UniSOT: A Unified Framework for Multi-Modality Single Object Tracking

Yinchao Ma, Yuyang Tang, Wenfei Yang, Tianzhu Zhang*, Xu Zhou, and Feng Wu

Abstract—Single object tracking aims to localize target object with specific reference modalities (bounding box, natural language or both) in a sequence of specific video modalities (RGB, RGB+Depth, RGB+Thermal or RGB+Event.). Different reference modalities enable various human-machine interactions, and different video modalities are demanded in complex scenarios to enhance tracking robustness. Existing trackers are designed for single or several video modalities with single or several reference modalities, which leads to separate model designs and limits practical applications. Practically, a unified tracker is needed to handle various requirements. To the best of our knowledge, there is still no tracker that can perform tracking with these above reference modalities across these video modalities simultaneously. Thus, in this paper, we present a unified tracker, UniSOT, for different combinations of three reference modalities and four video modalities with uniform parameters. Extensive experimental results on 18 visual tracking, vision-language tracking and RGB+X tracking benchmarks demonstrate that UniSOT shows superior performance against modality-specific counterparts. Notably, UniSOT outperforms previous counterparts by over 3.0% AUC on TNL2K across all three reference modalities and outperforms Un-Track by over 2.0% main metric across all three RGB+X video modalities.

Index Terms—Single object tracking, unified framework, reference modality, video modality

1 Introduction

Single object tracking focuses on localizing target object using specific reference modalities (bounding box, natural language or both) within sequences of particular video modalities (RGB, RGB+Depth, RGB+Thermal or RGB+Event). Different reference modalities provide various human-machine interactions [1]. Different video modalities are demanded in complex scenarios to enhance robustness [2]. Over the years, great progress has been achieved in single object tracking, which has spawned a wide range of applications in robotics, autonomous driving, human-machine interaction and so on [3]. Existing trackers can be divided into different categories by reference and video modalities.

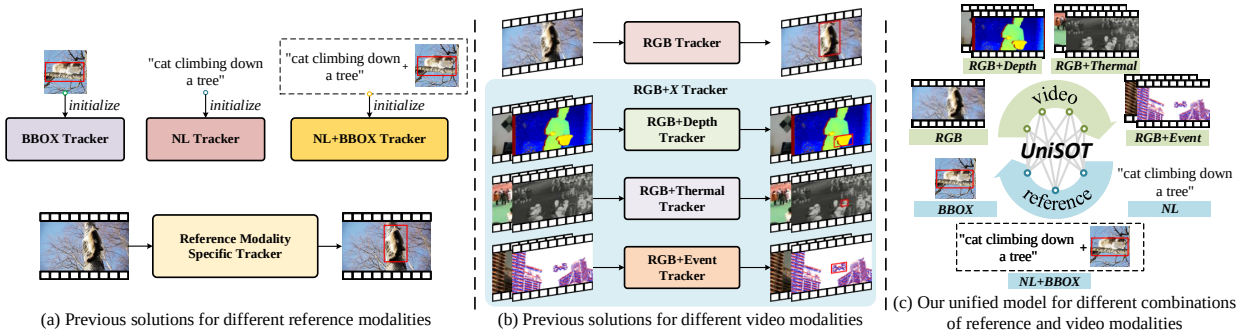
Based on *reference modalities*, trackers can be divided into three categories, as shown in Figure 1(a). **1) Bounding box** (**BBOX**). Most trackers [4], [5], [6] utilize the target bounding box as the reference. They commonly crop a template according to the given bounding box and localize target object in subsequent frames by interacting with the cropped template. **2) Natural language** (**NL**). Different from the above tracking paradigm, tracking by natural language specification [7] provides a novel manner of human-machine interaction. This task first localizes target object in the first frame based on the language description, and then tracks target object based on the language description and the predicted bounding box [8]. **3) Natural language and bounding box** (**NL+BBOX**). For providing more accurate target reference, some trackers [9], [10] specify target object by both natural language and bounding box. They embed language descriptions into visual features through dynamic filter [7], cross-correlation [10] or channel-wise attention [9], achieving more robust tracking.

Based on *video modalities*, mainstream trackers can be divided into two categories, as shown in Figure 1(b). **1) RGB**. Traditional trackers perform tracking on RGB sequences [11], [12], which contain rich appearance cues to discriminate target object. **2) RGB+X**. X means auxiliary video modality (depth, thermal or event). Recently, auxiliary modalities are drawing increasing attention in tracking tasks due to their complementary information with RGB images [13]. Most RGB+X trackers [14], [15], [16] integrate well-designed fusion modules to enhance RGB features with one specific auxiliary video modality ($e.g.$ RGB+Event), achieving more robust results in complex scenarios.

Previous reference or video modality-specific trackers lead to separate model designing, as shown in Figure 1. In practice, users may desire various reference modalities for convenience [7], while the demand for video modalities varies across scenarios [2]. It brings out the need for a unified tracker that can handle different combinations of reference and video modalities simultaneously. Some trackers attempt to unify partial reference modalities [1], [8] or video modalities [2], [17]. For example, JointNLT [8] can perform tracking with NL or NL+BBOX references on RGB sequences, and Un-Track can perform tracking with BBOX references on RGB+Depth, RGB+Thermal and RGB+Event sequences. However, there is still no tracker that can perform tracking with three reference modalities across four video modalities. The difficulties in designing such a unified tracker come from multiple aspects. From the reference modality aspect, the semantic gap across reference modalities brings the difficulty of consistent feature learning and stable object localization [9]. From the video modality aspect, it is difficult to learn video odality-aligned features across video modalities while retaining their specific cues [17].

By studying previous methods, we summarize two main issues that need to be considered to build a unified single object tracker. **1) How to design a tracking model for various reference modalities.** Recent trackers [1], [13] follow a neat but effective architecture with a feature extractor and a box head. The semantic

• **Corresponding Author. Yinchao Ma, YuYang Tang, Wenfei Yang, Tianzhu Zhang, Feng Wu are with the School of Information Science, University of Science and Technology of China, Hefei 230027, China (e-mail: imyc@mail.ustc.edu.cn; yuyangtang@mail.ustc.edu.cn, tzzhang@ustc.edu.cn; zhyd73@ustc.edu.cn; fengwu@ustc.edu.cn). Xu Zhou is with the Sangfor Research Institute.*



(a) Previous solutions for different reference modalities (b) Previous solutions for different video modalities (c) Our unified model for different combinations of reference and video modalities

Fig. 1. Comparison between previous solutions and UniSOT. There are various reference and video modalities in single object tracking tasks for different application scenarios. However, previous trackers are commonly tailored for a specific reference or video modality. (a) BBOX, NL, NL+BBOX trackers utilize the bounding box(BBOX), natural language(NL), or both(NL+BBOX) as target object reference to track on RGB sequences. (b) RGB, RGB+Depth, RGB+Thermal, RGB+Event trackers are designed for tracking in sequences of corresponding video modality with BBOX reference. (c) Unlike the customized solutions for specific reference or video modalities, we seek to design a unified tracker (UniSOT) which can perform tracking with different combinations of three reference modalities and four video modalities using uniform parameters, enabling generalized capability.

gap across reference modalities poses a challenge in consistent feature learning for the feature extractor and stable object localization for the box head. Thus, on the one hand, we need to *design a generalized feature extractor for different references.* Recent trackers introduce natural language reference through well-designed fusion modules, such as channel-wise attention [9] and Transformer blocks [8]. However, they ignore the semantic gap between different modalities, resulting in a tendency for these trackers to rely on semantic information in language references, which limits their generalization on bounding box references [9]. To this end, it is necessary to design a new reference-generalized feature extractor via vision-language alignment, achieving consistent feature learning for different reference modalities. On the other hand, we need to *design a robust box head across reference modalities.* Existing trackers design various box heads to estimate target object state, such as anchor-free head [18], corner-based head [8], and point-based head [19]. These heads commonly take the reference-enhanced features of the search region as input, and regress the target box through offline trained parameters. However, enhanced features of search region vary with reference modalities,while the box head processes all search regions identically despite variations in reference modality, which may lead to unstable results across reference modalities. Thus, we argue that it is better to design a reference-adaptive head, which can localize target object dynamically according to reference modalities for stable tracking. 2) **How to design a training strategy for various video modalities.** Recently, some RGB+X trackers [13], [17] fine-tune extra low-rank modules on frozen foundation models, avoiding overfitting on small RGB+X datasets while retaining foundational tracking capabilities. However, they commonly project auxiliary modalities into different low-rank spaces with an identical rank and fuse with RGB features. Such a design poses a challenge for video modality-aligned feature learning, and ignores the potential difference in the amount of information across video modalities that may desire various ranks. Thus, we seek to design a video modality-adaptive adaptation mechanism for video modality joint tuning, achieving both video modality-aligned and video modality-specific feature learning.

Motivated by the above discussions, we propose a unified tracking framework, termed UniSOT, with two designing modules for reference and video modality unification, respectively. **For reference modality designing**, we propose a reference-generalized feature extractor and a reference-adaptive box head, which are trained on large-scale RGB tracking datasets with different reference modalities. The *reference-generalized feature extractor* is a new architecture constructed based on Transformer [20], in which we extract features of vision and language 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 reference 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 reference modalities. The *reference-adaptive box head* dynamically mines scenario information from video contexts by reference features of different modalities and localizes target object in a contrastive way for stable tracking. Specifically, we propose a novel distribution-based cross-attention mechanism, which can adaptively mine features of target, distractor and background from historical frames by different reference modalities. Then, target object can be localized directly through feature comparison. **For video modality designing**, we freeze parameters trained on RGB tracking datasets with different reference modalities and then introduce incremental parameters to train with auxiliary video modalities. Specifically, we design a *rank-adaptive modality adaptation* (RAMA) mechanism inspired by AdaLoRA [21], which evaluates the importance of incremental parameters from video modality-shared and video modality-specific perspectives, and dynamically adjusts the parameter ranks for different video modalities. Thus, RAMA can learn not only video modality-aligned features through shared parameters but also video modality-specific features via adaptive ranks, enabling UniSOT to perform robust tracking with different auxiliary video modalities in a uniform parameter set. Extensive experimental results on 18 tracking benchmarks with various reference and video modalities demonstrate that UniSOT shows superior performance against reference or video modality-specific counterparts.

To summarize, the main contributions of this work are: (1) We propose a novel unified tracking framework for different combinations of reference and video modalities. To the best of our knowledge, UniSOT is the first tracker to support tracking with three reference modalities (NL, BBOX, NL+BBOX) in sequences of four video modalities (RGB, RGB+Depth, RGB+Thermal, RGB+Event) with uniform parameters, enabling more generalized application scenarios. (2) To address the challenges in unified tracking with different reference modalities, we design a reference-generalized feature extractor with multi-modal contrastive loss based on Transformer for consistent feature learning,

and a reference-adaptive box head to stabilize target localization via contrastive learning of dynamic scenarios. (3) To address the challenges in unified tracking with different video modalities, we propose a rank-adaptive modality adaptation mechanism inspired by AdaLoRA for both video modality-aligned and video modality-specific feature learning, which enables various video modalities in UniSOT and provides a novel training paradigm for incremental video modalities.

2 Related Work

In this section, we will introduce single object tracking methods based on different reference and video modalities.

2.1 Trackers based on Different Reference Modalities

Different reference modalities can provide information of target object from different aspects and enable various human-machine interactions. Based on the reference modalities, existing trackers can be divided into three categories, including BBOX, NL and NL+BBOX trackers.

BBOX. Most existing trackers localize target object in a video sequence according to the bounding box in the first frame [22]. These trackers commonly crop a target template from the first frame and estimate target object state in subsequent frames by interacting with the cropped template. Two-stream trackers [19], [23], [24] extract features from both template and search region and then localize target object through building feature interactions between them by cross-correlation layer [11], [25], [26], correlation filter [12], [27], [28] or attention blocks [6], [29], [30]. Recently, one-stream trackers [18], [31] achieve joint feature extraction and interaction using Transformer [20], [32], which simplify the tracking pipeline and achieve superior performance.

NL. Different from the above tracking paradigm, some trackers [7], [33], [34] specify target object purely based on natural language, which provides a novel human-computer interaction manner. Li *et al.* [7] first define this task and provide a baseline by combining a grounding model and a tracking model. Then, some trackers [33], [34], [35] follow this paradigm to design different models to solve the grounding task and tracking task separately. Recently, some trackers [8], [36] perform tracking and grounding using a unified model, which simplifies the overall tracking framework and enables end-to-end optimization.

NL+BBOX. To provide more accurate target references, some trackers [7], [9] specify target object through both the initial bounding box and natural language. Li *et al.* [7] first introduce natural language into tracking achieving more robust results than visual trackers, which demonstrates the potential of vision-language tracking. SNLT [10] embeds natural language into Siamese-based trackers as a convolutional kernel and localizes target object through cross-correlation. VLT [9] treats the natural language feature as a selector to weigh different visual feature channels, enhancing target-related channels for robust tracking. Also, JointNLT [8] 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 with bounding box (BBOX) reference [9]. To this end, we design a multi-modal contrastive loss to align features of different reference modalities into a unified semantic space for consistent feature learning. Meanwhile, a reference-adaptive box head is proposed to alleviate the instability of target localization across reference modalities. Thanks to effective designs, UniSOT achieves promising performance across all reference modalities with high FPS.

2.2 Trackers based on Different Video Modalities

Different video modalities can provide complementary information for target localization. Based on the video modalities, mainstream trackers can be divided into two categories, including RGB and RGB+X trackers.

RGB. Most existing trackers [11], [19], [26] are designed for RGB video modality, which contains rich appearance cues to discriminate target object. These trackers extract features for both the target template and search region, and then match features between them through correlation filter [28], [37], [38], [39], convolution filter [12], [24], [40], cross-correlation [11], [23], [41] or attention mechanism [6], [29], [42] to localize target object. Recent trackers [18], [31], [43] allow interaction between target template and search region in the feature extraction phase for target-aware feature learning. Although great progress has been achieved, these RGB trackers still have inherent shortcomings in challenging scenarios, which have few appearance cues such as occlusion, fast motion, low illumination and so on [13].

RGB+X. Auxiliary video modalities (depth, thermal and event) are drawing increasing attention due to their complementary information with RGB images. Traditionally, RGB-D trackers [16], [44], RGB-T trackers [15], [45], RGB-E trackers [14], [46] design video modality-specific modules to integrate auxiliary video modalities, achieving more robust performance under complex scenarios. Recently, some works [2], [13] seek to adapt foundation RGB trackers for RGB-X tracking tasks via video modality-specific fine-tuning, which can make full use of the pretrained knowledge of the foundation model. Furthermore, UnTrack [17] proposes a video modality-joint fine-tuning method via low-rank factorization and reconstruction techniques. However, it projects different video modalities into different low-rank spaces with an identical rank. Such a design poses a challenge for video modality-aligned feature learning, and ignores the potential difference in the amount of information across video modalities, which may desire various ranks. To address the above limitations, we design a rank-adaptive modality adaptation (RAMA) mechanism for video modality-joint tuning, which can dynamically adjust the rank of shared parameters for different auxiliary video modalities, enabling both video modality-shared and video modality-specific feature learning.

3 Our Method

In this section, we first introduce the overall architecture of UniSOT, which presents a simple but effective tracking pipeline for different combinations of reference and video modalities. The following two subsections introduce reference and video modality designing respectively.

3.1 Tracking Architecture

As shown in Figure 2, UniSOT can perform tracking with different combinations of reference modalities (NL, BBOX, NL+BBOX) and video modalities (RGB, RGB+Depth, RGB+Thermal, RGB+Event). The template is cropped based on the initial bounding box. The search region is cropped based on the last-frame bounding box. Given the language description l, we tokenize the sentence and embed each word with the text embedding

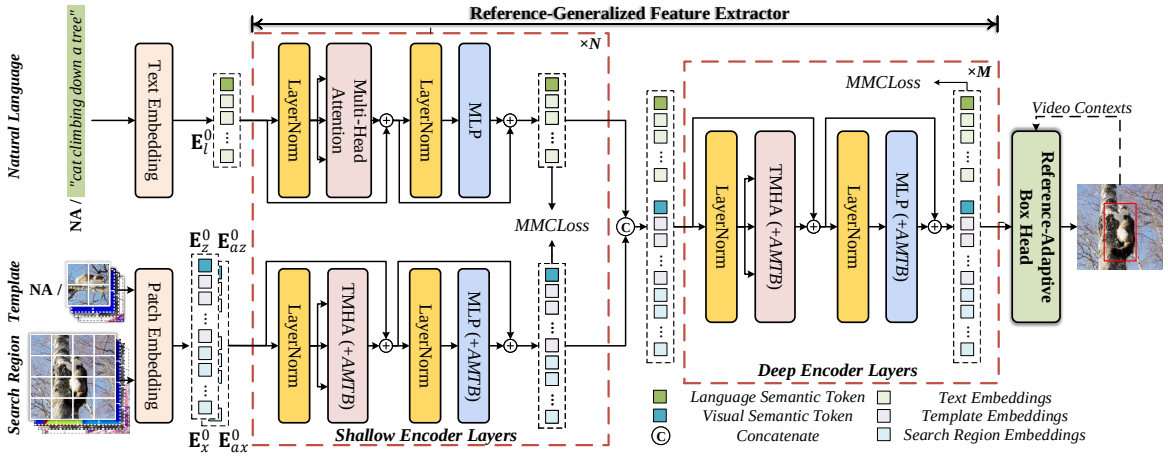


Fig. 2. A unified tracking framework for different target references. NA means "not available", which is filled with zeros. Natural language is not available for visual tracking task, and the template is not available for grounding task. Different from previous trackers designed for specific reference or video modalities, our UniSOT can simultaneously handle different combinations of reference and video modalities in a unified framework.

layer to obtain text embeddings $\mathbf{E}_l^0 \in \mathbb{R}^{N_l \times C}$. N_l is the maximum text length, C is the feature dimension. Given the RGB template $z \in \mathbb{R}^{3\times H_z \times W_z}$ with its auxiliary video modality $z_a \in \mathbb{R}^{3\times H_z \times W_z}$, the RGB search region (full image for grounding) $x \in \mathbb{R}^{3\times H_x \times W_x}$ with its auxiliary video modality $x_a \in \mathbb{R}^{3\times H_x \times W_x}$, we split and reshape them into a sequence of flattened 2D patches and then project them into a latent space. Learnable position embeddings are added to the corresponding patch embeddings, obtaining template embeddings $\mathbf{E}_z^0 \in \mathbb{R}^{N_z \times C}$ with its auxiliary modality embeddings $\mathbf{E}_{az}^0 \in \mathbb{R}^{N_z \times C}$, search region embeddings $\mathbf{E}_x^0 \in \mathbb{R}^{N_x \times C}$ with its auxiliary video modality embeddings $\mathbf{E}_{ax}^0 \in \mathbb{R}^{N_x \times C}$. Here, the position embeddings for the template and search region are independent to implicitly reflect their roles as different image types (template vs. search region), while those for different video modalities associated with the same image type are shared to align their position information. N_z and N_x are the patch numbers of the template and search region respectively. Here, the unavailable (**NA**) features for different reference and video modalities will be filled with zeros. We also prepend a language semantic token $\mathbf{T}_l^0 \in \mathbb{R}^{1\times C}$ and a visual semantic token $\mathbf{T}_v^0 \in \mathbb{R}^{1\times C}$ to text embeddings and image embeddings correspondingly, which are designed to capture the global semantics of different reference modalities. After that, text and image embeddings are fed into the reference-generalized feature extractor, which is built on Transformer architecture. Specifically, we extract language and visual features separately in shallow encoder layers and fuse them in deep encoder layers. A multi-modal contrastive loss (MMCLoss) is proposed to align vision and language features into a unified semantic space. Moreover, features of RGB modality and auxiliary video modality are interacted in auxiliary modality tuning blocks (AMTB). Finally, we feed the interacted text and image embeddings into the reference-adaptive box head, which can make full use of different reference modalities to mine ever-changing scenario features from video contexts and localize target object 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.

3.2 Reference Modality Designing

3.2.1 Reference-Generalized Feature Extractor

For unified tracking with different reference modalities, three critical issues must be carefully addressed in feature extraction: multi-modal feature learning and fusion, reference modality-compatible training, and multi-modal feature alignment. Thus, we design a new architecture for reference modality generalized feature extraction, termed reference-generalized feature extractor. As shown in Figure 2, the reference-generalized feature extractor is constructed based on Transformer [20], which consists of N shallow encoder layers and M deep encoder layers. Here, "shallow" and "deep" refer to the sequential position of the layers in the reference-generalized feature extractor. 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 reference modalities, and allow high-level semantics interaction for target localization.

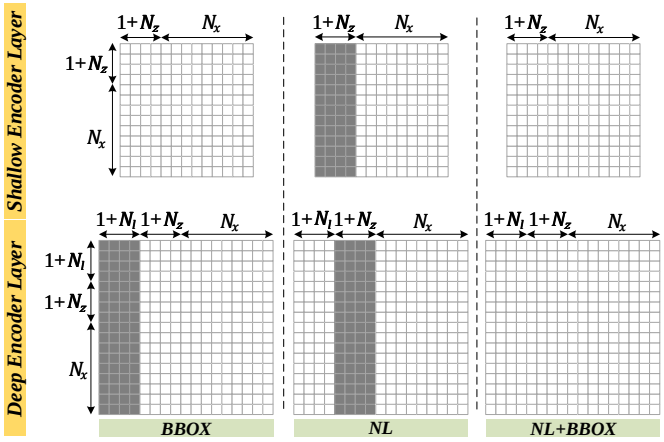
For modality-compatible training of different reference inputs, we fill the unavailable reference embeddings with zeros and propose a task-oriented multi-head attention mechanism (TMHA) to avoid task-irrelevant feature interactions, inspired by masking mechanism in Transformer [20]. For simplicity, we present the single-head formulas of TMHA below. 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,

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

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

where $\mathrm{MLP}(\cdot)$ is multi-layer perception, $\mathrm{LN}(\cdot)$ is layer normalization, $\widehat{\mathbf{E}}^i$ 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 reference modalities, where we fill the positions of unavailable features with $-\inf$ and other positions with 0.

Furthermore, during joint training with different reference modalities, we find that semantic tokens derived from different reference modalities exhibit significant discrepancies on the corresponding maps of the search region and demonstrate limited



(a) Attention Mask $\mathbf{M}_a$

Fig. 3. The attention mask of task-oriented multi-head attention for different target references. It enables different reference modalities to be trained jointly.

discriminability to the target object. This indicates that, within a unified framework, features from different reference modalities suffer from a substantial semantic gap and poor discriminative power, ultimately hindering their effectiveness in robust tracking. Thus, inspired by cross-modal alignment algorithms [47], [48], we specifically design a simple yet effective multi-modal contrastive loss (MMCLoss) to enhance the similarity between visual/language semantic tokens and target features in the search region while suppressing similarity with background distractors ($e.g.$, top N_{neg} scores out of target object box). By this design, we can align vision and language features into a unified semantic space and improve the discriminability of reference features. As shown in Figure 4, we first acquire the semantic token $\mathbf{T}^i$ of i^{th} encoder layer according to the reference modalities (NL, BBOX or NL+BBOX), which ensures compatibility with different reference modalities during joint training.

$$\mathbf{T}^i = \begin{cases} \mathbf{T}_l^i & NL, \\ \mathbf{T}_v^i & BBOX, \\ (\mathbf{T}_l^i + \mathbf{T}_v^i)/2 & NL + BBOX. \end{cases} \tag{3}$$

Given the semantic token $\mathbf{T}^i$ of i^{th} encoder layer, we compute the similarity $\mathbf{S}^i = [s^{i,1}; s^{i,2}, ..., s^{i,N_x}]$ between $\mathbf{T}^i$ and search region embeddings $\mathbf{E}_x^i = [f^{i,1}; f^{i,2}, ..., f^{i,N_x}]$. Formally,

$$s^{i,j} = \mathrm{sim}(\mathbf{T}^i, f^{i,j})/\tau, \mathrm{sim}(\mathbf{T}^i, f^{i,j}) = \frac{\mathbf{T}^i (f^{i,j})^\top}{||\mathbf{T}^i||_2 ||f^{i,j}||_2}, \tag{4}$$

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

$$\mathcal{L}_{mmc}^i = -\log\left(\frac{e^{s_p^i}}{e^{s_p^i} + \sum_{k=1}^{N_{neg}} e^{s_n^{i,k}}}\right). \tag{5}$$

Discussion. Unlike prior image-wise or video-wise contrastive learning losses [47], [48], [49], which focus on text-image or text-video holistic alignment, MMCLoss is a target-wise contrastive loss specifically tailored for unified tracking with diverse reference modalities. It focuses on the alignment between different reference modalities and target features, thereby aligning visual and language features into a unified semantic space and enhancing the discriminability of reference features for robust tracking.

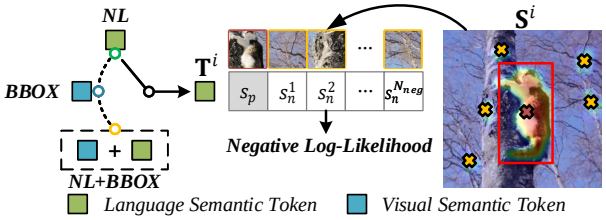


Fig. 4. The diagram of the multi-modal contrastive loss. We align different reference modalities with target object patch feature by contrasting hard background patches.

3.2.2 Reference-Adaptive Box Head

Inspired by OSTrack [18], 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, enhanced features of search region vary with reference modalities, while the box head processes all search regions identically despite variations in reference modality, which may lead to unstable results across reference modalities. Thus, we further design a reference-adaptive box head (RABH), which can adaptively mine ever-changing scenario features from video context by reference features of different modalities to assist target localization for robust tracking.

As shown in Figure 5(a), we treat the semantic token in the last encoder layer as the target prototype $\widehat{\mathbf{P}}_t = \mathbf{T}^{N+M}$ (as defined in Equation 3) and introduce learnable distractor prototype $\widehat{\mathbf{P}}_d$ and background prototype $\widehat{\mathbf{P}}_b$ to localize target object in a contrastive way. Here, $\widehat{\mathbf{P}}_d \in \mathbb{R}^{1 \times C}$ and $\widehat{\mathbf{P}}_b \in \mathbb{R}^{1 \times C}$ are shared across reference modalities, which provide a general capability to discriminate backgrounds when video contexts are unavailable, $e.g.$, localizing the target object in the first frame for natural language reference (NL). During tracking, video contexts can provide rich scenario cues to discriminate target object. 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. Historic search region embeddings with high-confidence target bounding boxes are saved as context embeddings $\mathbf{E}_c^{N+M}$ (as stated in Section 3.1). Given the template and context embeddings $\mathbf{E}_t = [\mathbf{E}_z^{N+M}; \mathbf{E}_c^{N+M}] \in \mathbb{R}^{(N_z+N_x) \times C}$ 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,

$$\mathbf{A}_{in} = \mathrm{Softmax}\left(\frac{\mathbf{T}^{N+M}\mathbf{E}_t^\top}{\sqrt{C}} + \mathbf{M}_t\right), \tag{6}$$

$$\mathbf{A}_{out} = \mathrm{Softmax}\left(\frac{\mathbf{T}^{N+M}\mathbf{E}_t^\top}{\sqrt{C}} + \widetilde{\mathbf{M}}_t\right), \tag{7}$$

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 $-inf$. $\widetilde{\mathbf{M}}_t$ 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 target object is a key factor affecting the tracking robustness [50], we divide features out of the target box into distractor and background from the perspective of

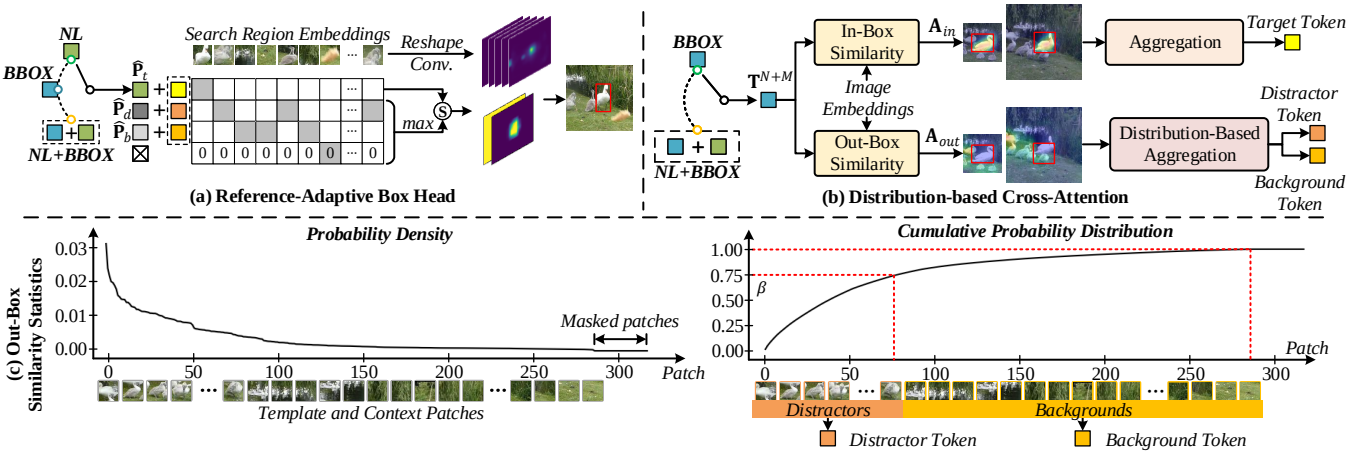


Fig. 5. (a) shows the schematic of the reference-adaptive box head, which can make full use of reference information to discriminate target object. (b) shows the structure of the distribution-based cross-attention. (c) illustrates the descending order of similarity between the semantic token and template/context background patches, which can be regarded as the probability of background features being mistakenly classified as the target object. The cumulative probability distribution clarifies how we apply the threshold β to separate distractor features from background features.

distribution. As shown in Figure 5(c), we rank the out-box probabilities $\mathbf{A}_{out}$ in descending order and sum them cumulatively to obtain the cumulative probability distribution. The probability density can be regarded as the probability of background features being mistakenly classified as the target object. We can find that some background patches have a similar appearance to the target object, which tends to confuse the tracker. Thus, we separate these distractor features from background features using a threshold β on the cumulative probability distribution to obtain distractor patches in the search region. Specifically, we assign the patch whose target cumulative probability distribution score is lower than β to 0 for $\mathbf{M}_d$ and other positions to $-inf$. $\widetilde{\mathbf{M}}_d$ is the opposite. Then, the distractor token and background token can be formulated as,

$$\mathbf{T}_d = \text{Softmax}(\frac{\mathbf{T}^{N+M}\mathbf{E}_t^\top}{\sqrt{C}} + \widetilde{\mathbf{M}}_t + \mathbf{M}_d)\mathbf{E}_t, \tag{8}$$

$$\mathbf{T}_b = \text{Softmax}(\frac{\mathbf{T}^{N+M}\mathbf{E}_t^\top}{\sqrt{C}} + \widetilde{\mathbf{M}}_t + \widetilde{\mathbf{M}}_d)\mathbf{E}_t. \tag{9}$$

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

$$\mathbf{P}_t = \widehat{\mathbf{P}}_t + \mathbf{T}_t, \mathbf{P}_d = \widehat{\mathbf{P}}_d + \mathbf{T}_d, \mathbf{P}_b = \widehat{\mathbf{P}}_b + \mathbf{T}_b, \tag{10}$$

$$\hat{\alpha}_t^i = \text{sim}(f^i, \mathbf{P}_t)/\tau, \tag{11}$$

$$\hat{\alpha}_b^i = \max\big(\text{sim}(f^i, \mathbf{P}_d)/\tau, \text{sim}(f^i, \mathbf{P}_b)/\tau, 0\big), \tag{12}$$

$$\alpha_t^i = e^{\hat{\alpha}_t^i}/(e^{\hat{\alpha}_t^i} + e^{\hat{\alpha}_b^i}). \tag{13}$$

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,

$$(\hat{x}, \hat{y}) = \Big((x_c + \hat{\mathbf{O}}(0, x_c, y_c)) \cdot p, (y_c + \hat{\mathbf{O}}(1, x_c, y_c)) \cdot p\Big), \tag{14}$$

$$(\hat{w}, \hat{h}) = \big(\hat{\mathbf{S}}(0, x_c, y_c) \cdot H_x, \hat{\mathbf{S}}(1, x_c, y_c) \cdot W_x\big). \tag{15}$$

3.2.3 Training Objective

We train UniSOT on RGB datasets with different reference modalities, which have rich data sources, enabling powerful foundational tracking capabilities. We call this the first training stage. Specifically, 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,

$$\mathcal{L}_{tgt} = \mathcal{L}_{bce}(\hat{\mathbf{L}}, \mathbf{L}) \tag{16}$$

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 OSTrack [18]. In summary, the overall objective function can be formulated as,

$$\mathcal{L} = \mathcal{L}_{tgt} + \mathcal{L}_{cls} + \mathcal{L}_{box} + \lambda_{mmc}\sum_{i=1}^{N+M}\mathcal{L}_{mmc}^i. \tag{17}$$

3.3 Video Modality Designing

In the first training stage, UniSOT enables the unification of reference modalities on RGB sequences. In this subsection, we design a rank-adaptive modality adaptation mechanism for video modality joint tuning, enabling UniSOT to freely integrate auxiliary video modalities for robust tracking.

3.3.1 Rank-Adaptive Modality Adaptation

Previous methods [13], [17] project auxiliary video modalities into different low-rank spaces with an identical rank and fuse with RGB features. Such a design poses a challenge for video modality-aligned feature learning and ignores the potential difference in the amount of information across video modalities, which may desire various ranks. To address the above limitation, we propose a novel solution for video modality joint tuning inspired by AdaLoRA [21], enabling both video modality-aligned and video modality-specific feature learning.

Formulation. To empower the capability of auxiliary video modalities, we indirectly train some dense layers in UniSOT by parameterizing their adaptation changes as incremental weights, while freezing the pretrained weights. Specifically, we design an auxiliary modality tuning block (AMTB) for video modality-joint tuning, as shown in Figure 6, which can fuse auxiliary video modalities ($e.g.$ event feature $\mathbf{E}_E$) to RGB feature $\mathbf{E}_R$ by

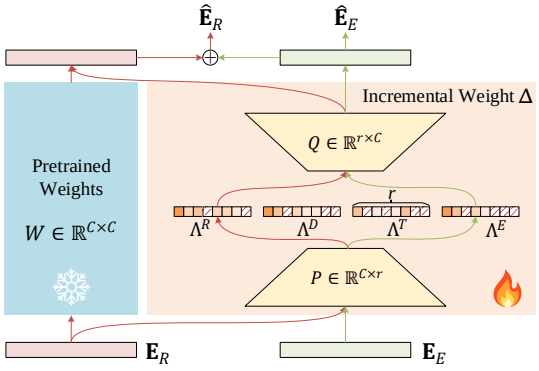


Fig. 6. The diagram of auxiliary modality tuning block (AMTB). Different video modalities share a uniform parameter set P and Q for video modality-aligned feature learning. Meanwhile, we assign various ranks across video modalities via singular values Λ to learn modality-specific cues.

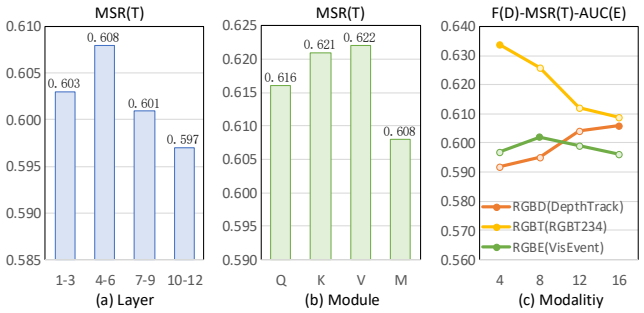


Fig. 7. Tuning performance comparison across tuned layers, modules and modalities. We inject AMTB into UniSOT and fix elements of Λs to 1. (a) and (b) show the tuning performance on RGBT234 dataset, which reflects the contribution gaps among incremental weights. (c) shows the performance for different video modalities and ranks, which indicates auxiliary video modalities desire various ranks to learn effective features. We adopt a common training configuration [13] for these experiments with consistent epochs and sample numbers.

incremental weights Δ. AMTB can be injected into the feature extractor to integrate auxiliary video modalities, as shown in Figure 8. Inspired by AdaLoRA [21], we parameterize the incremental weights Δ in the form of singular value decomposition $\Delta = P\Lambda Q$, where $P \in \mathbb{R}^{C\times r}$ and $Q \in \mathbb{R}^{r\times C}$ are orthogonal matrices, $\Lambda \in \mathbb{R}^{r\times r}$ contains singular values and r denotes the inherent rank of Δ. In practice, since Λ is diagonal, Λ can be expressed as a vector in $\mathbb{R}^r$. This weight format enables us to explicitly manipulate the rank of Δ by the number of non-zero elements in Λ. Further, we assign different Λ for different video modalities, termed Λ^R, Λ^D, Λ^T, Λ^E. Here, R, D, T and E correspond to RGB, depth, thermal and event modalities, similarly hereinafter. Different video modalities share a uniform parameter set P and Q, which facilitates video modality-aligned feature learning. Meanwhile, various Λ allow networks to learn features with different ranks for different video modalities to obtain video modality-specific cues. The output of the auxiliary modality tuning block can be formulated as,

$$\hat{\mathbf{E}}_a = ReLU(\mathbf{E}_a P\Lambda^a Q), \tag{18}$$

$$\hat{\mathbf{E}}_R = \mathbf{E}_R W + \mathbf{E}_R P\Lambda^R Q + \hat{\mathbf{E}}_a, \tag{19}$$

where $a \in \{D, T, E\}$ is the auxiliary video modality.

Parameter Importance. As shown in Figure 7, the performance of the model after fine-tuning varies with tuned layers, modules and modalities, which means the contribution gaps among incremental weights. Also, as shown in Figure 7(c), auxiliary video modalities desire various ranks to learn effective features, which poses a challenge for video modality-joint tuning. To this end, we seek to evaluate the contribution of parameters and allocate different ranks across layers, modules and modalities. So, we first define an importance function $s(\cdot)$ for incremental parameters:

$$I(w_{ij}) = |w_{ij}\nabla_{w_{ij}}\mathcal{L}|, \tag{20}$$

$$\overline{I}^{(t)} = \beta_1\overline{I}^{(t-1)} + (1-\beta_1)I^{(t)}, \tag{21}$$

$$\overline{U}^{(t)} = \beta_2\overline{U}^{(t-1)} + (1-\beta_2)|\overline{I}^{(t)} - I^{(t)}|, \tag{22}$$

$$s(w_{ij}) = \overline{I}^{(t)} \cdot \overline{U}^{(t)}. \tag{23}$$

Equation (20) essentially measures the change of loss when w_{ij} is zeroed out [51]. Equations (21–23) further stabilizes $I(w_{ij})$ through cross-batch sensitivity smoothing and uncertainty quantification [52], t means iteration step.

Modality-shared Rank Allocation. For the sake of description, we further denote incremental weights Δ_k as parameter tuples $G_{k,i} = \{P_{k,*i}, \lambda^R_{k,i}, \lambda^D_{k,i}, \lambda^T_{k,i}, \lambda^E_{k,i}, Q_{k,i*}\}$ containing the i^{th} singular values and vectors in k^{th} incremental weights. Here, $P_{k,*i} \in \mathbb{R}^{C\times 1}, Q_{k,i*} \in \mathbb{R}^{1\times C}$. We first design a modality-shared importance $S_{k,i}$ for each parameter tuple, which is based on the importance of singular values and vectors. Formally,

$$\begin{aligned} S_{k,i} = &\sum_{a\in\{R,D,T,E\}} s(\lambda^a_{k,i}) \\ &+ \frac{1}{d_1}\sum_{j=1}^{d_1} s(P_{k,ji}) + \frac{1}{d_2}\sum_{j=1}^{d_2} s(Q_{k,ij}). \end{aligned} \tag{24}$$

Then, we reserve top-n important singular values and zero out others according to $S_{k,i}$, adaptively constructing a compact modality-shared parameter space:

$$\hat{\lambda}^a_{k,i} = \begin{cases} \lambda^a_{k,i} & S_{k,i} \text{ is in the top-}n \text{ of } \mathcal{S}^1, \\ 0 & \text{otherwise}, \end{cases} \tag{25}$$

where $a \in \{R, D, T, E\}$ similarly hereinafter, n denotes the modality-shared rank budget and $\mathcal{S}^1$ contains the importance scores of all parameter tuples $G_{k,i}$.

Modality-specific Rank Allocation. As shown in Figure 7(c). Different video modalities desire various ranks to learn modality-specific cues and prevent overfitting. We further split Δ_k as modality-specific parameter tuples $G^a_{k,i} = \{P_{k,*i}, \lambda^a_{k,i}, Q_{k,i*}\}$ and evaluate the modality-specific importance $S^a_{k,i}$ for each parameter tuple to adjust the rank of weight for different video modalities. It can be formulated as,

$$S^a_{k,i} = s(\hat{\lambda}^a_{k,i}) + \frac{1}{d_1}\sum_{j=1}^{d_1} s(P_{k,ji}) + \frac{1}{d_2}\sum_{j=1}^{d_2} s(Q_{k,ij}). \tag{26}$$

Similarly, we reserve top-m important singular values and zero out others:

$$\hat{\lambda}^a_{k,i} = \begin{cases} \hat{\lambda}^a_{k,i} & S^a_{k,i} \text{ is in the top-}m \text{ of } \mathcal{S}^2, \\ 0 & \text{otherwise}, \end{cases} \tag{27}$$

where m denotes the modality-specific rank budget and $\mathcal{S}^2$ contains the importance scores of all modality-specific parameter tuples $G^a_{k,i}$.

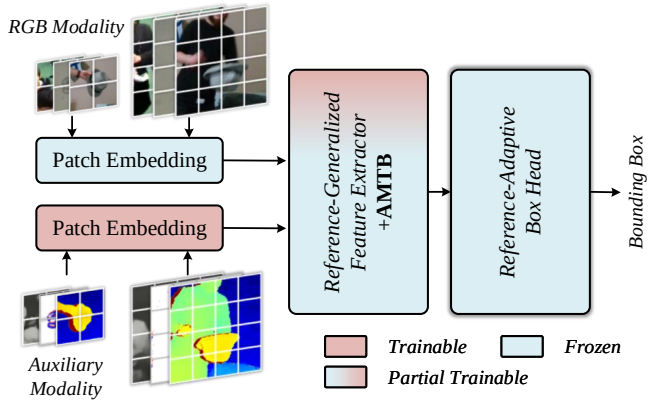


Fig. 8. The diagram of UniSOT. UniSOT integrates auxiliary modality tuning blocks (AMTB) for modality-joint tuning, which allows UniSOT to unify RGB+Depth, RGB+Thermal, and RGB+Event tracking with a uniform parameter set.

3.3.2 Fine-tuning

We further fine-tune UniSOT on RGB+X datasets, enabling various video modalities. We call this the second training stage. As shown in Figure 8, auxiliary video modalities are fed to UniSOT together with RGB images (template and search region). We freeze the parameters trained in the first training stage. Specifically, we add a trainable patch embedding layer for feature extraction of auxiliary video modality. Then, auxiliary modality tuning blocks are injected into the feature extractor for video modality-joint tuning. To enforce the orthogonality of P and Q, we add an orthogonal loss to the tuning objective $\mathcal{L}_{total}$ to enforce approximately orthogonal throughout training. $\mathcal{L}$ is consistent with the first stage. Formally,

$$\mathcal{L}_{orth} = ||P^\top P - I||_F^2 + ||Q^\top Q - I||_F^2, \tag{28}$$

$$\mathcal{L}_{total} = \mathcal{L} + \lambda_{orth}\mathcal{L}_{orth}. \tag{29}$$

Discussion. Modality-shared rank allocation constructs a compact modality-shared parameter space conducive to modality-aligned feature learning. Thanks to the feature alignment between RGB and auxiliary video modalities, these auxiliary video modalities can be integrated into vision-language tracking without any modification. It enables UniSOT to simultaneously handle different combinations of reference and video modalities in a unified framework with a uniform parameter set. Modality-specific rank allocation adaptively assigns different ranks across video modalities, allowing video modality-specific feature learning while avoiding overfitting. Unlike AdaLoRA, RAMA introduces a cross-modal rank allocation strategy to address the learning problem of video modality-shared and video modality-specific features during video modality-joint tuning. Meanwhile, RAMA is the first work to explore rank-adaptive allocation strategies in multi-modal joint fine-tuning, it empowers RGB-based models with multiple auxiliary video modalities through a single fine-tuning process, significantly streamlining training procedures. The resulting model enable multi-modal video tracking via a uniform parameter set, providing a reliable technical foundation for endowing RGB foundation models with multi-modal capabilities.

4 Experiment

Our tracker is implemented using Python 3.8.13 and Pytorch 1.10.1. The UniSOT are trained on a server with eight 24GB NVIDIA RTX 3090 GPUs and AMD EPYC 7713 64-Core Processor @ 2GHz with 503 GB RAM. The inference speed is tested on a single NVIDIA RTX2080 Ti GPU and Intel(R) Xeon(R) CPU E5-2695 v4 @ 2.10GHz with 251 GB RAM.

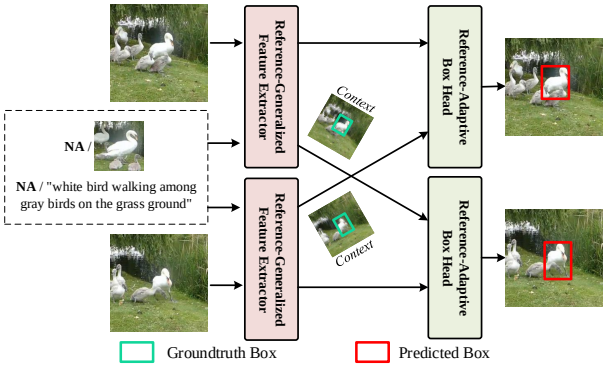


Fig. 9. The diagram of the UniSOT training process. We sample two search regions from a video, whose features serve as context features for each other in the box head.

4.1 Implementation Details

4.1.1 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, we present two variants, termed UniSOT-B and UniSOT-L. The number of encoder layers is set to N=6,M=6 for UniSOT-B and $N = 12$,$M = 12$ for UniSOT-L.

4.1.2 The First Training Stage.

Data Unit. As shown in Figure 9, we sample one template, two search regions and the corresponding language description as a training unit in the first training stage. Here, the template is not available for the natural language reference (NL). Natural language is not available for the bounding box reference (BBOX). The unavailable reference embeddings are filled with zeros in the reference-generalized feature extractor. The training units for bounding box reference (BBOX) are sampled from the training splits of LaSOT [53], GOT-10k [54], COCO2017 [55], TrackingNet [56], TNL2K [34], OTB99 [7]. The training units for natural language reference (NL) are sampled from the training splits of LaSOT, TNL2K, OTB99 and RefCOCOg-google [57]. The training units for both bounding box and natural language reference (NL+BBOX) are sampled from the training splits of LaSOT, TNL2K, OTB and RefCOCOg-google. The maximum sampling interval of search regions and target template is set to 200. Common data augmentation is used for model training, such as translation, horizontal flip, and color jittering [6]. Due to the flexibility of our framework, different reference modalities can be trained jointly, which provides a neat training pipeline.

Initialization. The language branch in shallow encoder layers of UniSOT-B is initialized with uncased version parameters of BERT-Base [58]. Other parameters in the reference-generalized feature extractor are initialized with ViT-Base parameters pretrained by MAE [59]. Correspondingly, UniSOT-L parameters are initialized with BERT-Large and ViT-Large parameters. The reference-adaptive box head is initialized with Xavier init [60].

TABLE 1
Comparison with state-of-the-art real-time visual trackers on TNL2K, AVisT, LaSOT, LaSOT$_{ext}$ and TrackingNet. The best two results are shown in bold and underline.

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											
UniSOT-L	**62.8**	**64.6**	**57.8**	**67.9**	**48.7**	**71.3**	**78.3**	**51.2**	**59.0**	**84.1**	82.9
APMT-SwinB [19]	-	-	-	-	-	70.2	77.0	-	-	83.6	82.2
OSTrack-384 [18]	55.9	-	-	-	-	71.1	77.6	50.5	57.6	83.9	**83.2**
OmniTracker-L [61]	-	-	-	-	-	69.1	75.4	-	-	83.4	82.3
Unicorn-ConvL [61]	-	-	-	-	-	68.5	74.1	-	-	83.0	82.2
MixFormer-L [31]	-	-	56.0	65.9	46.3	70.1	76.3	-	-	83.9	83.1
SimTrack-L/14 [43]	55.6	55.7	-	-	-	70.5	-	-	-	83.4	-
STARK-ST101 [6]	-	-	50.5	58.2	39.0	67.1	-	-	-	82.0	86.9
Basic Variants											
UniSOT-B	**60.9**	**61.9**	**56.5**	**66.0**	**45.1**	69.4	74.9	**49.2**	**55.8**	83.4	**82.1**
DiffusionTrack [62]	56.5	57.3	-	-	-	**70.7**	**77.3**	-	-	**83.6**	82.0
BIT [63]	-	-	-	-	-	68.6	73.2	48.8	56.8	83.0	81.0
MCFT [64]	-	-	-	-	-	68.5	74.5	-	-	82.4	81.7
APMT-SwinT [19]	-	-	-	-	-	68.4	74.2	-	-	81.7	79.8
SFTransT [65]	54.6	55.5	-	-	-	69.0	73.9	46.4	54.1	82.9	81.3
OSTrack-256 [18]	54.3	-	-	-	-	69.1	75.2	47.4	53.3	83.1	82.0
MixFormer-22k [31]	-	-	53.7	63.0	43.0	69.2	74.7	-	-	83.1	81.6
SimTrack-B/16 [43]	54.8	53.8	-	-	-	69.3	-	-	-	82.3	-
AiATrack [30]	-	-	-	-	-	69.0	73.8	47.7	55.4	82.7	80.4
STARK [6]	-	-	51.1	59.2	39.1	66.4	71.2	-	-	81.3	78.1
TransT [29]	50.7	51.7	49.0	56.4	37.2	64.9	73.8	-	-	81.4	80.3
TrDiMP [66]	-	-	48.1	55.3	33.8	63.9	61.4	-	-	78.4	73.1
Ocean [67]	38.4	37.7	38.9	43.6	20.5	56.0	56.6	-	-	-	-
KYS [68]	44.9	43.5	-	-	-	55.4	-	-	-	74.0	68.8
SiamFC++ [69]	38.6	36.9	-	-	-	54.4	54.7	-	-	75.4	70.5
SiamBAN [25]	41.0	41.7	-	-	-	51.4	52.1	-	-	-	-
DiMP [12]	44.7	43.4	41.9	45.7	26.0	56.9	56.7	39.2	45.1	74.0	68.7
SiamRPN++ [23]	41.3	41.2	39.0	43.5	21.2	49.6	49.1	34.0	39.6	73.3	69.4
ECO [28]	32.6	31.7	-	-	-	32.4	30.1	22.0	24.0	55.4	49.2
MDNet [70]	-	-	-	-	-	39.7	37.3	27.9	31.8	60.6	56.5
SiamFC [11]	29.5	28.6	-	-	-	33.6	33.9	23.0	26.9	57.1	53.3

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.

	Ocean	DiMP-50	PrDiMP-50	TransT	OSTrack-256	OSTrack-384	UniSOT-B	UniSOT-L
NFS	49.4	61.8	63.5	65.3	64.7	66.5	65.9	**67.6**
UAV123	57.4	64.3	68.0	68.1	68.3	70.7	69.3	**71.0**

TABLE 3
Comparison with state-of-the-art trackers on VOT2022.

Method	EAO ↑	Acc ↑	Rob ↑
Base Backbones			
UniSOT-B	**0.558**	**0.791**	**0.852**
SeqTrack-B256 [71]	0.513	0.789	0.806
OSTrack-256 [18]	0.516	0.777	0.801
MixFormer-22k [31]	0.531	0.775	0.842
SwinTrack-B [72]	0.524	0.788	0.803
TransT [29]	0.512	0.781	0.800
STARK-ST50 [6]	0.467	0.778	0.756
DiMP [12]	0.430	0.689	0.760
SiamFC [11]	0.255	0.562	0.543
Larger Backbone			
UniSOT-L	**0.561**	0.777	**0.855**
SeqTrack-L256 [71]	0.542	0.770	0.852
MixFormer-L [31]	0.541	**0.779**	0.835
STARK-ST101 [6]	0.492	0.775	0.782
VOT2022 Challenge (unpublished or variants)			
DAMT [73]	0.602	0.776	0.887
APMT-MR [73]	0.591	0.787	0.877
OSTrackSTB [73]	0.591	0.790	0.869
ADOTstb [73]	0.569	0.775	0.862
SRATransT [73]	0.560	0.764	0.864

TABLE 4
Evaluation of UniSOT on VOTS2024.

Method	Q ↑	Acc ↑	Rob ↑
UniSOT-L (+AlphaRefine)	**0.500**	0.705	**0.741**
UniSOT-B (+AlphaRefine)	0.481	0.699	0.716
MixFormer-L (+AlphaRefine) [74]	0.499	**0.713**	0.736
STARK (+AlphaRefine) [74]	0.462	0.695	0.685
MixFormerV2 (+AlphaRefine) [74]	0.458	0.678	0.693
ToMP (+AlphaRefine) [74]	0.442	0.681	0.671
Segmentation-based trackers			
DMAOT [75]	0.636	0.751	0.795
Cutie+SAM [76]	0.607	0.756	0.730
AOT [77]	0.550	0.698	0.767

Training Settings. As shown in Figure 9, two search regions interact with target references separately in the reference-generalized feature extractor. Then, these embeddings of two search regions serve as video contexts for each other to provide scenario cues in the reference-adaptive box head. Finally, the outputs of both search regions are constrained by the designed losses. The loss weights are set to $\lambda_{giou} = 2.0$, $\lambda_1 = 5.0$, $\lambda_{mmc} = 0.1$. In addition, our model is trained with a batch size of 64 and iterates 300 epochs with 3×10^4 samples per epoch. AdamW optimizer [78]

TABLE 5
Comparison with state-of-the-art vision-language trackers on LaSOT, TNL2K and OTB99 datasets.

Method	TNL2K		LaSOT		OTB99	
	AUC	P	AUC	P	AUC	P
NL						
UniSOT-L	**56.4**	**57.5**	**59.6**	**63.9**	**63.5**	**83.2**
UniSOT-B	53.9	54.4	57.2	61.0	60.1	79.1
QueryNLT	53.3	53.0	54.2	55.0	61.2	81.0
JointNLT	52.4	52.2	56.9	59.3	59.2	77.6
CTRNLT	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
RVTNLN	-	-	-	-	54.0	56.0
RTTNLD	-	-	-	-	54.0	78.0
TNLS-II	-	-	-	-	25.0	29.0
NL+BBOX						
UniSOT-L	**63.0**	**65.0**	**71.4**	**78.7**	71.1	92.0
UniSOT-B	61.2	62.8	69.4	75.9	69.3	89.9
QueryNLT	57.8	58.7	59.9	63.5	66.7	88.2
JointNLT	54.6	54.8	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
RVTNLN	25.0	27.0	50.0	56.0	67.0	73.0
RTTNLD	25.0	27.0	35.0	35.0	61.0	79.0
TNLS-III	-	-	-	-	55.0	72.0

with weight decay 10^{-4} is applied for model optimization. The learning rates start from 4×10^{-5} for the reference-generalized feature extractor and 4×10^{-4} for the reference-adaptive box head, then gradually reduce to 0 with cosine annealing [79].

4.1.3 The Second Training Stage.

Data Unit. We sample one template and two search regions as the first training stage does. Further, we also sample different auxiliary video modalities corresponding to the template and search regions for video modality-joint tuning. The inputs of auxiliary video modalities are converted to three channel formats as ViPT [13] does. The RGB+Depth, RGB+Thermal, RGB+Event data are sampled from DepthTrack [16], LasHeR [80] and VisEvent [46] respectively, which is same as UnTrack [17].

Fine-tuning Settings. We inject auxiliary modality tuning blocks (AMTB) into key, query, value projection of attention mechanisms and multi-layer perceptron. The inherent rank of incremental weights r is set as 32. The block-average modality-shared rank budget $\hat{n}$ is set as 16. The block-average modality-specific rank budget $\hat{m}$ is set as 32. Thus, given the number of AMTB N_{AMTB}, modality-shared rank budget $n = \hat{n} \times N_{AMTB}$ and modality-specific rank budget $m = \hat{m} \times N_{AMTB}$. We freeze the parameters of UniSOT trained in the first stage, and fine-tune incremental weights in AMTB with a batch size of 32 and iterate 60 epochs with 3×10^4 samples per epoch. λ_{orth} is set as 0.1. AdamW optimizer [78] with weight decay 10^{-4} is applied for model optimization. The learning rates start from 8×10^{-4} and then gradually reduce to 0 with cosine annealing.

4.1.4 Inference Details

Efficient Inference. We can remove unavailable references from inputs and replace all task-oriented multi-head attention with vanilla multi-head attention for efficient inference. In AMTB, we remove parameters in P and Q whose corresponding singular value Λ is zero to speed up inference for RGB+X inputs.

Scenario Token Update. Search region embeddings with high-confidence bounding boxes are saved as video contexts $\mathbf{E}_c^{N+M}$ to help with subsequent target localization. Specifically, if the target confidence score $\max[\hat{\mathbf{C}}(x,y)\hat{\mathbf{L}}(x,y)]$ is higher than 0.5, the corresponding search region embeddings will be saved as video contexts. The target token, distractor token and background token in the reference-adaptive box head are updated every 20 frames.

4.2 State-of-the-art Comparisons

Visual Tracking. We evaluate our trackers on seven visual tracking benchmarks, including TNL2K [34], AVisT [81], LaSOT [53], LaSOT$_{ext}$ [82], TrackingNet [56], NFS [83] and UAV123 [84], which are commonly used for visual tracker evaluation. The Area Under the Curve (AUC) of the success plot is the main metric for ranking trackers. As shown in Table 1-4, our UniSOT-B and UniSOT-L outperform trackers specifically designed for the visual tracking task on most benchmarks. Notably, UniSOT supports different reference modalities. These results demonstrate the effectiveness of UniSOT using bounding box reference (**BBOX**). 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 reference modalities and tend to rely on semantic information in language. Differently, UniSOT aligns vision and language into a unified semantic space, providing a unified tracking framework, which delivers superior performance for all three reference modalities.

Vision-Language Tracking. We further evaluate our UniSOT on vision-language tracking benchmarks, including TNL2K [34], LaSOT [53] and OTB99 [7] and compare with the latest trackers, including QueryNLT [36], JointNLT [8], CTRNLT [33], TNL2K [34], GTI [35], TNLS [7], VLTTT [9], RVTNLN [85], RTTNLD [86], SNLT [10]. As shown in Table 5, when specifying target object by natural language (**NL**), UniSOT-L surpasses the previous best tracker JointNLT with a large margin on three benchmarks. When initializing the tracker with both natural language and the bounding box (**NL+BBOX**), our UniSOT-L achieves the best performance on TNL2K and LaSOT. These results demonstrate the superiority of UniSOT for vision-language tracking.

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

Visual Grounding. Like VLTVG, we retrain UniSOT-B on train sets of RefCOCO [87], RefCOCO+ [87] and RefCOCOg [57] separately for a fair comparison, and report the Top-1 accuracy on corresponding test sets in Table 6. We compared UniSOT† with the grounding methods, including VLTVG [88], SeqTR [89], QRNet [90], TransVG [91], Ref-NMS [92], LBYL-Net [93], ReSC-Large [94], NMTree [95]. Here, UniSOT† means UniSOT is retrained with the general visual grounding setting [88]. The test image is scaled such that its long edge is 384. Although UniSOT is not specifically designed for grounding, UniSOT achieves the best performance on seven test sets, demonstrating the generalization of our framework.

Discussion. As mentioned above, existing trackers are designed for single or partial reference modalities and overspecialize on the specific reference modality. UniSOT not only can cope with three types of target reference (NL, BBOX, NL+BBOX) but also

TABLE 6
Comparison with state-of-the-art grounding methods on RefCOCO, RefCOCO+, RefCOCOg datasets. † means UniSOT is trained with the general visual grounding setting

Method	RefCOCO			RefCOCO+			RefCOCOg		
	val	*testA*	*testB*	*val*	*testA*	*testB*	*val-g*	*val-u*	*test-u*
UniSOT†	**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 7
Comparison with state-of-the-art RGBD trackers. * means only train with the first stage, similarly hereinafter.

Method	VOT-RGBD22			DepthTrack		
	EAO	Acc.	Rob.	F-score	Re	Pr
Unified Model with Uniform Parameters						
UniSOT-L	**73.8**	**82.1**	**89.1**	**63.2**	**62.5**	**63.9**
UniSOT-B	73.2	81.8	88.8	62.5	62.2	62.9
UniSOT-L*	68.9	81.6	84.3	59.4	60.6	58.3
UniSOT-B*	66.6	80.2	81.3	53.8	54.9	52.7
Un-Track [17]	71.8	82.0	86.4	61.0	61.0	61.0
Unified Model with Separate Parameters						
ViPT [13]	72.1	81.5	87.1	59.4	59.6	59.2
ProTrack [2]	65.1	80.1	80.2	57.8	57.3	58.3
Specialized Tracker for RGBD Tracking or VOT Challenge						
MixForRGBD [73]	77.9	81.6	94.6	-	-	-
SAMF [73]	76.2	80.7	93.6	-	-	-
ProMix [73]	72.2	79.8	90.0	-	-	-
SBT-RGBD [73]	70.8	80.9	86.4	-	-	-
SPT [44]	65.1	79.8	85.1	53.8	54.9	52.7
DMTrack [73]	65.8	75.8	85.1	-	-	-
DeT [16]	65.7	76.0	84.5	53.2	50.6	56.0
DDiMP [96]	-	-	-	48.5	56.9	50.3
STARK-RGBD [6]	64.7	80.3	79.8	-	-	-
DRefine [97]	59.2	77.5	76.0	-	-	-
ATCAIS [96]	55.9	76.1	73.9	47.6	45.5	50.0
Non-real Time Trackers						
SeqTrackv2-L384 [71]	74.8	82.6	91.0	62.3	62.6	62.5
SeqTrackv2-L256 [71]	74.9	81.3	91.8	62.8	63.0	62.5

TABLE 8
Comparison with state-of-the-art RGBT trackers.

Method	LasHeR		RGBT234	
	AUC	P	MSR	MPR
Unified Model with Uniform Parameters				
UniSOT-L	**54.8**	**68.4**	**65.0**	**87.4**
UniSOT-B	54.0	67.4	64.1	86.6
UniSOT-L*	43.4	55.9	63.3	84.2
UniSOT-B*	42.9	55.5	61.6	82.2
Un-Track [17]	51.3	64.6	62.5	84.2
Unified Model with Separate Parameters				
ViPT [13]	52.5	65.1	61.7	83.5
ProTrack [2]	42.0	53.8	59.9	79.5
Specialized Tracker for RGBT Tracking				
M3PT [98]	54.2	67.3	63.4	85.9
CMD [99]	46.6	59.0	58.4	82.4
APFNet [45]	36.2	50.0	57.9	82.7
CMPP [100]	-	-	57.5	82.3
JMMAC [101]	-	-	57.3	79.0
mfDiMP [102]	34.3	44.7	42.8	64.6
DAPNet [103]	31.4	43.1	-	-
CAT [104]	31.4	45.0	56.1	80.4
HMFT [105]	31.3	43.6	-	-
MaCNet [106]	-	-	55.4	79.0
FANet [107]	30.9	44.1	55.3	78.7
DAFNet [108]	-	-	54.4	79.6
SGT [109]	25.1	36.5	47.2	72.0
Non-real Time Trackers				
SeqTrackv2-L384 [71]	61.0	76.7	68.0	91.3
SeqTrackv2-L256 [71]	58.8	74.1	68.5	92.3

outperforms previous reference modality-specific counterparts on 12 visual and vision-language tracking benchmarks, which proves the effectiveness and generalization of UniSOT.

RGB-Depth Tracking. VOT-RGBD22 [73] is the most recent dataset in RGB-D tracking, consisting of 127 short-term RGB-D sequences. DepthTrack [16] is a comprehensive long-term RGBD tracking benchmark, which contains 150 training and 50 testing videos featuring 15 per-frame attributes. We evaluate UniSOT on DepthTrack and VOT-RGBD22 and compare it with the state-of-the-art RGBD trackers in Table 7. UniSOT-B achieves 73.2% EAO on VOT-RGBD22 and 62.5% F-score on DepthTrack, surpassing Un-Track with a large margin. With a stronger foundation tracker, UniSOT-L achieves higher performance on both VOT-RGBD22 and DepthTrack.

RGB-Thermal Tracking. LasHeR [80] is a large-scale dataset with high diversity for RGB+Thermal tracking, which comprises 979 videos for training and 245 videos for testing. RGBT234 [110] is a large-scale RGBT tracking dataset containing 234 videos and adopts MSR and MPR as evaluation criteria. As shown in Table 8, UniSOT-B outperforms Un-Track trackers with 2.7% AUC on LasHeR and 2.5% MSR on RGBT234. Meanwhile, UniSOT-L obtains strong performance of 54.8% AUC on LasHeR and 65.0% MSR on RGBT234 surpassing Un-Track with a large margin.

RGB-Event Tracking. VisEvent [46] is the largest benchmark for RGBE tracking, which contains 500 videos for training and 320 videos for testing. Table 9 shows that UniSOT-B achieves 60.7% AUC, which surpasses the Un-Track with 1.8% AUC. Furthermore, UniSOT-L achieves the AUC score of 62.0%, which performs best among real-time trackers.

Discussion. UniSOT-B and UniSOT-L are running at 48 FPS and 21 FPS with auxiliary video modalities. The notation UniSOT-L* and UniSOT-B* in Table 7-9 indicates that these models are trained only with RGB data in the first training stage. We can find that UniSOT only achieves limited performance without the second training stage. This is because RGB images have inherent shortcomings in challenging scenarios that have few appearance cues. After two stages of training, UniSOT achieves significantly robust performance with auxiliary video modalities, reflecting their necessity in complex scenarios. Further, our UniSOT achieves competitive performance among video modality-specific counterparts on all three auxiliary video modalities (RGBD, RGBT, RGBE). Notably, UniSOT only needs to be fine-tuned once to achieve high-performance tracking with all three auxiliary video

TABLE 9
Comparison with state-of-the-art RGBE trackers.

Method	VisEvent	
	AUC	P
Unified Model with Uniform Parameters		
UniSOT-L	**62.0**	**79.0**
UniSOT-B	60.7	78.0
UniSOT-L*	53.2	69.2
UniSOT-B*	52.6	69.6
Un-Track [17]	58.9	75.5
Unified Model with Separate Parameters		
ViPT [13]	59.2	75.8
ProTrack [2]	47.1	63.2
Specialized Tracker for RGBE Tracking		
TransT_E [29]	47.4	65.0
PrDiMP_E [111]	45.3	64.4
STARK_E [6]	44.6	61.2
MDNet_E [70]	42.6	66.1
VITAL_E [112]	41.5	64.9
ATOM_E [24]	41.2	60.8
SiamBAN_E [25]	40.5	59.1
SiamMask_E [113]	36.9	56.2
Non-real Time Trackers		
SiamRCNN_E [114]	49.9	65.9
SeqTrackv2-L384 [71]	63.4	80.0
SeqTrackv2-L256 [71]	63.0	79.4

TABLE 10
Analysis of different components in UniSOT.

Method	TNL2K						FPS
	BBOX		NL		NL+BBOX		
	AUC	P	AUC	P	AUC	P	
baseline	57.3	57.8	49.6	49.8	57.4	58.5	58
+MMCLoss	58.7	60.2	51.9	52.3	59.8	61.2	58
+RABH	**60.9**	**61.9**	**53.9**	**54.4**	**61.2**	**62.8**	57

modalities. Meanwhile, different video modalities share a unified parameter set, which avoids redundant video modality-specific parameter sets. Recently, Un-Track designs a unified architecture and learns video modality-specific features by projecting auxiliary video modalities into different low-rank spaces with an identical rank. Such a design poses a challenge for video modality-aligned feature learning and ignores the potential difference in the amount of information across video modalities, which may desire various ranks. Conversely, UniSOT dynamically adjusts the rank of shared parameters for both video modality-aligned and video modality-specific feature learning, enabling UniSOT to exceed Un-Track by a large margin. Moreover, UniSOT supports tracking with different reference modalities (NL, BBOX, NL+BBOX) in sequences of different video modalities (RGB, RGB+X) under a unified architecture with uniform parameters, which has more generalized application scenarios. More combinations of reference and video modalities are discussed in Supplementary Materials.

4.3 Ablation Study

The following experiments use UniSOT-B as the base model. The baseline is UniSOT without MMCLoss constraint and using the anchor-free head for localization. FPS is test under vision-language tracking.

Effectiveness of the Different Components. Table 10 shows the performance of UniSOT with different components. Due to the semantic gap between vision and language, different reference modalities may push the baseline model toward divergent optimization directions during joint training, resulting in compromised performance for all reference modalities. MMCLoss brings 1.4%, 2.3% and 2.4% AUC gains for BBOX, NL and BBOX+NL reference modalities 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 reference-adaptive box head (RABH) brings 2.2%, 2.0% and 1.4% AUC gains for BBOX, NL and BBOX+NL reference modalities respectively with minimal efficiency sacrifice. As shown in Figure 10, the anchor free head shows limited target localization results. The reason is that enhanced features of search region vary with reference modalities, while the box head is fixed in tracking, which leads to compromising results across reference modalities. Differently, we design a dynamic head (RABH), which can make full use of reference information to localize target object in a contrastive way, improving tracking performance across all reference modalities.

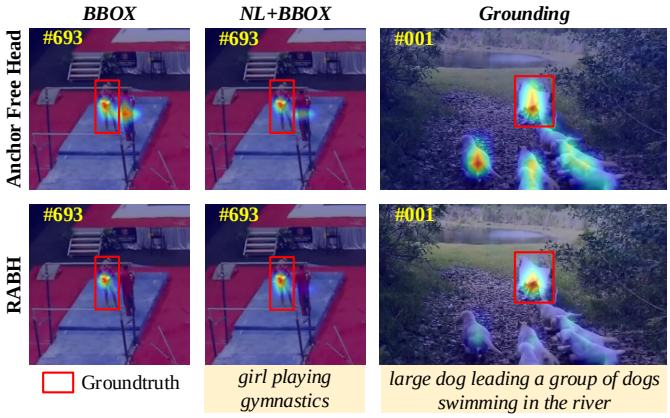


Fig. 10. Visualization of target localization results. RABH obtains more stable localization results. This is because RABH can adaptively capture scenario cues and localize target object in a contrastive way, thus suppressing distractors.

TABLE 11
Analysis of the reference-generalized feature extractor.

N	M	TNL2K						FPS
		BBOX		NL		NL+BBOX		
		AUC	P	AUC	P	AUC	P	
0	12	59.7	60.6	53.4	53.6	60.4	61.2	59
3	9	59.9	60.8	**54.1**	**54.5**	61.1	62.7	58
6	6	**60.9**	**61.9**	53.9	54.4	**61.2**	**62.8**	57
9	3	60.8	61.8	53.0	52.9	60.9	62.0	55
11	1	60.7	61.6	45.1	41.8	60.3	61.3	54

TABLE 12
Analysis of the multi-modal contrastive loss designing.

pos.	neg.	N_{neg}	TNL2K					
			BBOX		NL		NL+BBOX	
			AUC	P	AUC	P	AUC	P
avg	$rand$	9	59.4	60.2	51.8	51.1	59.5	60.2
ctr	$rand$	9	59.9	60.8	53.0	53.1	60.1	61.3
avg	top	9	60.3	61.1	53.5	53.5	60.4	61.7
ctr	top	9	**60.9**	**61.9**	**53.9**	**54.4**	**61.2**	**62.8**
ctr	top	1	59.7	60.6	52.5	52.6	59.9	61.1
ctr	top	5	60.4	61.2	53.5	53.8	60.7	62.1
ctr	top	13	60.8	61.8	53.8	54.2	**61.2**	62.7

Analysis of the Reference-Generalized Feature Extractor. As shown in Table 11, more separate layers (larger N) are beneficial for visual tracking and more fusion layers (larger M) are beneficial for vision-language tracking. Moreover, the efficiency is reduced by more separate layers as the serial inference of vision and language encoders. However, when we fuse visual and lan-

TABLE 13
Analysis of the distractor threshold β in the distribution-based cross-attention mechanism.

β	TNL2K					
	BBOX		NL		NL+BBOX	
	AUC	P	AUC	P	AUC	P
0.00	59.5	60.1	52.5	52.8	60.0	60.7
0.25	60.3	61.1	53.2	53.5	60.5	61.7
0.50	60.6	61.6	53.6	54.1	60.9	62.5
0.75	**60.9**	**61.9**	**53.9**	**54.4**	**61.2**	**62.8**
0.85	60.8	61.8	53.6	53.8	60.9	62.4

TABLE 14
Analysis of the Robustness under Extreme Scenarios.

Method	OTB99					
	BBOX		NL		NL+BBOX	
	AUC	P	AUC	P	AUC	P
Original Image	69.0	89.4	60.1	79.1	69.3	89.9
Low-contrast Image	68.3	88.5	59.6	78.6	68.7	89.2
Noisy Image	68.1	88.2	59.2	78.4	68.5	88.8

TABLE 15
Analysis of the model auxiliary modality tuning block (AMTB).

Module	DepthTrack			RGBT234		VisEvent	
	Pr.	Re.	F	MPR	MSR	Pr.	AUC
MCP*	60.6	59.6	60.1	84.1	61.5	76.4	59.3
MCP	60.2	59.0	59.6	83.7	61.2	75.9	58.9
LoRA	60.1	59.0	59.5	84.1	61.5	76.1	59.1
FixedRank	60.5	60.3	60.4	83.4	61.2	77.0	59.9
AMTB*	62.6	61.9	62.2	86.4	64.0	77.6	60.2
AMTB	**62.9**	**62.2**	**62.5**	**86.6**	**64.1**	**78.0**	**60.7**

TABLE 16
Analysis of the rank budget of AMTB.

index	$\hat{n}$	$\hat{m}$	DepthTrack			RGBT234		VisEvent		FPS
			Pr.	Re.	F	MPR	MSR	Pr.	AUC	
1	8	16	61.8	61.3	61.6	85.8	63.5	76.7	59.7	50
2	16	16	62.0	61.6	61.8	86.1	63.7	77.2	60.1	48
3	16	32	**62.9**	**62.2**	**62.5**	**86.6**	**64.1**	**78.0**	**60.7**	48
4	16	48	62.4	62.2	62.3	86.0	63.8	77.5	60.2	48
5	24	48	62.9	61.7	62.1	86.2	64.0	77.8	60.4	47

guage 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 reference modalities. Thus, we set N=6 and M=6 to balance the performance for all reference modalities and efficiency.

Analysis of the Multi-Modal Contrastive Loss. We study different ways to obtain the positive sample and the negative sample. As shown in Table 12, the best results are achieved when we sample the central score of target object 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 target object contains no backgrounds, which is more reliable for expressing target object. Further, more hard-negative samples can improve the discriminability of the semantic token, while an excess of such samples may introduce easy-to-distinguish samples, thereby diluting the weight of difficult-to-distinguish (distractor) samples in MMCLoss and weakening the discriminability optimization of the semantic token toward distractors.

Analysis of the Distractor Threshold. As shown in Table 13, the performance of UniSOT is insensitive to threshold β over a wide range. However, the performance drops overtly if we aggregate all background features into one token (β=0). This is because the distractor feature is vital to discriminate target object 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 Robustness under Extreme Scenarios. To evaluate the performance of UniSOT under sparse or noisy conditions, we constructed sparse and noisy datasets by reducing contrast and adding noise on OTB99 dataset. Experimental results in Table 14 demonstrate that UniSOT maintains comparable performance in both sparse and noisy datasets. These evaluations validate the robustness and generalization capability of our framework across diverse and extreme conditions.

Effectiveness of the Auxiliary Modality Tuning Block. We conduct experiments on the multi-modal fusion module in Table 15. $*$ means fine-tuning each video modality individually as ViPT does. MCP means that we use the MCP block of ViPT to fuse auxiliary video modalities. $LoRA$ means that we add features of auxiliary video modalities to the image features after patch embedding and utilize the LoRA [115] technique to fine-tune the foundation tracker. $FixedRank$ means that we fix singular values in the auxiliary modality tuning block (AMTB) to 1. Other training settings remain consistent with UniSOT. Our designed AMTB outperforms MCP and $LoRA$, which proves the effectiveness of the fusion designing. Meanwhile, AMTB surpasses MCP and $FixedRank$ with a large margin and outperforms AMTB with video modality-separate fine-tuning ($AMTB*$). This is because AMTB constructs a compact modality-shared parameter space via modality-shared rank allocation that facilitates video modality-aligned feature learning. Different auxiliary video modalities benefit from joint fine-tuning with modality-shared parameters to learn more generalized features. Meanwhile, AMTB adaptively assigns different ranks across video modalities via modality-specific rank allocation, enabling specialized feature learning while mitigating overfitting.

Analysis of the Rank Budget. We study the effect of rank budget of AMTB in Table 16. Some concepts need to be declared. The inherent rank r is set to 32. $\hat{n}$ is the block-average modality-shared rank budget and $\hat{m}$ is the block-average modality-specific rank budget as stated in Section 4.1.3. Comparing (1), (3) and (5), we can find that low $\hat{n}$ and $\hat{m}$ are not enough to support the tracker to learn effective features, while high $\hat{n}$ and $\hat{m}$ may lead to overfitting on the training set. Comparing (2), (3) and (4), we can find that a proper ratio of $\hat{n}$ and $\hat{m}$ is beneficial to the tracker performance. The reason is that low $\hat{m}/\hat{n}$ is not conducive to feature learning of modal alignment, while high $\hat{m}/\hat{n}$ limits the ability to learn video modality-specific features. Moreover, larger $\hat{n}$ will reduce inference efficiency. So we select $\hat{n} = 16$ and $\hat{m} = 8$ to keep good performance and efficiency.

Analysis of Incremental Video Modality Learning. We conduct further experiments to explicitly highlight RAMA's effectiveness in enabling incremental video modality learning. Specifically, we implement a two-stage training process on UniSOT: in the first stage, we train UniSOT with RGB combined with a single non-RGB video modality, and in the second stage, we extend the training to RGB plus multiple non-RGB video modalities (denoted as X), including depth, thermal and event. As detailed in Table 17, the first-stage results demonstrate strong performance on evaluations specific to the trained video modality, but exhibit limited generalization to other video modalities. Crucially, after the second-stage training incorporating multiple video modalities,

TABLE 17
Analysis of Incremental Video Modality Learning.

Module	DepthTrack			RGBT234		VisEvent	
	Pr.	Re.	F	MPR	MSR	Pr.	AUC
UniSOT-B (RGB+X)	62.9	62.2	62.5	86.6	64.1	78.0	60.7
stage 1: UniSOT-B (RGB+D)	62.6	61.9	62.2	82.0	61.3	69.2	51.9
stage 2: UniSOT-B (RGB+X)	62.9	62.3	62.6	86.6	64.1	77.9	60.7
stage 1: UniSOT-B (RGB+T)	51.8	54.2	53.1	86.4	64.0	68.8	51.3
stage 2: UniSOT-B (RGB+X)	62.8	62.2	62.5	86.6	64.1	78.0	60.6
stage 1: UniSOT-B (RGB+E)	50.6	53.7	52.1	80.6	60.3	77.6	60.2
stage 2: UniSOT-B (RGB+X)	62.8	62.1	62.4	86.5	64.1	78.1	60.8

TABLE 18
Evaluation of UniSOT with Degraded RGB Inputs.

Module	DepthTrack			RGBT234		VisEvent	
	Pr.	Re.	F	MPR	MSR	Pr.	AUC
UniSOT-B (Degraded RGB)	60.8	58.4	59.6	85.3	63.2	75.3	58.4
UniSOT-B	**62.9**	**62.2**	**62.5**	**86.6**	**64.1**	**78.0**	**60.7**

performance improved consistently across all video modality evaluations, regardless of the initial auxiliary video modality used in the first stage. This is because RAMA can adaptively allocate ranks among video modalities, thereby enabling incremental video modality learning. These results intuitively highlight RAMA's effectiveness in enabling incremental video modality learning.

Evaluation with Degraded RGB Inputs. We explored the model's robustness to degraded RGB inputs by applying contrast reduction to RGB data in multi-modal video datasets. Results in Table 18 demonstrate that UniSOT maintains comparable performance with degraded RGB inputs, which validates its ability to utilize extra cues of auxiliary video modalities for robust tracking. This is because RAMA allows the model to learn generalized representations from shared parameters and adaptively learn features crucial for different video modalities.

4.4 Visualization

Attention Maps and Target Score Maps. We visualize distractor attention maps and background attention maps in the distribution-based cross-attention. As shown in Figure 11, our designed distribution-based cross-attention can effectively divide out-box features into distractors and other backgrounds. By such design, the extracted distractor token can retain more detailed information to distinguish subsequent distractors, improving the robustness of target localization. Further, we visualize target score maps generated by separate distractor and background tokens or joint distractor and background token. Here, "separate" means we aggregate distractor and background features into different tokens, "joint" means we aggregate all out-box features into one token ($\beta = 0$ as discussed in Section 4.3). The comparison of target score maps also shows that separate distractor and background tokens can obtain more accurate localization results. These visualizations demonstrate the effectiveness of our proposed distribution-based cross-attention mechanism.

Response Map of Semantic Tokens. We further provide response maps of visual and language semantic tokens in the search region under both settings, with and without MMCLoss, as shown in Figure 12. In the absence of MMCLoss, two phenomena can be observed: first, the response distributions of visual and language semantic tokens exhibit inconsistency, reflecting the semantic gap between vision and language features. During multi-modal joint training, this inconsistency may cause the model to optimize in divergent directions for different reference modalities, resulting in compromised performance. Second, both visual and language semantic tokens demonstrate limited discriminability in the search region, reflecting the poor robustness of the reference features. Conversely, with MMCLoss, visual and language semantic tokens achieve consistent and discriminative responses, indicating that MMCLoss aligns vision and language features into a unified semantic space and facilitates consistent feature learning across reference modalities while enhancing the discriminability of reference features.

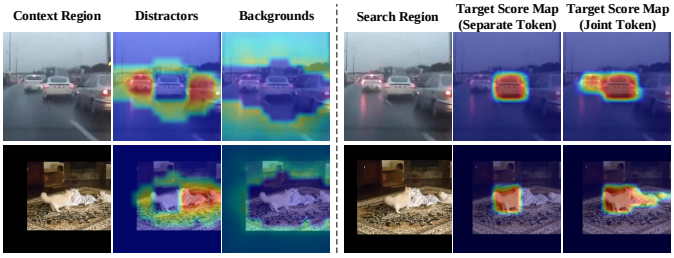


Fig. 11. The left images show the distractor and background attention maps on the context region. The right images show the target score maps generated by separate distractor and background tokens or joint distractor and background token ($\beta = 0.0$).

Center Score Maps with AMTB and LoRA. Comparative visualizations of center score map with AMTB or LoRA in Figure 14 demonstrate that our approach achieves more stable localization across auxiliary video modalities, as shown in Figure 13. This is because AMTB adaptively distributes parameters to simultaneously capture video modality-shared and video modality-specific features as shown in Figure 14, thereby improving multi-modal localization robustness through joint fine-tuning.

4.5 Applications & Limitations

Applications. Drones or professional recording equipment ($e.g.$, camera gimbal) typically use single-object tracking algorithms to continuously estimate the location of the target object during automatic tracking. In practice, users may prefer various reference modalities to specify the target object for convenience, while the demand for different video modalities varies across scenarios to enhance tracking robustness. The generalized capabilities of UniSOT provide users with significant convenience and functional flexibility within a single model, eliminating the need for multiple specialized systems. Additionally, we believe that the exploration of general capability models also potentially drives the development of artificial general intelligence (AGI).

Limitations As highlighted, UniSOT focuses on establishing a unified multi-modal tracking framework with strong generalization across benchmarks. Considering the simplicity of framework, UniSOT does not explicitly model the issues of target disappearance and re-detection, which remain an open challenge for future research. In practical applications, UniSOT can integrate off-the-shelf global detection modules [116] to address target disappearance/reappearance scenarios ($e.g.$, out of view), as STARK-LT [97] does.

5 Conclusion

In this work, we propose a novel unified tracking framework for different combinations of reference and video modalities. To the best of our knowledge, UniSOT is the first tracker to support

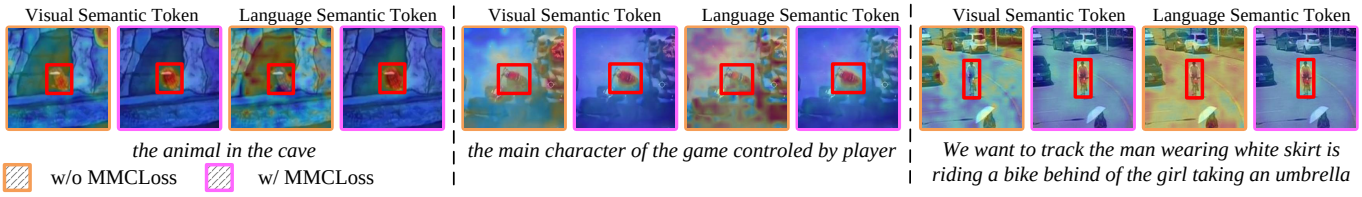


Fig. 12. The response maps of visual and language semantic tokens in the search region under both settings, with and without MMCLoss.

Fig. 13. The classification maps of UniSOT with RAMA and LoRA in RGB+Depth, RGB+Thermal and RGB+Event datasets.

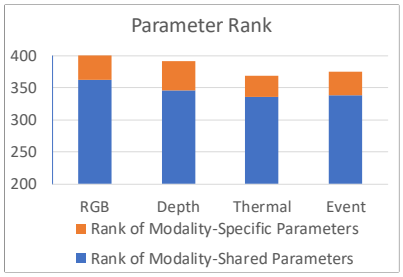


Fig. 14. The parameter rank distribution among video modalities in AMTBs.

TABLE 19
Analysis of the model training strategy of the first stage.

Pretrain	Ratio	TNL2K					
		BBOX		NL		NL+BBOX	
		AUC	P	AUC	P	AUC	P
BERT+MAE	3:1:3	60.6	61.5	53.8	54.3	60.8	62.5
BERT+MAE	4:1:4	60.9	61.9	**53.9**	**54.4**	61.2	62.8
BERT+MAE	5:1:5	**61.0**	**62.1**	53.4	53.7	**61.3**	**63.0**
MAE	4:1:4	60.4	60.7	52.3	51.8	60.1	61.5

TABLE 20
Analysis of the update interval of scenario tokens.

update interval	TNL2K					
	BBOX		NL		NL+BBOX	
	AUC	P	AUC	P	AUC	P
∞	60.3	61.1	52.4	52.2	60.2	61.5
50	60.8	61.8	53.4	53.9	61.0	62.5
20	**60.9**	**61.9**	**53.9**	**54.4**	**61.2**	**62.8**
10	**60.9**	61.8	53.7	54.1	61.1	62.5

tracking with three reference modalities (NL, BBOX, NL+BBOX) in sequences of four video modalities (RGB, RGB+Depth, RGB+Thermal, RGB+Event) under a unified architecture with uniform parameters, enabling more generalized application scenarios. Extensive experimental results on visual tracking, vision-language tracking, grounding and RGB-X (depth, thermal, event) tracking datasets demonstrate that UniSOT shows superior results against modality-specific counterparts.

TABLE 21
Analysis of Single Video Modality Input.

Module	DepthTrack			RGBT234		VisEvent	
	Pr.	Re.	F	MPR	MSR	Pr.	AUC
UniSOT-B* (RGB)	52.7	54.9	53.8	82.2	61.6	69.6	52.6
UniSOT-B (RGB)	53.2	55.2	54.2	82.4	61.7	69.8	52.7
UniSOT-B (X)	26.4	17.0	20.7	28.5	20.8	23.8	14.6
UniSOT-B (RGB+X)	**62.9**	**62.2**	**62.5**	**86.6**	**64.1**	**78.0**	**60.7**

6 Supplementary Materials

6.1 Extra Ablation Study

Analysis of the Training Strategy of the First Stage. As shown in Table 19, the ratio of different references (BBOX:NL:Both) has little effect on the performance of UniSOT. Moreover, we try to initialize all parameters in the backbone with ViT parameters 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.

Analysis of the Update Interval. The scenario tokens in the reference-adaptive box head are updated at regular intervals, which reduces computational costs and avoids error accumulation. As shown in Table 20, the performance is insensitive to the update interval in a certain range. If scenario tokens are not updated, the performance decreases overtly. It demonstrates that dynamic scenario cues are vital for robust tracking.

Analysis of Single Video Modality Input. We conduct experiments to evaluate the performance of RGB-only inputs, UniSOT-B (RGB), and the auxiliary video modality-only inputs, UniSOT-B (X), on the RGB+X tracking setup as shown in Table 21. UniSOT-B* means UniSOT only trained with the first training stage. As we can see, UniSOT-B achieves a slight performance improvement on RGB inputs after the second training stage (fine-tuning with AMTB on RGB+X dataset). These results demonstrate the robustness of the unified architecture when not using the auxiliary video modality. The performance drops significantly when only auxiliary video modality are used, as these modalities lack discriminative texture information, making target localization challenging. When both RGB and auxiliary video modalities serve as inputs, there is a notable performance boost. This proves its capability to effectively leverage multi-modal information for enhanced tracking performance.

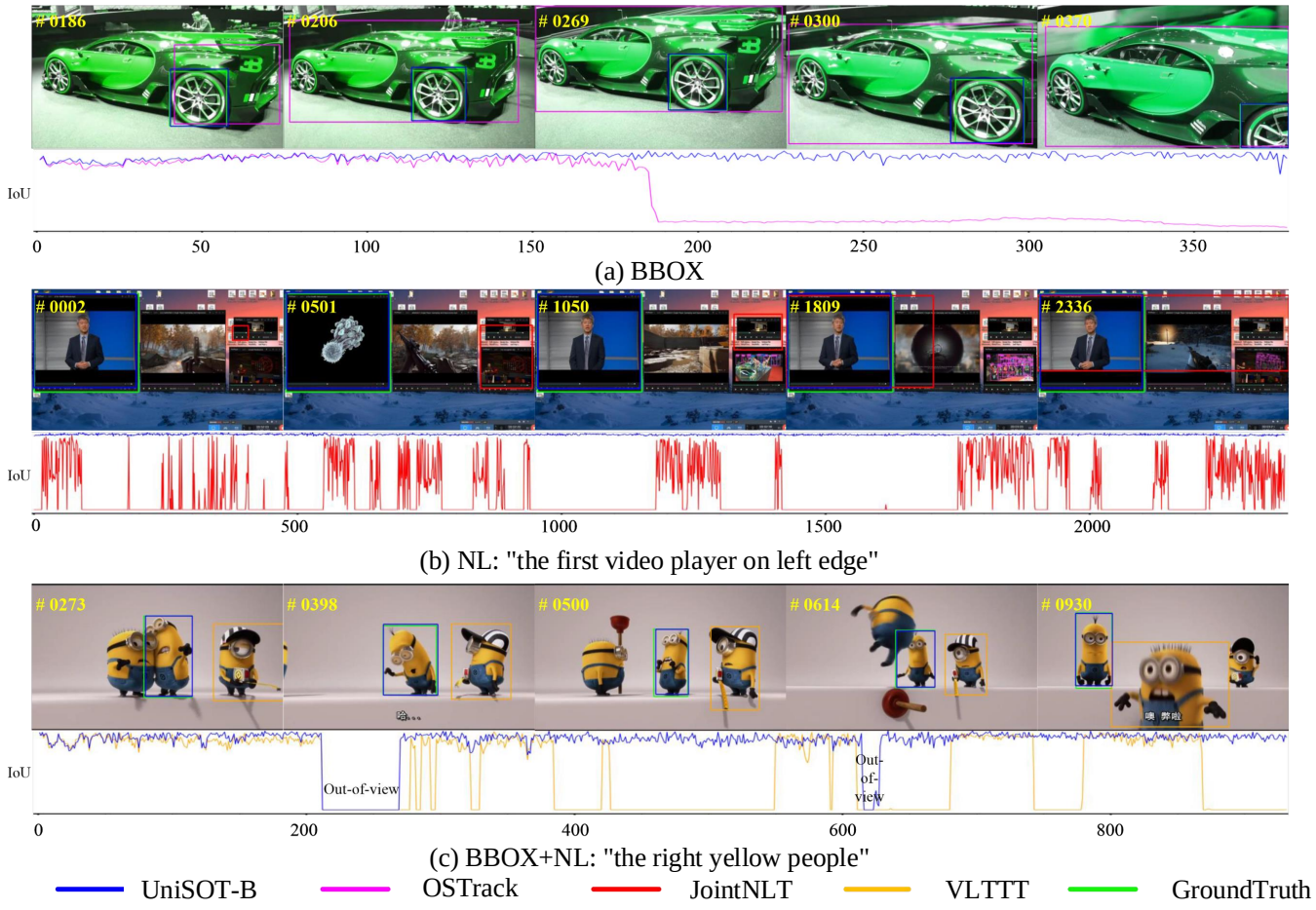


Fig. 15. Visualization of tracking results with different reference modalities. UniSOT achieves more robust tracking performance compared with modality-specific counterparts, which proves the generalization of UniSOT.

6.2 Visualization

Tracking Results. We visualize some tracking results with different reference modalities. As shown in Figure 15(a), compared with OSTrack [18], UniSOT can achieve more robust tracking in the complex scenario with the bounding box reference. This is because our designed reference-adaptive box head can utilize dynamically mined scenario cues to discriminate target object, reducing the impact of complex backgrounds. Meanwhile, as shown in Figure 15(b), UniSOT has excellent visual grounding ability, thus achieving more accurate localization results compared with JointNLT [8] using the natural language reference. Moreover, for tracking by language and box specification, VLTTT [9] suffers from frequent target drift, while UniSOT can maintain high-quality tracking results in long-term video, even after target object disappears and reappears. It benefits from modality-aligned feature learning, which can make full use of different modal references to enhance target features, so as to realize more robust tracking. Further, as shown in Figure 16, UniSOT achieves robust tracking results in challenging scenarios compared with UniSOT and previous state-of-the-art RGB-X trackers. This proves the advantage of the rank-adaptive designing that enables both modality-aligned and modality-specific feature learning. The above results intuitively confirm that our proposed framework can achieve superior performance across combinations of reference and video modalities.

More combinations of reference and video modalities. We further explore the combination of different reference and video modalities. As shown in Figure 17, UniSOT can achieve superior performance under different combinations of reference and video modalities. This proves that UniSOT can track with different reference modalities in sequences of different video modalities. UniSOT not only can simultaneously cope with three types of target references, enabling various human-machine interactions, but also can freely integrate different auxiliary modalities to enhance robustness in complex scenarios.

REFERENCES

[1] Y. Ma, Y. Tang, W. Yang, T. Zhang, J. Zhang, and M. Kang, "Unifying visual and vision-language tracking via contrastive learning," in *Proceedings of AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4107–4116.

[2] Y. Jinyu, L. Zhe, Z. Feng, L. Ales *et al.*, "Prompting for multi-modal tracking," in *ACM International Conference on Multimedia*, 2022, pp. 3492–3500.

[3] Z. Y. Chen Fei, Wang Xiaodong *et al.*, "Visual object tracking: A survey," *Computer Vision and Image Understanding*, vol. 222, p. 103508, 2022.

[4] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via structured multi-task sparse learning," *International Journal of Computer Vision*, vol. 101, no. 2, pp. 367–383, 2013.

[5] T. Zhang, C. Xu, and M.-H. Yang, "Robust structural sparse tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 473–486, 2018.

[6] F. J. Yan Bin, Peng Houwen *et al.*, "Learning spatio-temporal transformer for visual tracking," in *Proceedings of IEEE International Conference on Computer Vision*, 2021, pp. 10 448–10 457.

Fig. 16. Visualization of tracking results with different auxiliary modalities. UniSOT achieves more robust tracking performance compared with previous state-of-the-art trackers.

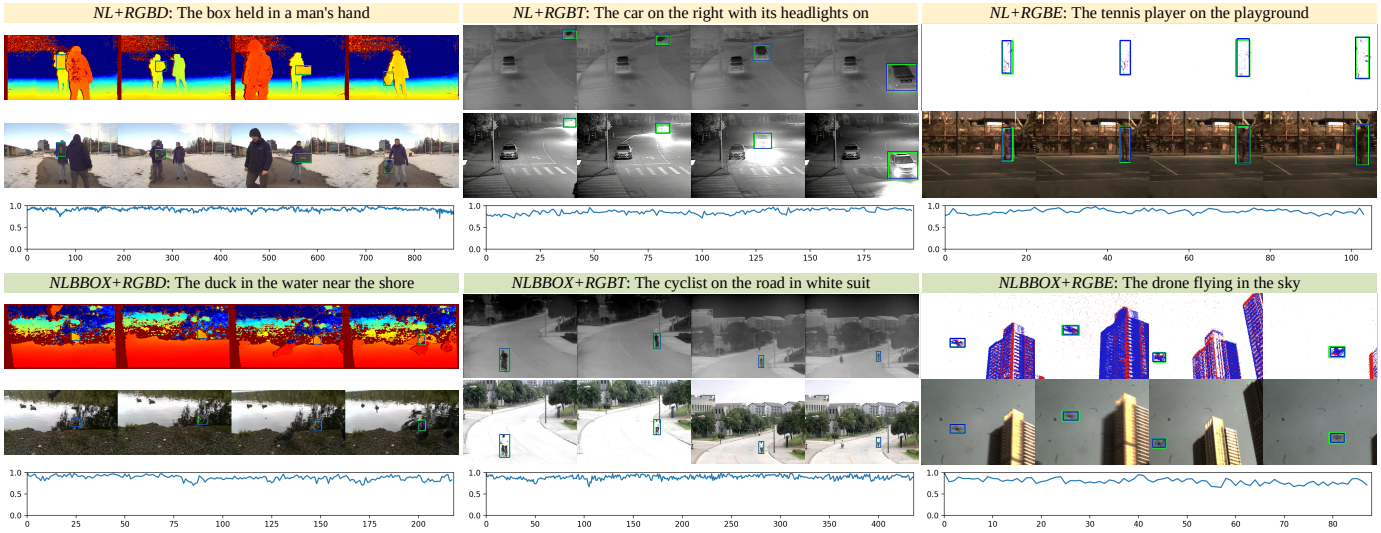


Fig. 17. Tracking with different combinations of reference and video modalities. The green box is groundtruth, blue box is the predicted box.

[7] G. E. Li Zhenyang, Tao Ran *et al.*, "Tracking by natural language specification," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6495–6503.

[8] M. K. Zhou Li, Zhou Zikun *et al.*, "Joint visual grounding and tracking with natural language specification," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 151–23 160.

[9] F. H. Guo Mingzhe, Zhang Zhipeng *et al.*, "Divert more attention to vision-language tracking," *Advances of Neural Information Processing Systems*, vol. 35, pp. 4446–4460, 2022.

[10] B. Q. Feng Qi, Ablavsky Vitaly *et al.*, "Siamese natural language tracker: Tracking by natural language descriptions with siamese trackers," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5851–5860.

[11] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *Proceedings of European Conference on Computer Vision Workshops*, 2016, pp. 850–865.

[12] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proceedings of IEEE International Conference on Computer Vision*, 2019, pp. 6181–6190.

[13] Z. Jiawen, L. Simiao, C. Xin, W. Dong *et al.*, "Visual prompt multi-modal tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9516–9526.

[14] J. Zhang, X. Yang, Y. Fu, X. Wei, B. Yin, and B. Dong, "Object tracking by jointly exploiting frame and event domain," in *Proceedings of IEEE International Conference on Computer Vision*, 2021, pp. 13 043–13 052.

[15] X. Wang, X. Shu, S. Zhang, B. Jiang, Y. Wang, Y. Tian, and F. Wu, "Mfgnet: Dynamic modality-aware filter generation for rgb-t tracking," *IEEE Transactions on Multimedia*, vol. 25, pp. 4335–4348, 2022.

[16] S. Yan, J. Yang, J. Käpylä, F. Zheng, A. Leonardis, and J.-K. Kämäräinen, "Depthtrack: Unveiling the power of rgbd tracking," in *Proceedings of IEEE International Conference on Computer Vision*, 2021, pp. 10 725–10 733.

[17] W. Zongwei, Z. Jilai, R. Xiangxuan, V. Florin-Alexandru *et al.*, "Single-model and any-modality for video object tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 156–19 166.

[18] M. B. Ye Botao, Chang Hong *et al.*, "Joint feature learning and relation modeling for tracking: A one-stream framework," *Proceedings of European Conference on Computer Vision*, pp. 341–357, 2022.

[19] Y. D. Ma Yinchao, He Jianfeng *et al.*, "Adaptive part mining for robust visual tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 11 443–11 457, 2023.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances of Neural Information Processing Systems*, 2017, pp. 5999–6009.

[21] Q. Zhang, M. Chen, A. Bukharin, N. Karampatziakis, P. He, Y. Cheng, W. Chen, and T. Zhao, "Adalora: Adaptive budget allocation for parameter-efficient fine-tuning," in *International Conference on Learning Representations*, 2023.

[22] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, 2015.

[23] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "Siamrpn++: Evolution of siamese visual tracking with very deep networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4277–4286.

[24] F. S. K. Martin Danelljan, Goutam Bhat *et al.*, "Atom: Accurate tracking by overlap maximization," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4655–4664.

[25] Z. Chen, B. Zhong, G. Li, S. Zhang, R. Ji *et al.*, "Siamese box adaptive network for visual tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6667–6676.

[26] Y. Cui, D. Guo, Y. Shao, Z. Wang, C. Shen, L. Zhang, and S. Chen, "Joint classification and regression for visual tracking with fully convolutional siamese networks," *International Journal of Computer Vision*, pp. 1–17, 2022.

[27] T. Xu, Z. Feng, X.-J. Wu *et al.*, "Adaptive channel selection for robust visual object tracking with discriminative correlation filters," *International Journal of Computer Vision*, vol. 129, pp. 1359–1375, 2021.

[28] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg, "Eco: Efficient convolution operators for tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6638–6646.

[29] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8122–8131.

[30] S. Gao, C. Zhou, C. Ma, X. Wang, and J. Yuan, "Aiatrack: Attention in attention for transformer visual tracking," in *Proceedings of European Conference on Computer Vision*, 2022, pp. 146–164.

[31] Y. Cui, C. Jiang, L. Wang, and G. Wu, "Mixformer: End-to-end tracking with iterative mixed attention," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 608–13 618.

[32] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of IEEE International Conference on Computer Vision*, 2021, pp. 22–31.

[33] Y. Li, J. Yu, Z. Cai, and Y. Pan, "Cross-modal target retrieval for tracking by natural language," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4931–4940.

[34] Z. Z. Wang Xiao, Shu Xiujun *et al.*, "Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 763–13 773.

[35] Y. Zhengyuan, K. Tushar, Chen, Tianlang *et al.*, "Grounding-tracking-integration," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 9, pp. 3433–3443, 2020.

[36] Y. Shao, S. He, Q. Ye, Y. Feng, W. Luo, and J. Chen, "Context-aware integration of language and visual references for natural language tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 208–19 217.

[37] K. Dai, D. Wang, H. Lu, C. Sun, and J. Li, "Visual tracking via adaptive spatially-regularized correlation filters," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4665–4674.

[38] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proceedings of European Conference on Computer Vision*, 2016, pp. 472–488.

[39] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 3, pp. 583–596, 2014.

[40] Y. Song, C. Ma, L. Gong, J. Zhang, R. Lau, and M.-H. Yang, "CREST: Convolutional residual learning for visual tracking," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 2574–2583.

[41] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8971–8980.

[42] Z. Fu, Q. Liu, Z. Fu, and Y. Wang, "Stmtrack: Template-free visual tracking with space-time memory networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 774–13 783.

[43] B. Chen, P. Li, L. Bai, L. Qiao, Q. Shen, B. Li, W. Gan, W. Wu, and W. Ouyang, "Backbone is all your need: A simplified architecture for visual object tracking," in *Proceedings of European Conference on Computer Vision*, 2022, pp. 375–392.

[44] X.-F. Zhu, T. Xu, Z. Tang, Z. Wu, H. Liu, X. Yang, X.-J. Wu, and J. Kittler, "Rgbd1k: A large-scale dataset and benchmark for rgb-d object tracking," in *Proceedings of AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 3870–3878.

[45] Y. Xiao, M. Yang, C. Li, L. Liu, and J. Tang, "Attribute-based progressive fusion network for rgbt tracking," in *Proceedings of AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2831–2838.

[46] X. Wang, J. Li, L. Zhu, Z. Zhang, Z. Chen, X. Li, Y. Wang, Y. Tian, and F. Wu, "Visevent: Reliable object tracking via collaboration of frame and event flows," *IEEE Transactions on Cybernetics*, vol. 54, no. 3, pp. 1997–2010, 2024.

[47] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, vol. 139, 2021, pp. 8748–8763.

[48] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances of Neural Information Processing Systems*, vol. 34, pp. 9694–9705, 2021.

[49] J. Yang, Y. Bisk, and J. Gao, "Taco: Token-aware cascade contrastive learning for video-text alignment," in *Proceedings of IEEE International Conference on Computer Vision*, 2021, pp. 11 562–11 572.

[50] C. Mayer, M. Danelljan, D. P. Paudel, and L. Van Gool, "Learning target candidate association to keep track of what not to track," in *Proceedings of IEEE International Conference on Computer Vision*, 2021, pp. 13 444–13 454.

[51] V. Sanh, T. Wolf, and A. Rush, "Movement pruning: Adaptive sparsity by fine-tuning," *Advances of Neural Information Processing Systems*, vol. 33, pp. 20 378–20 389, 2020.

[52] Q. Zhang, S. Zuo, C. Liang, A. Bukharin, P. He, W. Chen, and T. Zhao, "Platon: Pruning large transformer models with upper confidence bound of weight importance," in *International conference on machine learning*, 2022, pp. 26 809–26 823.

[53] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling, "LaSOT: A high-quality benchmark for large-scale single object tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5369–5378.

[54] L. Huang, X. Zhao, K. Huang *et al.*, "Got-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1562–1577, 2021.

[55] T.-Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proceedings of European Conference on Computer Vision*, 2014, pp. 740–755.

[56] M. Muller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem, "TrackingNet: A large-scale dataset and benchmark for object tracking in the wild," in *Proceedings of European Conference on Computer Vision*, 2018, pp. 310–327.

[57] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 11–20.

[58] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Conference of the North American Chapter of the Association for Computational Linguistics*, vol. 1, 2019, pp. 4171–4186.

[59] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[60] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[61] B. Yan, Y. Jiang, P. Sun, D. Wang, Z. Yuan, P. Luo, and H. Lu, "Towards grand unification of object tracking," in *Proceedings of European Conference on Computer Vision*, 2022, pp. 733–751.

[62] F. Xie, Z. Wang, and C. Ma, "Diffusiontrack: Point set diffusion model for visual object tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 2024, pp. 19 113–19 124.

[63] B. Sun, Z. Wang, S. Wang, Y. Cheng, J. Ning *et al.*, "Bidirectional interaction of cnn and transformer feature for visual tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.

[64] R. Wang, B. Zhong, and Y. Chen, "Motion-driven tracking via end-to-end coarse-to-fine verifying," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 2, pp. 1007–1019, 2023.

[65] C. Tang, X. Wang, Y. Bai, Z. Wu, J. Zhang, and Y. Huang, "Learning spatial-frequency transformer for visual object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 5102–5116, 2023.

[66] N. Wang, W. Zhou, J. Wang, and H. Li, "Transformer meets tracker: Exploiting temporal context for robust visual tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1571–1580.

[67] Z. Zhang, H. Peng, J. Fu, B. Li, W. Hu *et al.*, "Ocean: Object-aware anchor-free tracking," in *Proceedings of European Conference on Computer Vision*, 2020, pp. 771–787.

[68] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, "Know your surroundings: Exploiting scene information for object tracking," in *Proceedings of European Conference on Computer Vision*, 2020, pp. 205–221.

[69] X. Yinda, W. Zeyu, L. Zuoxin *et al.*, "SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2020, pp. 12 549–12 556.

[70] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4293–4302.

[71] X. Chen, B. Kang, J. Zhu, D. Wang, H. Peng, and H. Lu, "Unified sequence-to-sequence learning for single- and multi-modal visual object tracking," 2024. [Online]. Available: https://arxiv.org/abs/2304.14394

[72] L. Lin, H. Fan, Y. Xu, and H. Ling, "Swintrack: A simple and strong baseline for transformer tracking," *Advances of Neural Information Processing Systems*, 2021.

[73] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, J.-K. Kämäräinen, H. J. Chang, M. Danelljan, L. Č. Zajc, A. Lukežič *et al.*, "The tenth visual object tracking vot2022 challenge results," in *Proceedings of European Conference on Computer Vision*, 2022, pp. 431–460.

[74] M. Kristan, J. Matas, M. Danelljan, M. Felsberg, and et al., "The first visual object tracking segmentation vots2023 challenge results," in *Proceedings of IEEE International Conference on Computer Vision Workshops*, 2023, pp. 1788–1810.

[75] Z. Yang and Y. Yang, "Decoupling features in hierarchical propagation for video object segmentation," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.

[76] H. K. Cheng, S. W. Oh, B. Price, J.-Y. Lee, and A. Schwing, "Putting the object back into video object segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[77] Z. Yang, Y. Wei, and Y. Yang, "Associating objects with transformers for video object segmentation," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021.

[78] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2017.

[79] Ilya and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *International Conference on Learning Representations*, 2017.

[80] C. Li, W. Xue, Y. Jia, Z. Qu, B. Luo, J. Tang, and D. Sun, "Lasher: A large-scale high-diversity benchmark for rgbt tracking," *IEEE Transactions on Image Processing*, vol. 31, pp. 392–404, 2021.

[81] M. Noman, W. A. Ghallabi, D. Najiha, C. Mayer, A. Dudhane, M. Danelljan, H. Cholakkal, S. Khan, L. Van Gool, and F. S. Khan, "Avist: A benchmark for visual object tracking in adverse visibility," in *Proceedings of British Machine Vision Conference*, 2022.

[82] H. Fan, H. Bai, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, M. Huang, J. Liu, Y. Xu *et al.*, "Lasot: A high-quality large-scale single object tracking benchmark," *International Journal of Computer Vision*, vol. 129, no. 2, pp. 439–461, 2021.

[83] H. Kiani, Galoogahi, A. Fagg, C. Huang, D. Ramanan, S. Lucey *et al.*, "Need for speed: A benchmark for higher frame rate object tracking," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 1134–1143.

[84] M. Mueller, N. Smith, B. Ghanem *et al.*, "A benchmark and simulator for UAV tracking," in *Proceedings of European Conference on Computer Vision*, 2016, pp. 445–461.

[85] Q. Feng, V. Ablavsky, Q. Bai, and S. Sclaroff, "Robust visual object tracking with natural language region proposal network," *arXiv preprint arXiv:1912.02048*, vol. 1, no. 7, p. 8, 2019.

[86] Q. Feng, V. Ablavsky, Q. Bai, G. Li, and S. Sclaroff, "Real-time visual object tracking with natural language description," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 700–709.

[87] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *Proceedings of European Conference on Computer Vision*, 2016, pp. 69–85.

[88] L. Yang, Y. Xu, C. Yuan, W. Liu, B. Li, and W. Hu, "Improving visual grounding with visual-linguistic verification and iterative reasoning," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9499–9508.

[89] C. Zhu, Y. Zhou, Y. Shen, G. Luo, X. Pan, M. Lin, C. Chen, L. Cao, X. Sun, and R. Ji, "Seqtr: A simple yet universal network for visual grounding," in *Proceedings of European Conference on Computer Vision*, 2022, pp. 598–615.

[90] J. Ye, J. Tian, M. Yan, X. Yang, X. Wang, J. Zhang, L. He, and X. Lin, "Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 502–15 512.

[91] J. Deng, Z. Yang, T. Chen, W. Zhou, and H. Li, "Transvg: End-to-end visual grounding with transformers," in *Proceedings of IEEE International Conference on Computer Vision*, 2021, pp. 1769–1779.

[92] L. Chen, W. Ma, J. Xiao, H. Zhang, and S.-F. Chang, "Ref-nms: Breaking proposal bottlenecks in two-stage referring expression grounding," in *Proceedings of AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1036–1044.

[93] B. Huang, D. Lian, W. Luo, and S. Gao, "Look before you leap: Learning landmark features for one-stage visual grounding," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 888–16 897.

[94] Z. Yang, T. Chen, L. Wang, and J. Luo, "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*, 2020, pp. 387–404.

[95] D. Liu, H. Zhang, F. Wu, and Z.-J. Zha, "Learning to assemble neural module tree networks for visual grounding," in *Proceedings of IEEE International Conference on Computer Vision*, 2019, pp. 4673–4682.

[96] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, J.-K. Kämäräinen, M. Danelljan, L. Č. Zajc, A. Lukežič, O. Drbohlav *et al.*, "The eighth visual object tracking vot2020 challenge results," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, 2020, pp. 547–601.

[97] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, R. Pflugfelder, J.-K. Kämäräinen, H. J. Chang, M. Danelljan, L. Cehovin, A. Lukežič *et al.*, "The ninth visual object tracking vot2021 challenge results," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2711–2738.

[98] Q. Wang, Y. Bai, and H. Song, "Middle fusion and multi-stage, multi-form prompts for robust rgb-t tracking," *Neurocomputing*, vol. 596, p. 127959, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231224007306

[99] T. Zhang, H. Guo, Q. Jiao, Q. Zhang, and J. Han, "Efficient rgb-t tracking via cross-modality distillation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 5404–5413.

[100] C. Wang, C. Xu, Z. Cui, L. Zhou, T. Zhang, X. Zhang, and J. Yang, "Cross-modal pattern-propagation for RGB-T tracking," in *CVPR*, 2020, pp. 7064–7073.

[101] P. Zhang, J. Zhao, C. Bo, D. Wang, H. Lu, and X. Yang, "Jointly modeling motion and appearance cues for robust RGB-T tracking," *IEEE Transactions on Image Processing*, pp. 3335–3347, 2021.

[102] L. Zhang, M. Danelljan, A. Gonzalez-Garcia, J. van de Weijer, and F. Shahbaz Khan, "Multi-modal fusion for end-to-end RGB-T tracking," in *Proceedings of IEEE International Conference on Computer Vision Workshops*, 2019, pp. 2252–2261.

[103] Y. Zhu, C. Li, B. Luo, J. Tang, and X. Wang, "Dense feature aggregation and pruning for RGBT tracking," in *ACM International Conference on Multimedia*, 2019, pp. 465–472.

[104] C. Li, L. Liu, A. Lu, Q. Ji, and J. Tang, "Challenge-aware RGBT tracking," in *Proceedings of European Conference on Computer Vision*, 2020, pp. 222–237.

[105] P. Zhang, J. Zhao, D. Wang, H. Lu, and X. Ruan, "Visible-thermal uav tracking: A large-scale benchmark and new baseline," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8886–8895.

[106] H. Zhang, L. Zhang, L. Zhuo, and J. Zhang, "Object tracking in RGB-T videos using modal-aware attention network and competitive learning," *Sensors*, p. 393, 2020.

[107] Y. Zhu, C. Li, J. Tang, and B. Luo, "Quality-aware feature aggregation network for robust RGBT tracking," *IEEE Transactions on Intelligent Vehicles*, pp. 121–130, 2020.

[108] Y. Gao, C. Li, Y. Zhu, J. Tang, T. He, and F. Wang, "Deep adaptive fusion network for high performance RGBT tracking," in *Proceedings of IEEE International Conference on Computer Vision Workshops*, 2019, pp. 91–99.

[109] C. Li, N. Zhao, Y. Lu, C. Zhu, and J. Tang, "Weighted sparse representation regularized graph learning for RGB-T object tracking," in *ACM International Conference on Multimedia*, 2017, pp. 1856–1864.

[110] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "Rgb-t object tracking: Benchmark and baseline," *Pattern Recognition*, vol. 96, p. 106977, 2019.

[111] M. Danelljan, L. V. Gool, and R. Timofte, "Probabilistic regression for visual tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7183–7192.

[112] Y. Song, C. Ma, X. Wu, L. Gong, L. Bao, W. Zuo, C. Shen, R. W. Lau, and M.-H. Yang, "VITAL: VIsual Tracking via Adversarial Learning," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 10, 2022, pp. 121 230–121 248.

[113] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1328–1338.

[114] P. Voigtlaender, J. Luiten, P. H. S. Torr, and B. Leibe, "Siam R-CNN: Visual tracking by re-detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6577–6587.

[115] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *International Conference on Learning Representations*, 2022.

[116] L. Huang, X. Zhao, and K. Huang, "Globaltrack: A simple and strong baseline for long-term tracking," *Proceedings of AAAI Conference on Artificial Intelligence*, vol. 34, pp. 11 037–11 044, 04 2020.