which can simultaneously handle all three reference settings (BBOX, NL, NL+BBOX) with the same parameters. The proposed UVLTrack enjoys several merits. First, we design a modality-unified feature extractor for joint visual and language feature learning and propose a multi-modal contrastive loss to align the visual and language features into a unified semantic space. Second, a modality-adaptive box head is proposed, which makes full use of the target reference to mine ever-changing scenario features dynamically from video contexts and distinguish the target in a contrastive way, enabling robust performance in different reference settings. Extensive experimental results demonstrate that UVLTrack achieves promising performance on seven visual tracking datasets, three vision-language tracking datasets, and three visual grounding datasets. Codes and models will be open-sourced at <https://github.com/OpenSpaceAI/UVLTrack>.

Introduction

Single object tracking is one of the fundamental research topics in computer vision, aiming to locate the target object in a video sequence according to the reference specified by the initial bounding box (BBOX) (Alper et al. 2006), natural language (NL) (Li et al. 2017), or both (NL+BBOX) (Wang et al. 2021b). It has a wide range of applications in robotics, video surveillance, autonomous driving, human-computer interaction and so on (Chen et al. 2022b). Although great progress has been achieved for specific reference settings, it is still challenging to design a unified tracker that performs well across all three reference settings.

Most trackers (Li et al. 2019; Xu et al. 2020; Yan et al. 2021) utilize the target bounding box in the first frame as the reference (**BBOX**). They commonly crop a template according to the given bounding box and locate the target in

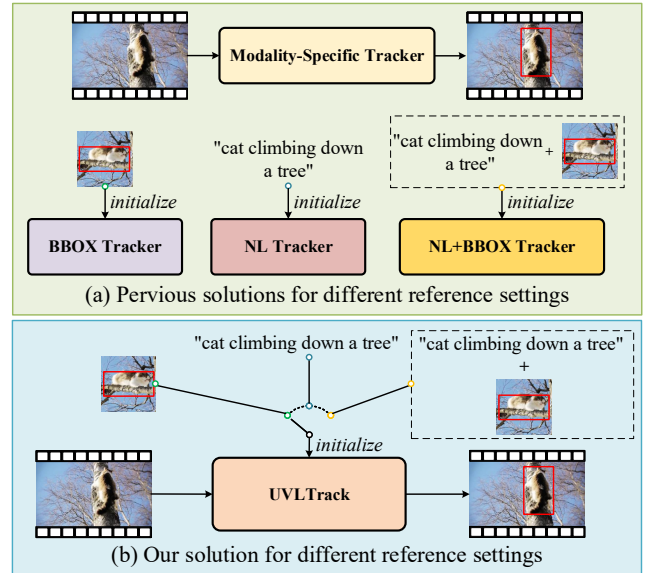


Figure 1: Comparison between previous solutions and UVLTrack. BBOX, NL, NL+BBOX tracker means the tracker is designed to utilize the bounding box, natural language, or both as the target reference respectively. Our UVLTrack can simultaneously handle three different reference settings.

subsequent frames by interacting with the cropped template. Representatively, one-stream trackers (Ye et al. 2022a; Cui et al. 2022) combine feature extraction and interaction of the template and search region in Transformer architectures, achieving superior performance. However, the bounding box has no direct target semantics, which may lead to ambiguity (Wang et al. 2021b). Different from the above tracking paradigm, tracking by natural language specification (Li et al. 2017) provides a novel manner of human-computer interaction, which specifies the target based on the natural language reference (**NL**). This task can be roughly divided into two steps. 1) locating the target in the first frame based on the language description. 2) tracking the target based on the language description and the predicted bounding box. Recently, JointNLT (Zhou et al. 2023) proposes a unified network to jointly conduct locating and tracking, enabling

*Corresponding author

end-to-end model optimization. For providing more accurate target reference, some trackers (Guo et al. 2022; Feng et al. 2021) specify the target by both natural language and bounding box (**NL+BBOX**). They embed language descriptions into visual features through dynamic filter (Li et al. 2017), cross-correlation (Feng et al. 2021) or channel-wise attention (Guo et al. 2022), achieving more robust tracking. Language description brings rich target semantics for tracking. However, due to the semantic gap between different modalities, trackers designed with natural language show limited performance in the bounding box reference setting.

The modality of the target reference varies with the application scenario. However, previous trackers overspecialize on the specific modalities, limiting their generalization, as shown in Figure 1. To address the above limitation, we seek to combine visual and vision-language tracking into a unified framework. It has two main benefits. First, the unified tracker can simultaneously cope with three types of target references, enabling a wider range of application scenarios. Second, we can utilize richer target references to optimize models, thereby improving their generalization ability. By studying previous methods, we summarize two key issues that need to be considered to design the unified visual and vision-language tracking framework.

1) **Modality-aligned feature learning.** Most vision-language trackers introduce natural language reference through well-designed fusion modules. Typically, VLT (Guo et al. 2022) designs a *ModaMixer* to fuse the language feature and visual feature by channel-wise attention. *JointNLT* (Zhou et al. 2023) builds feature interactions between language and vision through self-attention mechanisms. However, they ignore the semantic gap between different modalities, resulting in a tendency for trained visual-language trackers to rely on semantic information in language references, which limits their performance in pure bounding box reference setting (Guo et al. 2022). To this end, it is necessary to design a new modality-aligned feature extractor, achieving consistent feature learning for different modal references. 2) **Modality-adaptive target localization.** Existing trackers design various static box heads to estimate the target state, such as anchor-free head (Ye et al. 2022a), corner-based head (Zhou et al. 2023), and point-based head (Ma et al. 2023). These heads commonly take the reference-enhanced features of the search region as input, and regress the target box through offline trained parameters. However, various reference modalities increase the difficulty of static head training, which may lead to compromised results in different reference settings. Thus, we argue that it is better to design a dynamic head, which can make full use of different modal references to mine ever-changing scenario features from video contexts to discriminate the target.

Motivated by the above discussions, we propose a unified framework for visual and vision-language tracking, termed *UVLTrack*, which mainly consists of a modality-unified feature extractor and a modality-adaptive box head. The **modality-unified feature extractor** is constructed based on Transformer architecture, in which we extract features of different modalities separately in shallow encoder layers and fuse them in deep encoder layers. Such a design avoids the

confusion of low-level feature modeling between different modalities and allows high-level semantics interaction. Besides, we design a multi-modal contrastive loss to align visual and language features into a unified semantic space, so as to realize consistent feature learning for different modal references. The **modality-adaptive box head** dynamically mines ever-changing scenario information from video contexts and localize the target in a contrastive way. Specifically, we propose a novel distribution-based cross-attention mechanism, which can make full use of different modal references to adaptively mine features of target, distractor and background from historic scenarios. Then, the target can be localized directly through feature comparison. By introducing dynamic scenario information, *UVLTrack* can achieve more robust tracking under different modal references.

To summarize, the main contributions of this work are: (1) We propose a novel unified tracker, *UVLTrack*, for visual and vision-language tracking, which can simultaneously cope with three types of target reference (**BBOX**, **NL**, **NL+BBOX**). (2) We design a modality-unified feature extractor for joint visual and language feature learning, and design a multi-modal contrastive loss to align different modal features into a unified semantic space. (3) We propose a modality-adaptive box head to dynamically mine scenario features by different modal references and localize the target in a contrastive way, which helps *UVLTrack* achieve robust performance across all reference settings. (4) Extensive experimental results on seven visual tracking datasets, three vision-language tracking datasets, and three visual grounding datasets demonstrate that *UVLTrack* shows promising performance compared with modality-specific counterparts.

Related Work

Visual Tracking

Visual tracking aims to locate the target in a video sequence according to the given bounding box in the first frame (**BBOX**). Visual trackers commonly crop a target template from the first frame and estimate the target state in subsequent frames by interacting with the cropped template. Siamese-based trackers (Li et al. 2019; Guo et al. 2020) extract features from both template and search region with the siamese network and locate the target through well-designed matching modules. Discriminative Correlation Filter based trackers (Danelljan et al. 2019; Bhat et al. 2019) learn a correlation filter from historic target states in the video to discriminate the target from backgrounds. Recently, one-stream trackers (Cui et al. 2022; Ye et al. 2022a) achieve joint feature extraction and interaction using Transformer architectures (Wu et al. 2021; Vaswani et al. 2017), which simplify the tracking pipeline and achieve superior performance. However, these visual trackers are designed purely based on visual features, which cannot flexibly introduce high-level language semantics to reduce visual ambiguity.

Vision-Language Tracking

Natural language can provide clear target semantics to avoid visual ambiguity. Thus, some trackers seek to utilize natural

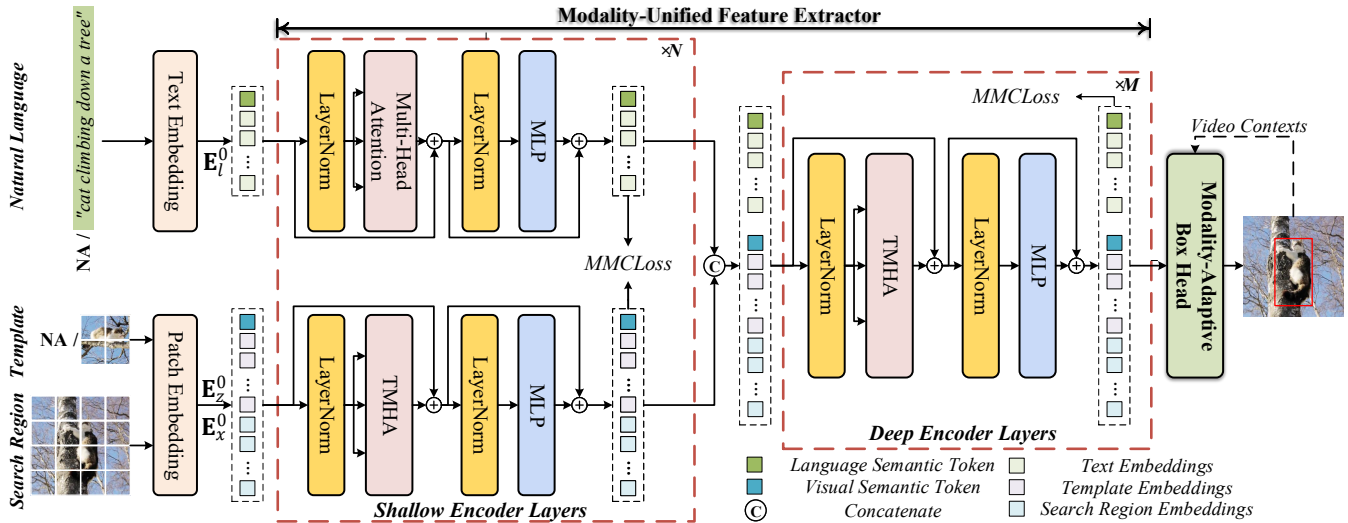


Figure 2: A unified tracking framework for different target references. NA means “not available”. Natural language is not available for visual tracking task and template is not available for grounding task. Different from previous trackers designed for specific reference modalities, our UVLTrack can simultaneously handle all target reference settings (BBOX, NL, NL+BBOX).

language to specify the tracking target. **Tracking by natural language specification** (NL) provides a novel human-computer interaction manner, which specifies the target purely based on natural language. Li *et al.* first define this task and provides a baseline by combining a grounding model and a tracking model. Then, some trackers (Li et al. 2022; Wang et al. 2021b; Yang et al. 2020b) follow this paradigm to design different models to solve the grounding task and tracking task separately. Recently, JointNLT (Zhou et al. 2023) performs tracking and grounding using a unified model, which simplifies the overall framework and enables end-to-end optimization. **Tracking by language and box specification** (NL+BBOX) specifies the target through both the initial bounding box and natural language. Li *et al.* (Li et al. 2017) firstly introduce natural language into tracking achieving more robust results than visual tracker, which demonstrates the potential of vision-language tracking. SNLT (Feng et al. 2021) embeds natural language into Siamese-based trackers as a convolutional kernel and locates the target through cross-correlation. VLT (Guo et al. 2022) treats the natural language feature as a selector to weigh different visual feature channels, enhancing target-related channels for robust tracking. Also, JointNLT (Zhou et al. 2023) introduces natural language by interacting language and visual features in Transformer blocks. However, due to the semantic gap between vision and language, these trackers trained with natural language show limited performance in pure bounding box reference setting (Guo et al. 2022). To this end, we design a multi-modal contrastive loss to align features of different modalities into a unified semantic space. Meanwhile, a dynamic head, modality-adaptive box head, is proposed to alleviate the difficulty of static head training for different modal references. Thanks to the effective designs, our UVLTrack achieves promising performance across all reference settings with high FPS.

Method

In this section, we first introduce the overall architecture of UVLTrack, which presents a simple but effective pipeline for unified visual and vision-language tracking. The following two subsections introduce details of the modality-unified feature extractor and the modality-adaptive box head. In the last subsection, we introduce the training objectives.

Tracking Architecture

As shown in Figure 2, UVLTrack can take different modal references as input, including natural language, template, or both. The template is cropped based on the initial bounding box. Given the language description l , we tokenize the sentence and embed each word with the text embedding layer to obtain text embeddings $E_l^0 \in \mathbb{R}^{N_l \times C}$. N_l is the maximum text length. Given the template $z \in \mathbb{R}^{3 \times H_z \times W_z}$ and search region (full image for grounding) $x \in \mathbb{R}^{3 \times H_x \times W_x}$, they are split and reshaped into a sequence of flattened 2D patches and then linearly projected into latent space. Learnable position embeddings are added to the corresponding patch embeddings, obtaining template embeddings $E_z^0 \in \mathbb{R}^{N_z \times C}$ and search region embeddings $E_x^0 \in \mathbb{R}^{N_x \times C}$. N_z and N_x are the patch number of the template and search region respectively. We also prepend a language semantic token $T_l^0 \in \mathbb{R}^{1 \times C}$ and a visual semantic token $T_v^0 \in \mathbb{R}^{1 \times C}$ to text embeddings and image embeddings correspondingly, which are designed to capture the global semantics of different modalities. After that, text and image embeddings are fed into the modality-unified feature extractor, which is built based on Transformer architecture. Specifically, we extract language and visual features separately in shallow encoder layers and fuse them in deep encoder layers, which avoids the confusion in low-level feature modeling between different modalities and enables high-level semantics interaction. Moreover, a multi-modal contrastive loss is proposed to align differ-

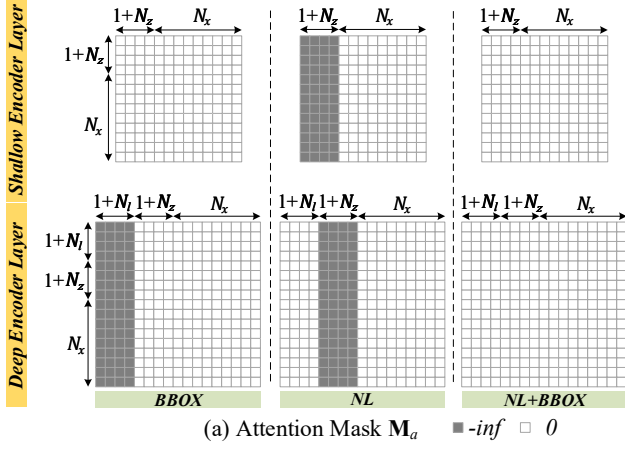


Figure 3: The attention mask of task-oriented multi-head attention for different target references.

ent modal features into a unified semantic space. Finally, we feed the enhanced text and image embeddings into the modality-adaptive box head, which can make full use of different modal references to mine ever-changing scenario features from video contexts and locate the target in a contrastive way. Further, search region embeddings with high-confidence target bounding boxes are saved as video contexts \mathbf{E}_c^{N+M} to help with subsequent target localization.

Modality-Unified Feature Extractor

As shown in Figure 2, the modality-unified feature extractor is designed based on Transformer architecture, which consists of N shallow encoder layers and M deep encoder layers. We extract visual and language features separately in shallow encoder layers and fuse them in deep encoder layers, which can avoid the confusion of low-level feature modeling for different modalities and allow high-level semantics interaction for target localization.

For parallel training of different reference inputs, we fill the unavailable (NA) reference embeddings with zeros and propose a task-oriented multi-head attention mechanism (TMHA) to avoid task-irrelevant feature interactions. We only present the single-head formulas of TMHA below for the sake of simplicity. Given the input of i^{th} encoder layer \mathbf{E}^{i-1} , key \mathbf{K}^i , query \mathbf{Q}^i and value \mathbf{V}^i arise from \mathbf{E}^{i-1} through layer normalization and linear projections. Then, we filter task-irrelevant feature interactions in attention mechanisms through masking. The output of i^{th} encoder layer can be formulated as,

$$\hat{\mathbf{E}}^i = \text{Softmax}\left(\frac{\mathbf{Q}^i(\mathbf{K}^i)^\top}{\sqrt{C}} + \mathbf{M}_a\right)\mathbf{V}^i + \mathbf{E}^{i-1}, \quad (1)$$

$$\mathbf{E}^i = \text{MLP}(\text{LN}(\hat{\mathbf{E}}^i)) + \hat{\mathbf{E}}^i, \quad (2)$$

where $\text{MLP}(\cdot)$ is multi-layer perception, $\text{LN}(\cdot)$ is layer normalization, $\hat{\mathbf{E}}^i$ is a intermediate variable, \mathbf{E}^i is the output of i^{th} encoder layer. \mathbf{M}_a is the attention mask, which is related to the input reference type. Figure 3 shows the details of the attention mask for different target references.

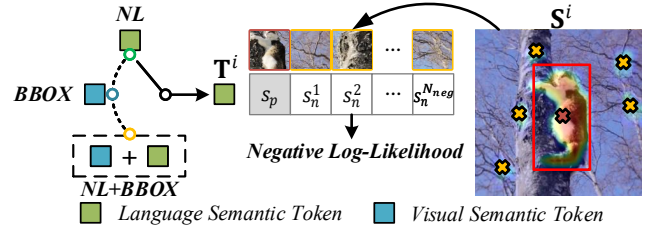


Figure 4: The diagram of the multi-modal contrastive loss.

Further, we propose a multi-modal contrastive (MMC) loss to align different modal features into a unified semantic space. As shown in Figure 4, given the semantic token \mathbf{T}^i of i^{th} encoder layer, we compute the similarity $\mathbf{S}^i = [s^{i,1}; s^{i,2}, \dots, s^{i,N_x}]$ between \mathbf{T}^i and search region embeddings $\mathbf{E}_x^i = [f^{i,1}; f^{i,2}, \dots, f^{i,N_x}]$. Formally,

$$s^{i,j} = \text{sim}(\mathbf{T}^i, f^{i,j}) / \tau, \quad \text{sim}(\mathbf{T}^i, f^{i,j}) = \frac{\mathbf{T}^i (f^{i,j})^\top}{\|\mathbf{T}^i\|_2 \|f^{i,j}\|_2}, \quad (3)$$

where τ is a temperature parameter, $\|\cdot\|_2$ means l_2 norm. According to \mathbf{S}^i , we select the central score of the target s_p^i as positive sample score and top N_{neg} scores out of the target box $[s_n^{i,k}]_{k=1}^{N_{neg}}$ as negative sample scores. Finally, the multi-modal contrastive loss can be formulated as follows,

$$\mathcal{L}_{mmc}^i = -\log\left(\frac{e^{s_p^i}}{e^{s_p^i} + \sum_{k=1}^{N_{neg}} e^{s_n^{i,k}}}\right). \quad (4)$$

Different modal features can be aligned to a unified semantic space in a contrastive way, which helps consistent feature learning for different modal references.

Modality-Adaptive Box Head

Inspired by OStrack (Ye et al. 2022a), we reshape the reference enhanced embeddings of search region \mathbf{E}_x^{N+M} into a 2D feature map and feed it into a three-branch convolutional network to regress a center score map $\hat{\mathbf{C}} \in (0, 1)^{\frac{H_x}{p} \times \frac{W_x}{p}}$, an offset map $\hat{\mathbf{O}} \in [0, 1]^{2 \times \frac{H_x}{p} \times \frac{W_x}{p}}$ and a normalized box size map $\hat{\mathbf{S}} \in (0, 1)^{2 \times \frac{H_x}{p} \times \frac{W_x}{p}}$, where p is the size of image patches. However, we find that the center score map is unstable for different modal references, which seriously affects tracking robustness. The underlying reason is that various reference modalities increase the difficulty of static head training. Thus, we further propose a dynamic head, modality-adaptive box head (MABH), which can make full use of reference information to mine ever-changing scenario features from video context to discriminate the target.

As shown in Figure 5(c), we treat the semantic token in the last encoder layer as the target prototype $\hat{\mathbf{P}}_t = \mathbf{T}^{N+M}$ and introduce learnable distractor prototype $\hat{\mathbf{P}}_d$ and background prototype $\hat{\mathbf{P}}_b$ to locate the target in a contrastive way. For tracking tasks, video contexts can provide rich scenario cues to discriminate the target. Thus, as shown in Figure 5(b), we propose a novel distribution-based attention mechanism to mine features of the target, distractor and background from historic frames. Given the template and context embeddings $\mathbf{E}_t = [\mathbf{E}_z^{N+M}; \mathbf{E}_c^{N+M}] \in$

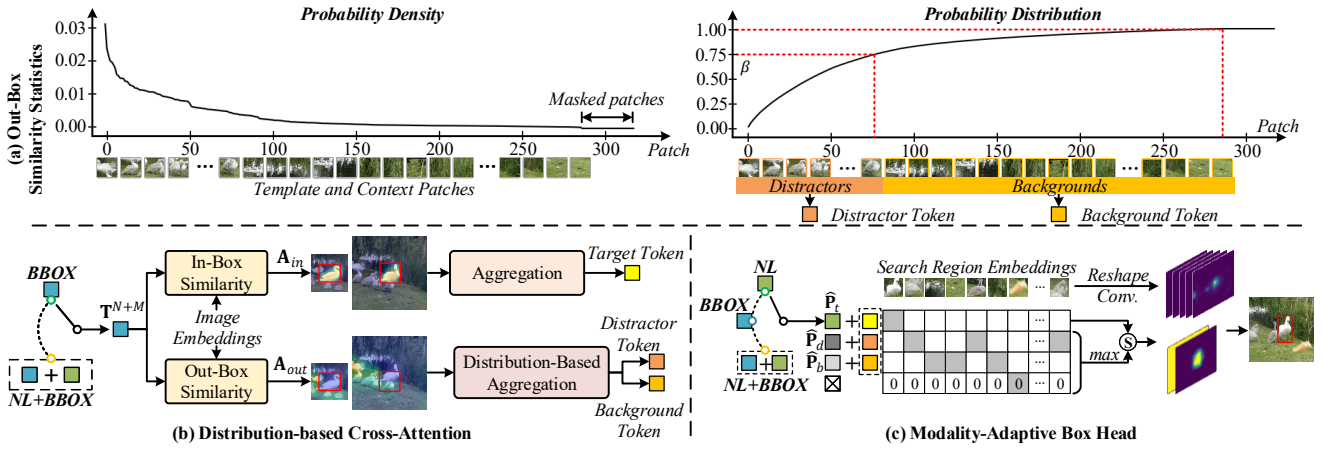


Figure 5: (a) shows the out-box similarity statistics. (b) shows the structure of the distribution-based cross-attention. (c) shows the schematic of the modality-adaptive box head, which can make full use of reference information to discriminate the target.

$\mathbb{R}^{(N_z+N_x) \times C}$ and the target masks $\mathbf{M}_t = [\mathbf{M}_z; \mathbf{M}_c] \in \mathbb{R}^{1 \times (N_z+N_x)}$, we compute in-box similarity \mathbf{A}_{in} and out-box similarity \mathbf{A}_{out} between the target semantic token \mathbf{T}^{N+M} and \mathbf{E}_t to obtain the probability that the patch belongs to the target. Formally,

$$\mathbf{A}_{in} = \text{Softmax}\left(\frac{\mathbf{T}^{N+M} \mathbf{E}_t^\top}{\sqrt{C}} + \mathbf{M}_t\right), \quad (5)$$

$$\mathbf{A}_{out} = \text{Softmax}\left(\frac{\mathbf{T}^{N+M} \mathbf{E}_t^\top}{\sqrt{C}} + \tilde{\mathbf{M}}_t\right), \quad (6)$$

where \mathbf{M}_t is obtained by assigning the position in the target box to 0 and the position out of the target box to $-inf$. $\tilde{\mathbf{M}}_t$ is the opposite. Then, the target token \mathbf{T}_t is obtained by in-box similarity aggregation $\mathbf{T}_t = \mathbf{A}_{in} \mathbf{E}_t$. Considering that the distractor with a similar appearance to the target is a key factor affecting the tracking robustness (Mayer et al. 2021), we divide features out of the target box into distractor and background from the perspective of distribution. Specifically, as shown in Figure 5(a), we rank the out-box probabilities \mathbf{A}_{out} in descending order and sum them cumulatively to obtain the probability distribution. We divide patches by threshold β to obtain the distractor mask \mathbf{M}_d and background mask $\tilde{\mathbf{M}}_d$. For distractor mask \mathbf{M}_d , we assign the patch whose target probability distribution score is lower than β to 0 and other positions to $-inf$. $\tilde{\mathbf{M}}_d$ is the opposite. Then, the distractor token and background token can be formulated as,

$$\mathbf{T}_d = \text{Softmax}\left(\frac{\mathbf{T}^{N+M} \mathbf{E}_t^\top}{\sqrt{C}} + \tilde{\mathbf{M}}_t + \mathbf{M}_d\right) \mathbf{E}_t, \quad (7)$$

$$\mathbf{T}_b = \text{Softmax}\left(\frac{\mathbf{T}^{N+M} \mathbf{E}_t^\top}{\sqrt{C}} + \tilde{\mathbf{M}}_t + \tilde{\mathbf{M}}_d\right) \mathbf{E}_t. \quad (8)$$

After obtaining $\mathbf{T}_t, \mathbf{T}_d, \mathbf{T}_b$, we add them to scenario prototypes $\hat{\mathbf{P}}_t, \hat{\mathbf{P}}_d, \hat{\mathbf{P}}_b$ to supplement the dynamic scenario information. Given search region embeddings $\mathbf{E}_x^{N+M} = [f^1; f^2; \dots; f^{N_x}]$, we compute the corresponding target sim-

ilarity $\hat{\mathbf{L}} = [\alpha_t^1, \alpha_t^2, \dots, \alpha_t^{N_x}]$ as follows,

$$\mathbf{P}_t = \hat{\mathbf{P}}_t + \mathbf{T}_t, \mathbf{P}_d = \hat{\mathbf{P}}_d + \mathbf{T}_d, \mathbf{P}_b = \hat{\mathbf{P}}_b + \mathbf{T}_b, \quad (9)$$

$$\hat{\alpha}_t^i = \text{sim}(f^i, \mathbf{P}_t) / \tau, \quad (10)$$

$$\hat{\alpha}_b^i = \max(\text{sim}(f^i, \mathbf{P}_d) / \tau, \text{sim}(f^i, \mathbf{P}_b) / \tau, 0), \quad (11)$$

$$\alpha_t^i = e^{\hat{\alpha}_t^i} / (e^{\hat{\alpha}_t^i} + e^{\hat{\alpha}_b^i}). \quad (12)$$

Here, we append a zero to the background score computation, which avoids unseen objects having relatively high target score α_t^i after the softmax operation. Finally, given the position $(x_c, y_c) = \text{argmax}_{(x,y)} \hat{\mathbf{C}}(x,y) \hat{\mathbf{L}}(x,y)$, the bounding box of target $\hat{b} = (\hat{x}, \hat{y}, \hat{w}, \hat{h})$ can be formulated as,

$$(\hat{x}, \hat{y}) = \left((x_c + \hat{\mathbf{O}}(0, x_c, y_c)) \cdot p, (y_c + \hat{\mathbf{O}}(1, x_c, y_c)) \cdot p \right), \quad (13)$$

$$(\hat{w}, \hat{h}) = (\hat{\mathbf{S}}(0, x_c, y_c) \cdot H_x, \hat{\mathbf{S}}(1, x_c, y_c) \cdot W_x). \quad (14)$$

Training Objective

We treat patches in the target box as positive samples and others as negative samples to generate the groundtruth \mathbf{L} for target score map $\hat{\mathbf{L}}$. Then, the binary cross-entropy loss is adopted for the target score map constraint. Formally,

$$\mathcal{L}_{tgt} = \mathcal{L}_{bce}(\hat{\mathbf{L}}, \mathbf{L}) \quad (15)$$

The training objectives of center score map \mathcal{L}_{cls} and bounding box $\mathcal{L}_{box} = \lambda_1 \mathcal{L}_1 + \lambda_{giou} \mathcal{L}_{giou}$ are consistent with OS-Track (Ye et al. 2022a). In summary, the overall objective function can be formulated as,

$$\mathcal{L} = \mathcal{L}_{tgt} + \mathcal{L}_{cls} + \mathcal{L}_{box} + \lambda_{mmc} \sum_{i=1}^{N+M} \mathcal{L}_{mmc}^i. \quad (16)$$

Experiment

Our tracker is implemented using Python 3.8.13 and Pytorch 1.10.1. The experiments are conducted on a server with eight 24GB NVIDIA RTX 3090 GPUs. Visualization and qualitative results are present in **Supplementary Materials**.

Method	TNL2K		AVisT			LaSOT		LaSOT _{ext}		TrackingNet	
	AUC	P	AUC	OP _{0.5}	OP _{0.75}	AUC	P	AUC	P	AUC	P
Performance-oriented Variants											
UVLTrack-L	64.8	68.8	57.8	67.9	48.7	71.3	78.3	51.2	59.0	84.1	82.9
OTrack-384 (Ye et al. 2022a)	<u>55.9</u>	-	-	-	-	71.1	77.6	<u>50.5</u>	<u>57.6</u>	83.9	83.2
MixFormer-L (Cui et al. 2022)	-	-	<u>56.0</u>	<u>65.9</u>	<u>46.3</u>	70.1	76.3	-	-	<u>83.9</u>	<u>83.1</u>
SimTrack-L/14 (Chen et al. 2022a)	55.6	55.7	-	-	-	70.5	-	-	-	83.4	-
Basic Variants											
UVLTrack-B	62.7	65.4	56.5	66.0	45.1	69.4	74.9	49.2	55.8	83.4	82.1
OTrack-256 (Ye et al. 2022a)	54.3	-	-	-	-	69.1	75.2	47.4	53.3	<u>83.1</u>	<u>82.0</u>
MixFormer-22k (Cui et al. 2022)	-	-	<u>53.7</u>	<u>63.0</u>	<u>43.0</u>	69.2	74.7	-	-	<u>83.1</u>	81.6
SimTrack-B/16 (Chen et al. 2022a)	<u>54.8</u>	<u>53.8</u>	-	-	-	<u>69.3</u>	-	-	-	82.3	-
AiATrack (Gao et al. 2022)	-	-	-	-	-	69.0	73.8	<u>47.7</u>	<u>55.4</u>	82.7	80.4
STARK (Yan et al. 2021)	-	-	51.1	59.2	39.1	66.4	71.2	-	-	81.3	78.1
TransT (Chen et al. 2021b)	50.7	51.7	49.0	56.4	37.2	64.9	73.8	-	-	81.4	80.3
TrDiMP (Wang et al. 2021a)	-	-	48.1	55.3	33.8	63.9	61.4	-	-	78.4	73.1
STMTrack (Fu et al. 2021)	-	-	-	-	-	60.6	63.3	-	-	80.3	76.7
PrDiMP (Danelljan et al. 2020)	47.0	45.9	43.3	48.0	28.7	59.8	60.8	-	-	75.8	70.4
SiamFC++ (Xu et al. 2020)	38.6	36.9	-	-	-	54.4	54.7	-	-	75.4	70.5
Ocean (Zhang et al. 2020)	38.4	37.7	38.9	43.6	20.5	56.0	56.6	-	-	-	-
DiMP (Bhat et al. 2019)	44.7	43.4	41.9	45.7	26.0	56.9	56.7	39.2	45.1	74.0	68.7
SiamRPN++ (Li et al. 2019)	41.3	41.2	39.0	43.5	21.2	49.6	49.1	34.0	39.6	73.3	69.4
SiamFC (Bertinetto et al. 2016)	29.5	28.6	-	-	-	33.6	33.9	23.0	26.9	57.1	53.3

Table 1: Comparison with state-of-the-art visual trackers on TNL2K, AVisT, LaSOT, LaSOT_{ext} and TrackingNet. The best two results are shown in bold and underline

	Ocean	DiMP-50	PrDiMP-50	TransT	OTrack-256	OTrack-384	UVLTrack-B	UVLTrack-L
NFS	49.4	61.8	63.5	65.3	64.7	<u>66.5</u>	65.9	67.6
UAV123	57.4	64.3	68.0	68.1	68.3	<u>70.7</u>	69.3	71.0

Table 2: Comparison with state-of-the-art trackers on NFS, UAV123 datasets in terms of overall AUC score. The best two results are shown in bold and underline

Implementation Details

Network Details. We crop the template and search region by 2^2 and 4^2 times the target bounding box area and resize them to 128×128 and 256×256 respectively. The test image for first frame grounding is scaled such that its long edge is 256. Image patch size $p=16$. For language, the max length of the sentence N_l is set to 40. To demonstrate the scalability of UVLTrack, we present two variants, termed UVLTrack-B and UVLTrack-L. The number of encoder layers is set to $N=6, M=6$ for UVLTrack-B and $N=12, M=12$ for UVLTrack-L. The language branch in shallow encoder layers is initialized with uncased parameters of BERT (Devlin et al. 2019). Other parameters in the modality-unified feature extractor are initialized with ViT parameters pre-trained by MAE (He et al. 2022). The modality-adaptive box head is initialized with Xavier init (Glorot et al. 2010).

Training Details. We train our model on the training splits of LaSOT (Fan et al. 2019), GOT-10k (Huang et al. 2019), COCO2017 (Lin et al. 2014), TrackingNet (Muller et al. 2018), TNL2K (Wang et al. 2021b), OTB99 (Li et al. 2017) and RefCOCOg-google (Mao et al. 2016). Common data augmentation is used for model training, such as translation, horizontal flip, and color jittering. Due to the flexibility of our framework, different modal references can be trained jointly, which provides a neat training pipeline. The

loss weights are set to $\lambda_{giou}=2.0$, $\lambda_1=5.0$, $\lambda_{mmc}=0.1$.

State-of-the-art Comparisons

Visual Tracking. We evaluate our trackers on seven visual tracking benchmarks, including TNL2K (Wang et al. 2021b), AVisT (Noman et al. 2022), LaSOT (Fan et al. 2019), LaSOT_{ext} (Fan et al. 2021), TrackingNet (Muller et al. 2018), NFS (Kiani et al. 2017) and UAV123 (Mueller et al. 2016), which are commonly used for visual tracker evaluation. The Area Under the Curve (AUC) of the success plot is the main metric to rank trackers. As shown in Table 1 and 2, our UVLTrack-B outperforms the best visual tracking counterparts. Further, UVLTrack-L achieves new state-of-the-art performance on all seven visual tracking benchmarks. These results demonstrate the effectiveness of UVLTrack under the bounding box reference setting (**BBOX**).

Discussion. We evaluate recent vision-language trackers initialized by the target bounding box on LaSOT. JointNLT achieves 54.5% AUC and VLTTT achieves 53.4% AUC. These vision-language trackers show limited performance without natural language. This is because these trackers ignore the semantic gap between different modalities and tend to rely on semantic information in language. Differently, UVLTrack aligns vision and language into a unified semantic space, providing a unified tracking framework, which de-

Method	RefCOCO			RefCOCO+			RefCOCOg		
	<i>val</i>	<i>testA</i>	<i>testB</i>	<i>val</i>	<i>testA</i>	<i>testB</i>	<i>val-g</i>	<i>val-u</i>	<i>test-u</i>
UVLTrack*	85.47	87.56	81.73	74.60	79.70	65.64	73.86	75.94	74.86
VLTVG	84.77	87.24	80.49	74.19	78.93	65.17	72.98	76.04	74.18
SeqTR	83.72	86.51	81.24	71.45	76.26	64.88	71.50	74.86	74.21
QRNet	84.01	85.85	82.34	72.94	76.17	63.81	71.89	73.03	72.52
TransVG	81.02	82.72	78.35	64.82	70.70	56.94	67.02	68.67	67.73
Ref-NMS	80.70	84.00	76.04	68.25	73.68	59.42	-	70.55	70.62
LBYL-Net	79.67	82.91	74.15	68.64	73.38	59.49	62.70	-	-
ReSC-Large	77.63	80.45	72.30	63.59	68.36	56.81	63.12	67.30	67.20
NMTree	76.41	81.21	70.09	66.46	72.02	57.52	64.62	65.87	66.44

Table 3: Comparison with state-of-the-art grounding methods on RefCOCO, RefCOCO+, RefCOCOg datasets.

Method	TNL2K		LaSOT		OTB99	
	AUC	P	AUC	P	AUC	P
NL						
UVLTrack-L	58.2	60.9	59.6	63.9	63.5	83.2
UVLTrack-B	55.7	57.2	57.2	61.0	60.1	79.1
JointNLT	54.6	55.0	56.9	59.3	59.2	77.6
CTRNL	14.0	9.0	52.0	51.0	53.0	72.0
TNL2K-1	11.0	6.0	51.0	49.0	19.0	24.0
GTI	-	-	47.8	47.6	58.1	73.2
TNLS-II	-	-	-	-	25.0	29.0
NL+BBOX						
UVLTrack-L	64.9	69.3	71.4	78.7	71.1	92.0
UVLTrack-B	63.1	66.7	69.4	75.9	69.3	89.9
JointNLT	56.9	58.1	60.4	63.6	65.3	85.6
VLTTT	53.1	53.3	67.3	72.1	76.4	93.1
SNLT	27.6	41.9	54.0	57.6	66.6	80.4
TNL2K-2	42.0	42.0	51.0	55.0	68.0	88.0
TNLS-III	-	-	-	-	55.0	72.0

Table 4: Comparison with state-of-the-art vision-language trackers on LaSOT, TNL2K and OTB99 datasets.

livers superior performance for all three reference settings.

Vision-Language Tracking. We further evaluate our UVLTrack on vision-language tracking benchmarks, including TNL2K (Wang et al. 2021b), LaSOT (Fan et al. 2019) and OTB99 (Li et al. 2017) and compare with the latest trackers, including JointNLT (Zhou et al. 2023), CTRNL (Li et al. 2022), TNL2K (Wang et al. 2021b), GTI (Yang et al. 2020b), TNLS (Li et al. 2017), VLTTT (Guo et al. 2022), SNLT (Feng et al. 2021). As shown in Table 4, when specifying the target by natural language (NL), UVLTrack-L surpasses the previous best tracker JointNLT with a large margin on three benchmarks. Also, our UVLTrack-L achieves the best performance on TNL2K and LaSOT when initializing the tracker with both natural language and the bounding box (NL+BBOX). These results demonstrate the superiority of UVLTrack for vision-language tracking.

Efficiency. UVLTrack-B runs at 58 FPS for visual tracking and 57 FPS for vision-language tracking. UVLTrack-L runs at 28 FPS for visual tracking and 27 FPS for vision-language tracking. Compared with JointNLT (39 FPS), UVLTrack-B achieves better performance with $1.46\times$ speed.

Visual Grounding. Like VLTVG, we retrain UVLTrack-B on train sets of RefCOCO (Yu et al. 2016), RefCOCO+ (Yu

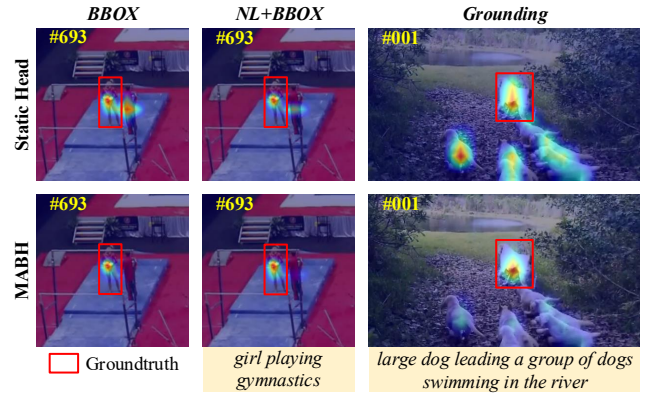


Figure 6: Visualization of target localization results.

Method	TNL2K					
	BBOX		NL		NL+BBOX	
	AUC	P	AUC	P	AUC	P
baseline	59.4	59.9	51.6	51.8	59.7	60.6
+MMCLoss	60.6	62.8	53.8	54.7	62.0	64.9
+MABH	62.7	65.4	55.7	57.2	63.1	66.7

Table 5: Analysis of different components in UVLTrack.

et al. 2016) and RefCOCOg (Mao et al. 2016) separately, and report the Top-1 accuracy on corresponding test sets in Table 3. We compared UVLTrack with the latest grounding methods, including VLTVG (Yang et al. 2022), SeqTR (Zhu et al. 2022), QRNet (Ye et al. 2022b), TransVG (Deng et al. 2021), Ref-NMS (Chen et al. 2021a), LBYL-Net (Huang et al. 2021), ReSC-Large (Yang et al. 2020a), NMTree (Liu et al. 2019). The test image is scaled such that its long edge is 384. Compared to visual grounding methods, our UVLTrack achieves the best performance on seven test sets, demonstrating the generalization ability of our framework.

Ablation Study

The following experiments use UVLTrack-B as the base model. The baseline is UVLTrack without MMCLoss constraint and using the static anchor-free head for localization.

Effectiveness of the Different Components. Table 5 shows the performance of UVLTrack with different components.

N	M	TNL2K					
		BBOX		NL		NL+BBOX	
		AUC	P	AUC	P	AUC	P
0	12	61.3	63.8	55.1	56.1	62.3	65.1
3	9	61.6	64.2	55.9	57.3	63.0	66.6
6	6	62.7	65.4	55.7	57.2	63.1	66.7
9	3	62.6	65.3	54.6	55.4	62.6	65.8
11	1	62.5	65.1	46.2	42.3	61.9	64.6

Table 6: Analysis of the modality-unified feature extractor.

MMCLoss brings 1.2%, 2.2% and 2.3% AUC gains for BBOX, NL and BBOX+NL reference settings respectively. This is because MMCLoss can align different modal features into a unified semantic space, which enables consistent feature learning for different reference modalities. The modality-adaptive box head (MABH) brings 2.1%, 1.9% and 1.1% AUC gains for BBOX, NL and BBOX+NL reference settings respectively. As shown in Figure 6, the static anchor free head shows unstable target localization results. The reason is that various reference modalities increase the difficulty of static head training, which leads to compromised results. Differently, we design a dynamic head (MABH), which can make full use of reference information to locate the target in a contrastive way, improving tracking performance across all reference settings.

Analysis of the Modality-Unified Feature Extractor. As shown in Table 6, more separate layers (larger N) are beneficial for visual tracking and more fusion layers (larger M) are beneficial for vision-language tracking. However, when we fuse visual and language features in all encoder layers, the performance is suboptimal for vision-language tracking. This is because early fusion breaks the low-level feature modeling for different modalities. Thus, we set $N=6$ and $M=6$ to balance the performance for all reference settings.

Analysis of the Multi-Modal Contrastive Loss. We study different ways to obtain the positive sample and the negative sample. As shown in Table 7, the best results are achieved when we sample the central score of the target as the positive sample and the top 9 scores out of the target box as negative samples. The underlying reason is that the central feature of the target contains no backgrounds, which is more reliable to express the target. Further, more hard-negative samples can improve the discriminability of the semantic token and align different modal features into a compact semantic space.

Analysis of the Distractor Threshold. As shown in Table 8, the performance of UVLTrack is insensitive to threshold β over a wide range. However, the performance drops overtly if we aggregate all background features into one token ($\beta=0$). This is because the distractor feature is vital to discriminate the target in complex scenarios. If we aggregate all background features together, the distractor features will be smoothed by other background features, which is not conducive for the tracker to distinguish distractors.

Analysis of the Training Strategy. As shown in Table 9, the ratio of different references (BBOX:NL:Both) has little effect on the performance of UVLTrack. Moreover, we try to initialize all parameters in the backbone with ViT pa-

pos.	neg.	N_{neg}	TNL2K					
			BBOX		NL		NL+BBOX	
			AUC	P	AUC	P	AUC	P
<i>avg</i>	<i>rand</i>	9	61.4	63.8	53.6	54.0	61.5	64.2
<i>ctr</i>	<i>rand</i>	9	61.8	64.5	54.9	55.9	62.1	65.4
<i>avg</i>	<i>top</i>	9	62.1	64.6	55.2	56.3	62.3	65.5
<i>ctr</i>	<i>top</i>	9	62.7	65.4	55.7	57.2	63.1	66.7
<i>ctr</i>	<i>top</i>	1	61.6	64.1	54.5	55.4	61.8	65.1
<i>ctr</i>	<i>top</i>	5	62.3	64.9	55.4	56.7	62.6	66.0
<i>ctr</i>	<i>top</i>	13	62.6	65.4	55.6	57.0	63.1	66.6

Table 7: Analysis of the multi-modal contrastive loss design.

β	TNL2K					
	BBOX		NL		NL+BBOX	
	AUC	P	AUC	P	AUC	P
0.00	61.3	63.8	54.3	55.3	61.8	64.7
0.25	62.1	64.7	55.2	56.5	62.4	65.8
0.50	62.5	65.3	55.4	56.8	62.8	66.3
0.75	62.7	65.4	55.7	57.2	63.1	66.7
0.85	62.6	65.3	55.3	56.6	62.7	66.2

Table 8: Analysis of the distractor threshold β in the distribution-based cross-attention mechanism.

Pretrain	Ratio	TNL2K					
		BBOX		NL		NL+BBOX	
		AUC	P	AUC	P	AUC	P
BERT+MAE	3:1:3	62.3	65.0	55.6	57.0	62.8	66.5
BERT+MAE	4:1:4	62.7	65.4	55.7	57.2	63.1	66.7
BERT+MAE	5:1:5	62.8	65.6	55.2	56.5	63.2	67.0
MAE	4:1:4	62.1	64.2	53.8	54.5	61.9	65.1

Table 9: Analysis of the training strategy.

rameters pretrained by MAE, which reduces overall performance. This is because pretrained BERT parameters bring initial language modeling capabilities to the tracker to understand natural language.

Conclusion

In this work, we propose a novel unified tracker (UVLTrack) for visual and vision-language tracking, which can simultaneously cope with three types of target reference (BBOX, NL, NL+BBOX). Specifically, we design a modality-unified feature extractor for joint visual and language feature learning, and propose a multi-modal contrastive loss to align different modal features into a unified semantic space. Further, a modality-adaptive box head is proposed to localize the target dynamically with scenario features, enabling robust performance for different reference settings. UVLTrack achieves promising results on seven visual tracking, three vision-language tracking, three visual grounding datasets.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62306294, 62071122, 62121002) and Youth Innovation Promotion Association CAS (2018166).

References

- Alper, Y.; Omar, J.; Mubarak, S.; et al. 2006. Object tracking: A survey. *ACM Computing Surveys*, 38(4): 13–es.
- Bertinetto, L.; Valmadre, J.; Henriques, J. F.; Vedaldi, A.; and Torr, P. H. S. 2016. Fully-Convolutional Siamese Networks for Object Tracking. In *Proceedings of the European Conference on Computer Vision Workshops*.
- Bhat, G.; Danelljan, M.; Gool, L. V.; and Timofte, R. 2019. Learning Discriminative Model Prediction for Tracking. In *Proceedings of the IEEE International Conference on Computer Vision*.
- Chen, B.; Li, P.; Bai, L.; Qiao, L.; Shen, Q.; Li, B.; Gan, W.; Wu, W.; and Ouyang, W. 2022a. Backbone is All Your Need: A Simplified Architecture for Visual Object Tracking. In *Proceedings of the European Conference on Computer Vision*.
- Chen, F.; Wang, X.; Zhao, Y.; Lv, S.; and Niu, X. 2022b. Visual object tracking: A survey. *Computer Vision and Image Understanding*, 222: 103508.
- Chen, L.; Ma, W.; Xiao, J.; Zhang, H.; and Chang, S.-F. 2021a. Ref-nms: Breaking proposal bottlenecks in two-stage referring expression grounding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 1036–1044.
- Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; and Lu, H. 2021b. Transformer Tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
- Cui, Y.; Cheng, J.; Wang, L.; and Wu, G. 2022. MixFormer: End-to-End Tracking with Iterative Mixed Attention. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
- Danelljan, M.; Bhat, G.; Khan, F. S.; and Felsberg, M. 2019. ATOM: Accurate Tracking by Overlap Maximization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
- Danelljan, M.; Gool, L. V.; Timofte, R.; et al. 2020. Probabilistic regression for visual tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
- Deng, J.; Yang, Z.; Chen, T.; Zhou, W.; and Li, H. 2021. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1769–1779.
- Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.
- Fan, H.; Bai, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Huang, M.; Liu, J.; Xu, Y.; et al. 2021. Lasot: A high-quality large-scale single object tracking benchmark. *International Journal of Computer Vision*, 129(2): 439–461.
- Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H.; Xu, Y.; Liao, C.; and Ling, H. 2019. LaSOT: A High-Quality Benchmark for Large-Scale Single Object Tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
- Feng, Q.; Ablavsky, V.; Bai, Q.; and Sclaroff, S. 2021. Siamese natural language tracker: Tracking by natural language descriptions with siamese trackers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5851–5860.
- Fu, Z.; Liu, Q.; Fu, Z.; and Wang, Y. 2021. Stmtrack: Template-free visual tracking with space-time memory networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 13774–13783.
- Gao, S.; Zhou, C.; Ma, C.; Wang, X.; and Yuan, J. 2022. Aia-track: Attention in attention for transformer visual tracking. In *Proceedings of the European Conference on Computer Vision*, 146–164. Springer.
- Glorot, X.; et al. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*.
- Guo, D.; Wang, J.; Cui, Y.; Wang, Z.; and Chen, S. 2020. SiamCAR: Siamese Fully Convolutional Classification and Regression for Visual Tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
- Guo, M.; Zhang, Z.; Fan, H.; and Jing, L. 2022. Divert more attention to vision-language tracking. *Advances in Neural Information Processing Systems*, 35: 4446–4460.
- He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 16000–16009.
- Huang, B.; Lian, D.; Luo, W.; and Gao, S. 2021. Look before you leap: Learning landmark features for one-stage visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16888–16897.
- Huang, L.; Zhao, X.; Huang, K.; et al. 2019. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Kiani, H.; Galoogahi, F.; Fagg, A.; Huang, C.; Ramanan, D.; Lucey, S.; et al. 2017. Need for speed: A benchmark for higher frame rate object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*.
- Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; and Yan, J. 2019. SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
- Li, Y.; Yu, J.; Cai, Z.; and Pan, Y. 2022. Cross-modal target retrieval for tracking by natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4931–4940.
- Li, Z.; Tao, R.; Gavves, E.; Snoek, C. G.; and Smeulders, A. W. 2017. Tracking by natural language specification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6495–6503.
- Lin, T.-Y.; Maire, M.; Belongie, S. J.; Bourdev, L. D.; Girshick, R. B.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision*.

- Liu, D.; Zhang, H.; Wu, F.; and Zha, Z.-J. 2019. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4673–4682.
- Ma, Y.; He, J.; Yang, D.; Zhang, T.; and Wu, F. 2023. Adaptive Part Mining for Robust Visual Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11–20.
- Mayer, C.; Danelljan, M.; Paudel, D. P.; and Van Gool, L. 2021. Learning target candidate association to keep track of what not to track. In *Proceedings of the IEEE International Conference on Computer Vision*, 13444–13454.
- Mueller, M.; Smith, N.; Ghanem, B.; et al. 2016. A benchmark and simulator for UAV tracking. In *Proceedings of the European Conference on Computer Vision*.
- Muller, M.; Bibi, A.; Giancola, S.; Alsubaihi, S.; and Ghanem, B. 2018. TrackingNet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European Conference on Computer Vision*.
- Noman, M.; Ghallabi, W. A.; Najiha, D.; Mayer, C.; Dudhane, A.; Danelljan, M.; Cholakkal, H.; Khan, S.; Van Gool, L.; and Khan, F. S. 2022. Avist: A benchmark for visual object tracking in adverse visibility. *arXiv preprint arXiv:2208.06888*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Advances of Neural Information Processing Systems*.
- Wang, N.; Zhou, W.; Wang, J.; and Li, H. 2021a. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1571–1580.
- Wang, X.; Shu, X.; Zhang, Z.; Jiang, B.; Wang, Y.; Tian, Y.; and Wu, F. 2021b. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 13763–13773.
- Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; and Zhang, L. 2021. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, 22–31.
- Xu, Y.; Wang, Z.; Li, Z.; Yuan, Y.; and Yu, G. 2020. SiamFC++: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Yan, B.; Peng, H.; Fu, J.; Wang, D.; and Lu, H. 2021. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, 10448–10457.
- Yang, L.; Xu, Y.; Yuan, C.; Liu, W.; Li, B.; and Hu, W. 2022. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9499–9508.
- Yang, Z.; Chen, T.; Wang, L.; and Luo, J. 2020a. Improving one-stage visual grounding by recursive sub-query construction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 387–404. Springer.
- Yang, Z.; Kumar, T.; Chen, T.; Su, J.; and Luo, J. 2020b. Grounding-tracking-integration. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(9): 3433–3443.
- Ye, B.; Chang, H.; Ma, B.; and Shan, S. 2022a. Joint Feature Learning and Relation Modeling for Tracking: A One-Stream Framework. *Proceedings of the European Conference on Computer Vision*.
- Ye, J.; Tian, J.; Yan, M.; Yang, X.; Wang, X.; Zhang, J.; He, L.; and Lin, X. 2022b. Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15502–15512.
- Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, 69–85. Springer.
- Zhang, Z.; Peng, H.; Fu, J.; Li, B.; and Hu, W. 2020. Ocean: Object-aware Anchor-free Tracking. In *Proceedings of the European Conference on Computer Vision*.
- Zhou, L.; Zhou, Z.; Mao, K.; and He, Z. 2023. Joint Visual Grounding and Tracking with Natural Language Specification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23151–23160.
- Zhu, C.; Zhou, Y.; Shen, Y.; Luo, G.; Pan, X.; Lin, M.; Chen, C.; Cao, L.; Sun, X.; and Ji, R. 2022. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, 598–615. Springer.