SDSTrack: Self-Distillation Symmetric Adapter Learning for Multi-Modal Visual Object Tracking

Xiaojun Hou[1], Jiazheng Xing[1], Yijie Qian[1], Yaowei Guo[1], Shuo Xin[1], Junhao Chen[1], Kai Tang[1], Mengmeng Wang[1], Zhengkai Jiang[2], Liang Liu[3*], Yong Liu[1*]
[1]Zhejiang University [2]Youtu Lab, Tencent [3]Huzhou Institute, Zhejiang University
[1]{xiaojunhou, jiazhengxing, yijieqian, yaoweiguo, shuoxin, chenjunhao kaitang, mengmengwang}@zju.edu.cn [1*]yongliu@iipc.zju.edu.cn
[2]zhengkjiang@tencent.com [3*]leonliuz@zju.edu.cn

2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-5300-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPR52733.2024.02507

Abstract

Multimodal Visual Object Tracking (VOT) has recently gained significant attention due to its robustness. Early research focused on fully fine-tuning RGB-based trackers, which was inefficient and lacked generalized representation due to the scarcity of multimodal data. Therefore, recent studies have utilized prompt tuning to transfer pre-trained RGB-based trackers to multimodal data. However, the modality gap limits pre-trained knowledge recall, and the dominance of the RGB modality persists, preventing the full utilization of information from other modalities. To address these issues, we propose a novel symmetric multimodal tracking framework called SDSTrack. We introduce lightweight adaptation for efficient fine-tuning, which directly transfers the feature extraction ability from RGB to other domains with a small number of trainable parameters and integrates multimodal features in a balanced, symmetric manner. Furthermore, we design a complementary masked patch distillation strategy to enhance the robustness of trackers in complex environments, such as extreme weather, poor imaging, and sensor failure. Extensive experiments demonstrate that SDSTrack outperforms state-of-the-art methods in various multimodal tracking scenarios, including RGB+Depth, RGB+Thermal, and RGB+Event tracking, and exhibits impressive results in extreme conditions. Our source code is available at: https://github.com/hoqolo/SDSTrack.

1. Introduction

RGB-based visual object tracking has garnered considerable research attention in recent years, and several works have achieved impressive tracking performance [4–7, 9, 13, 36, 53, 57, 61]. However, the performance of RGB-based trackers often degrades in complex conditions, particularly due to the degradation in imaging quality. This limitation is critical in real-world applications, especially in safety-sensitive scenarios like autonomous driving. Consequently, there is a growing interest in incorporating multimodal images to obtain more comprehensive information, which has also gained considerable attention in other areas such as object detection [23, 24, 46] and semantic segmentation [1–3, 25, 37, 44, 68, 75], among others.

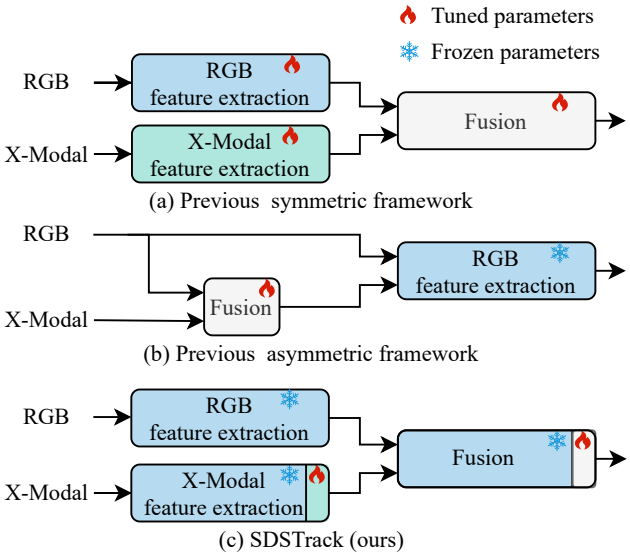

Figure 1. **Previous frameworks *vs.* SDSTrack**. (a) Previous symmetric framework [58] has lots of training parameters and risk of overfitting. (b) Previous asymmetric framework [71] regards RGB as the primary modality and X-Modal as the auxiliary modality with prompt tuning. (c) Our proposed SDSTrack utilizes adapter-based tuning to fine-tune the pre-trained RGB-based tracker in a symmetric manner. "X-Modal" denotes modalities other than RGB, which can be Depth, Thermal, Event, *etc*.

*Corresponding authors.

2575-7075/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPR52733.2024.02507

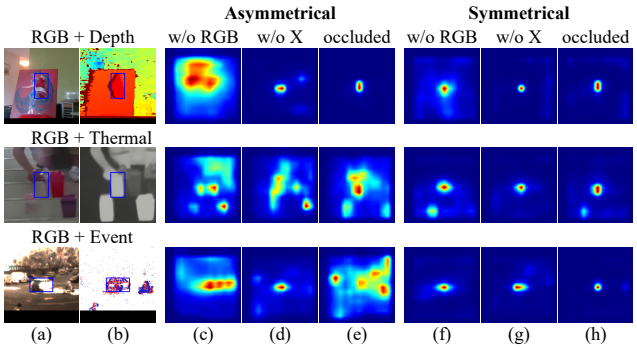

Figure 2. **Input modality dependency comparison of multi-modal object trackers**. (a)-(b) Ground truth of RGB flow and X-modal flow. (c)-(e) Score maps of ViPT [71] under RGB drop, X drop, and multimodal random occlusion conditions. (f)-(h) Score maps of our SDSTrack under RGB drop, X drop, and multimodal random occlusion conditions.

Several previous studies [41, 58, 72–74] have extended pre-trained RGB-based trackers [4, 57, 61] to other modalities, often using a symmetrical framework where information flow is symmetrical, and the pre-trained models are fully fine-tuned (see Fig. 1(a)). However, full fine-tuning introduces numerous training parameters and encounters limitations due to the scarcity of multimodal data, resulting in a lack of generalized representation. As an alternative, some methods [59, 71] employ an asymmetrical framework and utilize prompt tuning to transfer pre-trained RGB-based trackers to other modalities (see Fig. 1(b)), which is a parameter-efficient manner to overcome the multimodal data scarcity limitation.

In these asymmetrical trackers, the dominance of the RGB modality persists, while other modalities serve as auxiliary modalities. We conducted a toy experiment to investigate the performance of trackers utilizing different frameworks in extreme scenarios such as modalities drop and modalities occlusion. As shown in Fig. 2, the results indicate a significant decrease in the performance of asymmetric trackers when the RGB modality is dropped, a moderate decrease when other modalities are dropped, and a noticeable decrease when both modalities are occluded. In contrast, symmetric trackers exhibit relatively robust performance even in challenging conditions. This experiment demonstrates that while the asymmetric structure reduces training costs, it heavily relies on the dominant modality. This reliance hinders the full utilization of information from other modalities, resulting in a lack of robustness.

To address the aforementioned issues, we propose a novel method for multimodal tracking, termed **SDSTrack**. As illustrated in Fig. 1(c), we integrate a symmetric framework with parameter-efficient fine-tuning (PEFT) to effectively explore and fuse multimodal information. We also introduce a complementary masked patch distillation strategy based on self-distillation learning to enhance robustness and accuracy. Specifically, the core concept of PEFT involves freezing the pre-trained model and training only a limited number of additional parameters. This approach ensures the effective transfer of the original pre-trained model to other domains while minimizing training costs. Building on this, we employ the idea of adapter-based tuning to transfer the feature extraction capability of RGB-based trackers to other domains and effectively fuse multimodal features. Moreover, we adopt a symmetrical structure, *i.e.*, there is no precedence among modalities, which prevents the tracker from relying excessively on a specific modality. Furthermore, we design a complementary masked patch distillation strategy by adding random complementary masks to the original patches during training, creating two paths. These paths share a network and perform self-distillation, enabling the model to explore information within multimodal images and improve its tracking capability, even under extreme conditions.

In summary, we make the following contributions:

• We propose a novel method called **SDSTrack** with a symmetrical framework. We use lightweight multimodal adaptation for parameter-efficient fine-tuning to transfer the feature extraction capability of the pre-trained model to other modalities and fuse multimodal features effectively.

• SDSTrack applies a complementary masked patch distillation strategy based on self-distillation learning to improve the robustness and accuracy of tracking results under extreme conditions.

• Extensive experiments on several benchmarks for various modality combinations, such as RGB+Depth, RGB+Thermal, and RGB+Event, show our SDSTrack achieves state-of-the-art performance and impressive tracking results in extreme conditions.

2. Related Works

2.1. Multimodal object tracking

Visual object tracking(VOT) has gained substantial research attention in the past decade and has found applications in various fields such as autonomous driving [12, 21], mobile robotics [49, 50], video surveillance [18, 30, 31], human-computer interaction [60, 70], *etc*. The goal of VOT is to track a specific target object, *i.e.*, given a target bounding box location in the first frame, the tracker must recognize and locate this target in subsequent frames.

In recent years, due to the unstable reliability of RGB-only data in challenging scenarios, more and more studies are exploring the potential of multimodal data to address this issue. For example, CMX [64] and zhang *et al*. [65] have utilized different modalities such as RGB, Lidar, Depth, Event, Thermal, *etc*. for semantic segmentation. Similarly, in visual object tracking, researchers have dedi-

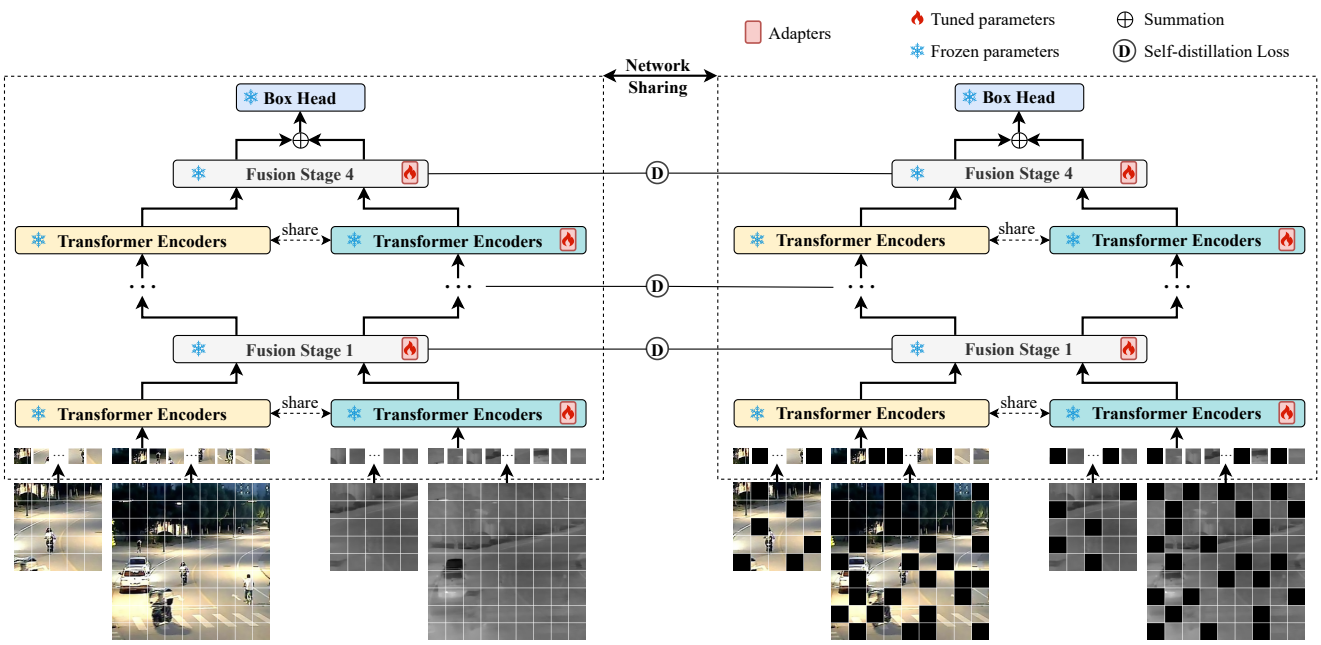

Figure 3. **The Overall pipeline of our SDSTrack**. Firstly, we apply a random masking technique to the RGB-X patch embeddings in a complementary manner, ensuring that at least one modality remains valid. Next, the masked and clean data undergo four feature extraction and fusion stages. These stages have a symmetrical structure comprising ViT blocks and adapters. In each stage, the fused features from both the masked and clean paths undergo self-distillation, which improves the accuracy and robustness of the model. Finally, the RGB and X features are combined and forwarded to the head network to obtain the prediction results.

cated considerable effort to fusing multimodal images. For example, APFNet [54] proposes an attribute-based progressive fusion network that addresses five typical challenges by designing different branches to extract corresponding features and then fusing them. MACFT [41] introduces a fusion method based on a mixed attention mechanism and uses the KL divergence loss function to enforce consistency between modal features. CEUTrack [74] proposes a cross-modal masking modeling strategy to facilitate modal interaction and avoid attention mechanisms from aggregating homogeneous modal information. In addition, some works handle combinations of different modalities using a single model. For example, ProTrack [59] fuses two modalities using a weighted sum, allowing the model to adapt to RGB+depth/thermal/event combinations. However, it lacks in-depth multimodal exploration. Similarly, ViPT [71] proposes a multimodal prompting method. However, its asymmetric design does not fully exploit information from other modalities, leading to a lack of robustness.

2.2. Parameter-efficient Fine-tuning for Vision Models

With the development of large models [15, 43, 62], parameter-efficient fine-tuning, *i.e.*, PEFT, has gained increasing importance. PEFT is initially employed in natural language processing [19, 20, 29, 63] and has recently been widely explored in computer vision. Its core concept is to keep the pre-trained model frozen and only train a few additional parameters. This approach is particularly useful when there is limited data available for fine-tuning. Popular PEFT methods include adapter-based tuning and prompt tuning. Adapter-based tuning [19] involves inserting trainable adapters into pre-trained models. On the other hand, prompt tuning [29] fine-tunes the model by introducing learnable prompt tokens. In the field of multimodal visual object tracking (VOT), where multimodal tracking data is scarce, full fine-tuning is prone to overfitting and restricts the learning of generalized multimodal representations. Therefore, PEFT holds great potential in addressing these issues. However, there is limited research specifically focused on multimodal VOT using PEFT. ProTrack [59] introduces the concept of prompt in multimodal VOT but does not apply it practically. ViPT [71] employs prompt tuning in multimodal VOT but lacks effective knowledge recall from pre-trained models due to modality gap and assigns different importance levels to different modalities, which can lead to dependence on the primary modality. Currently, there is no research in this field based on adapter-based tuning, which is a pluggable knowledge transfer method.

3. Method

In this paper, we propose a novel method called SDSTrack for multimodal visual object tracking. The overall pipeline

is illustrated in Fig. 3. By training lightweight adapters, SDSTrack efficiently transfers the feature extraction capability of pre-trained RGB-based trackers to other modalities and fuses multimodal features effectively. SDSTrack can enhance robustness and accuracy through complementary masked patch distillation, even in extreme conditions. In the following, we use X to refer to modalities other than RGB, which can be Depth, Thermal, Event, *etc*.

3.1. Symmetric Multimodal Adaptation (SMA)

To reduce the number of trainable parameters, mitigate the risk of overfitting, and effectively fuse different modalities, we propose symmetric multimodal adaptation (SMA) for adapting pre-trained RGB-based trackers to multimodal object tracking. Inspired by the parameter-efficient fine-tuning techniques [22, 45, 56], we freeze the pre-trained model and only train the proposed SMA, which comprises cross modal adaptation and multimodal fusion adaptation.

In this paper, we choose the classic one-stream RGB-based model, *e.g*., OSTrack [61], as the pre-trained model, comprising a ViT [11] backbone and a prediction head. Each ViT block consists of Multi-head Self-Attention (MSA), LayerNorm (LN), MultiLayer Perceptron (MLP), and residual connections. The input of SDSTrack consists of a pair of template and search frames, *i.e*., template frames $\mathbf{z}_{\text{image}}^{\text{rgb}}, \mathbf{z}_{\text{image}}^{\text{X}} \in \mathbb{R}^{H_z \times W_z \times 3}$, and search frames $\mathbf{x}_{\text{image}}^{\text{rgb}}, \mathbf{x}_{\text{image}}^{\text{X}} \in \mathbb{R}^{H_x \times W_x \times 3}$, which are then projected into patch embeddings $\hat{\mathbf{z}}_{\text{rgb}}, \hat{\mathbf{z}}_{\text{X}} \in \mathbb{R}^{N_z \times D}$ and $\hat{\mathbf{x}}_{\text{rgb}}, \hat{\mathbf{x}}_{\text{X}} \in \mathbb{R}^{N_x \times D}$ by Patch Embed Layers. Then, the patch embeddings $\hat{\mathbf{z}}_{\text{rgb}}$ and $\hat{\mathbf{x}}_{\text{rgb}}$ are concated as $\mathbf{H}_{\text{rgb}}^{(0)} = [\hat{\mathbf{z}}_{\text{rgb}}; \hat{\mathbf{x}}_{\text{rgb}}] \in \mathbb{R}^{(N_z+N_x)\times D}$, $\hat{\mathbf{z}}_{\text{X}}$ and $\hat{\mathbf{x}}_{\text{X}}$ are concated as $\mathbf{H}_{\text{X}}^{(0)} = [\hat{\mathbf{z}}_{\text{X}}; \hat{\mathbf{x}}_{\text{X}}] \in \mathbb{R}^{(N_z+N_x)\times D}$, and the computation of a ViT block can be formulated as:

$$\mathbf{H}'^{(l)}_{\text{rgb}} = \mathbf{H}_{\text{rgb}}^{(l-1)} + \text{MSA}\left(\text{LN}\left(\mathbf{H}_{\text{rgb}}^{(l-1)}\right)\right) \tag{1}$$

$$\mathbf{H}_{\text{rgb}}^{(l)} = \mathbf{H}'^{(l)}_{\text{rgb}} + \text{MLP}\left(\text{LN}\left(\mathbf{H}'^{(l)}_{\text{rgb}}\right)\right) \tag{2}$$

where $\mathbf{H}_{\text{rgb}}^{(l-1)}$ and $\mathbf{H}_{\text{rgb}}^{(l)}$ represent the output of the $l-1$-th and l-th ViT block, respectively.

3.1.1 Cross Modal Adaptation

Since the pre-trained model we used is trained on RGB data, there exists a modality gap to the X-modal. Therefore, we propose Cross Modal Adaptation (CMA), which transfers the feature extraction capability from the RGB domain to the X domain. The core components of CMA are adapters. As shown in Fig. 4(a), the adapter is a bottleneck architecture, which consists of two fully connected (FC) layers, a GELU [17] activation layer, and a residual connection. The first FC layer (FC Down) projects the input to a lower dimension, and the second FC layer (FC Up) projects it back to the original dimension. As illustrated in Fig. 4(b), we insert adapters into ViT block after MSA and parallel to the MLP, and the calculation process is written by:

$$\mathbf{H}'^{(l)}_{\text{X}} = \mathbf{H}_{\text{X}}^{(l-1)} + \text{Adapter}\left(\text{MSA}\left(\text{LN}\left(\mathbf{H}_{\text{X}}^{(l-1)}\right)\right)\right) \tag{3}$$

$$\begin{aligned}\mathbf{H}_{\text{X}}^{(l)} = \mathbf{H}'^{(l)}_{\text{X}} &+ \text{MLP}\left(\text{LN}\left(\mathbf{H}'^{(l)}_{\text{X}}\right)\right) \\ &+ r \cdot \text{Adapter}\left(\text{LN}\left(\mathbf{H}'^{(l)}_{\text{X}}\right)\right)\end{aligned} \tag{4}$$

where $\mathbf{H}_{\text{X}}^{(l-1)}$ and $\mathbf{H}_{\text{X}}^{(l)}$ are the output of the $l-1$-th and l-th ViT block, r is a scaling factor that regulates the influence of the adapter's output weight.

3.1.2 Multimodal Fusion Adaptation

Previous methods often involve customized fusion modules to achieve multimodal fusion because it is commonly believed that the feature extraction backbone cannot fuse multimodal data. However, introducing the modules can lead to more tunable parameters or limited fusion performance.

To address this problem, we present a new strategy: *reuse part of the pre-trained ViT blocks in the pre-trained model to achieve multimodal fusion*. Vision Transformer (ViT) processes images in a sequential modeling manner, *i.e*., it divides and projects images into a series of tokens, which are then processed through multiple layers of transformer models, capturing contextual dependency relationships among tokens. Inspired by this, we believe multimodal fusion can also be regarded as the interaction among multimodal tokens. Therefore, we propose Multimodal Fusion Adaptation (MFA), whose core components are also adapters. Specifically, we reuse the last ViT block of each stage of encoders and integrate the adapters into the blocks as fusion stages. As shown in Fig. 4(c), we insert adapters into the ViT block after MSA and parallel to the MLP. In each fusion stage, we concatenate RGB features $\mathbf{H}_{\text{rgb}}^{(l)}$ and X-modal features $\mathbf{H}_{\text{X}}^{(l)}$ as the input. Then, we feed them into components like MSA and MLP, which can learn the relationship among multimodal tokens. To avoid overfitting and preserve the modeling capabilities of the pre-trained model, we introduce an attention mask into the MSA at each fusion stage, which sets values of RGB- and X-modal self-attention maps to zero. The computation in the i-th fusion stage can be formulated as:

$$\mathbf{H}'^{(i)} = \overline{\mathbf{H}}^{(i)} + \text{Adapter}\left(\text{MSA}\left(\text{LN}\left(\overline{\mathbf{H}}^{(i)}\right)\right)\right) \tag{5}$$

$$\begin{aligned}\mathbf{H}^{(i)} = \mathbf{H}'^{(i)} &+ \text{MLP}\left(\text{LN}\left(\mathbf{H}'^{(i)}\right)\right) \\ &+ r \cdot \text{Adapter}\left(\text{LN}\left(\mathbf{H}'^{(i)}\right)\right)\end{aligned} \tag{6}$$

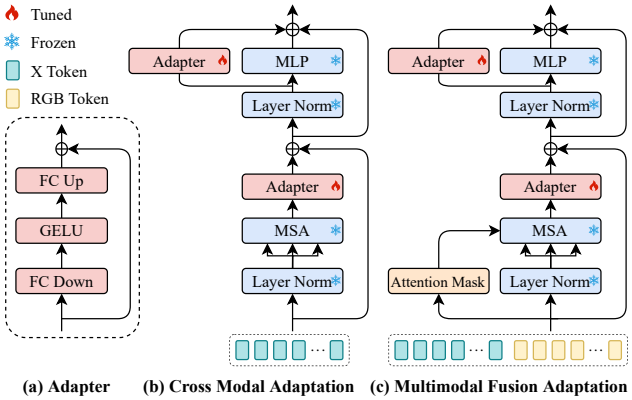

Figure 4. **The components of symmetric multimodal adaptation (SMA)**. (a) The structure of the adapter. (b) Cross Modal Adaptation for X-modal image feature extraction (c) Multimodal Fusion Adaptation for multimodal image feature fusion.

where $\overline{\mathbf{H}}^{(i)}$ is the concatenated multimodal features to be input into i-th fusion stage. In MSA, the attention mask is:

$$\mathbf{M} = \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \tag{7}$$

Hence, the attention map $\mathbf{A}$ after the softmax activation layer can be expressed as follows:

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{A}_{\mathrm{RGB,X}} \\ \mathbf{A}_{\mathrm{X,RGB}} & \mathbf{0} \end{bmatrix} \tag{8}$$

The attention mask not only prevents focusing on specific modality tokens, thus effectively facilitating multimodal interaction, but also maintains the feature extraction capability of the original pre-trained model without disturbance.

3.2. Complementary Masked Patch Distillation

As shown in Fig. 2, many multimodal trackers [71] rely on specific modalities and perform poorly in extreme scenarios. Additionally, these trackers fail to adequately explore the complementary information between modalities. To address these limitations, we propose a method called Complementary Masked Patch Distillation. This method aims to explore the nature of multimodal data and enhance the robustness of trackers. It involves random complementary patch mask strategy and self-distillation learning.

3.2.1 Random Complementary Patch Mask (RCPM)

Recently, masking strategies have found widespread application in natural language processing [10, 39] and visual applications [16, 44, 52, 55]. In particular, the use of image masking strategies has become prevalent in pre-training large-capacity backbone models, allowing them to acquire generic representations for downstream tasks. One of the key principles behind masking strategies is to enable the model to reconstruct the semantic information of masked patches, thereby enhancing its understanding of the underlying data. Inspired by this, we employ a masking strategy to explore the complementary information between modalities, facilitating multimodal image fusion and enabling the model to avoid over-dependency on a specific modality, thus improving robustness in extreme scenarios. Specifically, we first obtain clean multimodal patch embeddings through image projection. Then, we randomly mask the patch embeddings of one modality, *i.e.*, assigning their values to zero. Correspondingly, we perform random masking in the unmasked portions in the other modality. On one hand, this strategy ensures the validity of at least one modality, ensuring the accuracy of trackers. On the other hand, it enhances the model's ability to handle extreme situations, thereby improving the robustness of trackers.

3.2.2 Self-Distillation Learning (SD)

Merely obtaining masked embeddings through the previous RCPM can be seen as a form of data augmentation, but it is insufficient to significantly enhance the model's robustness. We propose a self-distillation learning (SD) strategy to address this limitation. The main idea behind self-distillation learning is to leverage the clean features, which ensure the model's accuracy, to supervise and guide the masked features, thereby improving the model's robustness. This strategy aims to teach the model to extract complementary information from multiple modalities by referring to the clean features and adapt to extreme scenarios where modalities may be occluded or even missing. During the training phase, we obtain both clean and masked embeddings using RCPM, resulting in two paths: the clean path and the masked path. These two paths share all the network parameters. At each fusion stage, we calculate the similarity between the fusion features from these two paths and express it using the mean squared error. This similarity calculation is used as the self-distillation loss, which encourages the model to learn to extract useful information from both the clean and masked features. The following formula expresses the self-distillation loss:

$$L_{\mathrm{SD}} = \frac{1}{s}\sum_{i=1}^{s} \mathrm{MSE}\left(\mathbf{H}^{(i)}, \widetilde{\mathbf{H}}^{(i)}\right) \tag{9}$$

where $\mathbf{H}^{(i)}$ and $\widetilde{\mathbf{H}}^{(i)}$ mean the fusion features of clean data and masked data in the i-th fusion stage, respectively, and MSE is the mean squared error.

3.3. Prediction Head and Supervised Loss

We utilize and freeze the same head as OSTrack [61], a full convolutional network(FCN) consisting of L stacked Conv-BN-ReLU layers. To avoid dependence on specific modality, we take the summation of the RGB- and X-search features as the input for the head, which is different from other

	CA3DMS [38]	TSDM [69]	DAL [42]	ATCAIS [26]	DDiMP [26]	DeT [58]	OSTrack [61]	SPT [72]	ProTrack [59]	ViPT [71]	**SDSTrack (Ours)**
Pr(↑)	0.218	0.393	0.512	0.500	0.503	0.560	0.536	0.527	0.583	0.592	**0.619**
Re(↑)	0.228	0.376	0.369	0.455	0.469	0.506	0.522	0.549	0.573	0.596	**0.609**
F-score(↑)	0.223	0.384	0.429	0.476	0.485	0.532	0.529	0.538	0.578	0.594	**0.614**

Table 1. **Overall performance** on the DepthTrack [58] test set.

	ATOM [9]	DRefine [27]	DMTracker [28]	DeT [58]	OSTrack [61]	SPT [72]	SBT_RGBD [28]	ProTrack [59]	ViPT [71]	**SDSTrack (Ours)**
EAO(↑)	0.505	0.592	0.658	0.657	0.676	0.651	0.708	0.651	0.721	**0.728**
Accuracy(↑)	0.698	0.775	0.758	0.760	0.803	0.798	0.809	0.801	**0.815**	0.812
Robustness(↑)	0.688	0.760	0.851	0.845	0.833	0.851	0.864	0.802	0.871	**0.883**

Table 2. **Overall performance** on the VOT-RGBD2022 [28] dataset.

	mfDiMP [66]	SGT [32]	DAFNet [14]	FANet [73]	CAT [34]	JMMAC [67]	CMPP [48]	APFNet [54]	ProTrack [59]	ViPT [71]	**SDSTrack (Ours)**
MPR(↑)	0.646	0.720	0.796	0.787	0.804	0.790	0.823	0.827	0.795	0.835	**0.848**
MSR(↑)	0.428	0.472	0.544	0.553	0.561	0.573	0.575	0.579	0.599	0.617	**0.625**

Table 3. **Overall performance** on the RGBT234 [33] dataset.

methods [59, 61, 71] that only use RGB information. During training, SDSTrack employs the same loss function as OSTrack [61] to supervise clean and masked data tracking. The loss functions are represented as follows:

$$L_{\text{CLEAN}} = L_{\text{cls}} + \lambda_{\text{iou}} L_{\text{iou}} + \lambda_{L_1} L_1 \tag{10}$$

$$L_{\text{MASK}} = L_{\text{cls}} + \lambda_{\text{iou}} L_{\text{iou}} + \lambda_{L_1} L_1 \tag{11}$$

where $\lambda_{iou} = 2$ and $\lambda_{L_1} = 5$ are the regularization parameters as in [57]. Finally, the overall loss function is:

$$L_{\text{track}} = L_{\text{CLEAN}} + \alpha L_{\text{MASK}} + \beta L_{\text{SD}} \tag{12}$$

where $\alpha = 0.3$ and $\beta = 0.25$ are hyperparameters for balancing different loss functions.

4. Experiments

In this section, we first describe the experiment implementation details of the proposed SDSTrack. We then compare SDSTrack with other state-of-the-art methods on multiple benchmark datasets. Our tracker is evaluated on Depthtrack [58] and VOT22RGBD [28] for RGB-D tracking, LasHeR [35] and RGBT234 [33] for RGB-T tracking, and VisEvent [51] for RGB-E tracking. Lastly, we conduct robustness performance and ablation studies.

4.1. Implementation Details

Training. Our proposed SDSTrack can apply to various modality combinations. We utilize the training sets of DepthTrack [58] for RGB-D tracking, LasHeR [35] for RGB-T tracking, and VisEvent [51] for RGB-E tracking. The models are trained on 4 NVIDIA 3090Ti GPUs with a global batch size of 64. We utilize the AdamW [40] optimizer, with a weight decay set to 10^{-4}. The learning rate is set to 3×10^{-5} for RGB-D and 6×10^{-5} for RGB-T and RGB-E. The models are trained for 15, 40, and 50 epochs for RGB-D, RGB-T, and RGB-E, respectively. Each epoch involves sampling 60k samples.

Inference. During inference, the data follows the path of the clean data without involving the RCPM module. The tracking speed, tested on an NVIDIA 3090Ti GPU, is approximately 20.86 frames per second (*fps*).

4.2. Comparison with State-of-the-arts

DepthTrack. DepthTrack [58] is a comprehensive RGB-D tracking benchmark that addresses numerous challenges. It consists of 150 sequences for training and 50 sequences for testing. The evaluation metrics include precision (Pr), recall (Re), and F-score, which is calculated as $F = \frac{2Re \times Pr}{Re + Pr}$. As shown in Tab. 1, our SDSTrack achieves new state-of-the-art performance, surpassing previous SOTA 2.7% in precision, 1.3% in recall, and 2.0% in F-score.

VOT-RGBD2022. VOT-RGBD2022 [28] is the most recent dataset in RGB-D tracking, comprising 127 short-term RGB-D sequences. The performance metric includes Accuracy, Robustness, and Expected Average Overlap (EAO). As illustrated in Tab. 2, our proposed SDSTrack obtains improved robustness while maintaining good precision, with a 1.2% improvement in Robustness compared to the previous state-of-the-art performance.

RGBT234. RGBT234 [33] is a large-scale RGB-T tracking dataset that contains 234 pairs of videos. MSR and MPR are adopted for performance evaluation. As shown in Tab. 3, our SDSTrack obtains new SOTA 84.8% in MPR and 62.5% in MSR.

LasHeR. LasHeR [35] is a large-scale RGB-T dataset that comprises 979 video pairs for training and 245 pairs for testing. As shown in Fig. 5, our SDStrack surpasses all previous SOTA trackers, obtaining the top performance

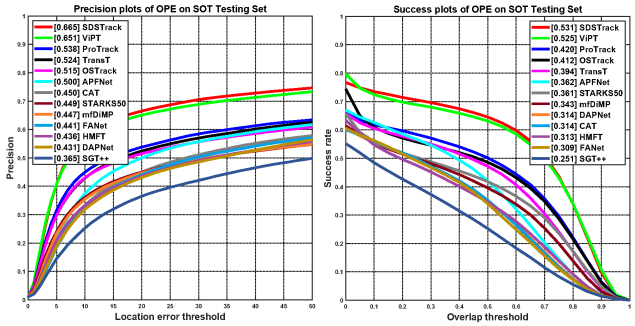

Figure 5. **Overall performance** on the LasHeR [35] test set.

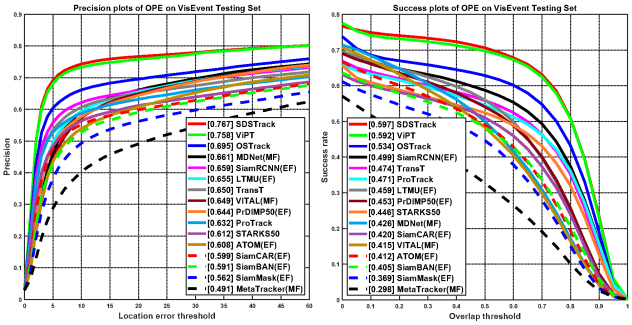

Figure 6. **Overall performance** on the VisEvent [51] test set.

of 66.5% and 53.1% in precision and success, respectively, where precision exceeds the previous best result by 1.4%.

VisEvent. VisEvent [51] is the largest dataset for RGB-E tracking. It contains 500 pairs of videos for training and 320 pairs of videos for testing. We compare our proposed SDSTrack with existing state-of-the-art RGB-E trackers. As reported in Fig. 6, our SDSTrack obtains new state-of-the-art precision and success of 76.7% and 59.7%, respectively.

4.3. Exploration Study and Analysis

Robustness performance. To analyze the tracking performance of our method in extreme scenarios, we conducted experiments involving different scenarios, namely: (1) Dropping RGB frames with a 50% probability. (2) Dropping X-modal frames with a 50% probability. (3) Randomly occluding multimodal images by setting certain pixels to pure black. As shown in Tab. 4, our SDSTrack demonstrates superior robustness compared to ViPT [71], particularly in scenarios where the RGB modality is lost. Specifically, when RGB drops, SDSTrack achieves significant improvements of 3.3% in F-score for RGB-D tracking, 23.2% and 16.4% in precision and success for RGB-T tracking, 11.6% and 7.9% in precision and success for RGB-E tracking. Therefore, SDSTrack reduces reliance on specific modalities to a certain extent, making it effectively applicable to more challenging scenarios.

challenge	tracker	DepthTrack [58]			LasHeR [35]		VisEvent [51]	
		Pr	Re	F-score	Pre	Suc	Pre	Suc
w/o RGB	ViPT [71]	0.361	0.318	0.338	0.306	0.268	0.473	0.348
	Ours	**0.396**	**0.340**	**0.365**	**0.538**	**0.432**	**0.589**	**0.427**
w/o X	ViPT [71]	0.572	0.570	0.571	**0.553**	**0.451**	0.723	0.562
	Ours	**0.587**	**0.577**	**0.582**	0.552	0.448	**0.741**	**0.574**
occlusion	ViPT [71]	0.494	0.473	0.483	0.414	0.340	0.627	0.439
	Ours	**0.514**	**0.497**	**0.505**	**0.545**	**0.421**	**0.639**	**0.448**

Table 4. **The performance comparison in challenging conditions** on DepthTrack [58], LasHeR [35] and VisEvent [51].

Furthermore, we perform analysis of various challenging attributes, such as illumination variation, motion blur, out-of-view, *etc*. As shown in Fig. 7, our SDSTrack also achieves the best tracking performance in these extreme

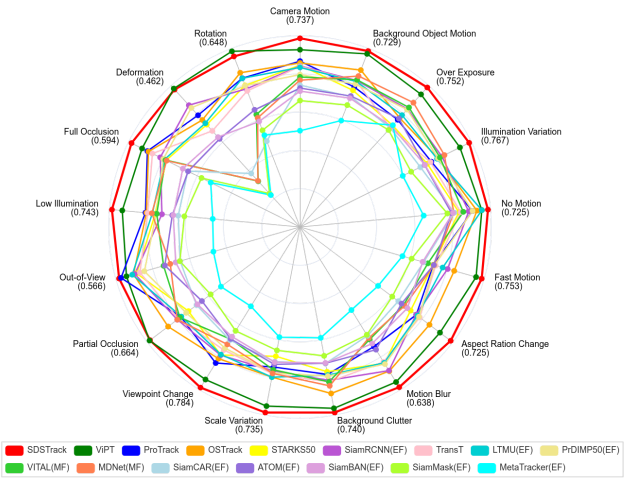

Figure 7. **Precision scores of different attributes** on the VisEvent [51] test set.

scenarios, demonstrating improved robustness. Please refer to the supplementary materials for more details.

Effectiveness of proposed components. We analyze the effect of the proposed symmetric multimodal adaptation (SMA), random complementary patch mask (RCPM), and self-distillation learning (SD). Our baseline is a symmetrical structure without the mask path and SMA. It reuses the 3rd, 6th, 9th, and 11th ViT blocks as the fusion stages. The RGB and X features of the last fusion stage are added for feeding into the prediction head, where only the LN before the head and Patch Embed Layers are tuned. As shown in Tab. 5, the baseline equipped with SMA (Variant 1) achieves significant improvement, surpassing the baseline by 1.9% in F-score for DepthTrack, 4.8% and 3.0% in precision and success for RGBT234, and 3.2% and 3.3% in precision and success for VisEvent. Further adding the RCPM (Variant 2) is also effective, especially improving the F-score by 1.9% for DepthTrack. Finally, SD (Variant 3) improves 2.0% in F-score for DepthTrack, 0.6% and 1.0% in precision and success for RGBT234, and 0.2% in precision and success for VisEvent.

Effectiveness of the Adaptation. To enable a compre-

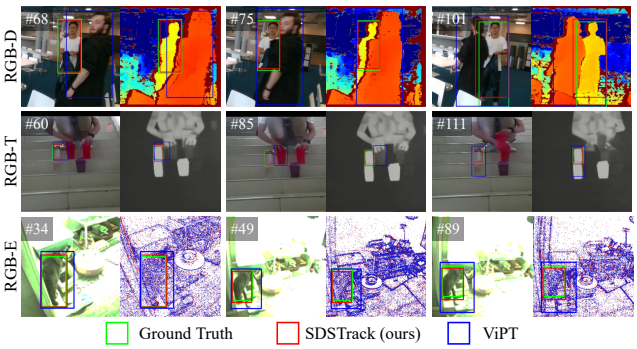

Figure 8. **Visualization results** of RGB-D, RGB-T and RGB-E.

Variants	SMA	RCPM	SD	DepthTrack [58]			RGBT234 [33]		VisEvent [51]	
				Pr	Re	F-score	Pre	Suc	Pre	Suc
baseline				0.553	0.556	0.554	0.799	0.592	0.728	0.562
1	✓			0.578	0.569	0.573	0.847	0.622	0.760	0.595
2	✓	✓		0.599	0.588	0.594	0.842	0.615	0.765	0.595
3	✓	✓	✓	0.619	0.609	0.614	0.848	0.625	0.767	0.597

Table 5. **Ablation studies** on the effect of the components. SMA, RCPM, and SD denote symmetric multimodal adaptation, random complementary patch mask, and self-distillation learning.

hensive comparison with our adaptation, we employed two training methods, freezing and fully fine-tuning (FFT), on the baseline model. The results of these comparisons are presented in Tab. 6. Remarkably, the baseline equipped with CMA surpasses the frozen- and FFT- methods by introducing only a few training parameters. Furthermore, we observed that only introducing MFA also improves performance significantly. Using both CMA and MFA contributes to the best performance.

Method	Model Param (M)	Tuned Param (M)	DepthTrack [58]			LasHeR [35]		VisEvent [51]	
			Pr	Re	F-score	Pre	Suc	Pre	Suc
Frozen	93.60	0.59	0.524	0.521	0.523	0.492	0.398	0.709	0.541
FFT	93.60	93.60	0.570	0.545	0.558	0.625	0.496	0.729	0.561
CMA	100.70	7.69	0.569	0.558	0.563	0.632	0.506	0.746	0.578
MFA	100.70	7.69	0.618	0.599	0.608	0.650	0.520	0.756	0.590
CMA + MFA	107.80	14.79	0.619	0.609	0.614	0.665	0.531	0.767	0.597

Table 6. **Ablation studies on the effect of the Adaptation**. CMA and MFA denote cross modal adaptation and multimodal fusion adaptation in Symmetric Multimodal Adaptation (SMA).

Effectiveness of components on robustness. We analyze the components' effectiveness on the model's robustness, and the results are presented in Tab. 7. We find that SMA effectively enhances the robustness of the model while using RCPM alone provides a limited improvement, which can be considered a form of data augmentation. However, when RCPM is combined with SD, the model's robustness is significantly improved.

Comparison on inference speed. We compare the inference speed of SDSTrack with previous methods. As shown in Tab. 8, our method enables real-time tracking (20.86 *fps*) while achieving SOTA accuracy and robustness.

Variants	SMA	RCPM	SD	w/o RGB		w/o X		occlusion	
				Pre	Suc	Pre	Suc	Pre	Suc
0	×	×	×	0.298	0.251	0.313	0.269	0.435	0.357
1	✓	×	×	0.472	0.389	0.525	0.410	0.512	0.399
2	✓	✓	×	0.481	0.390	0.529	0.423	0.523	0.415
3	✓	✓	✓	0.538	0.432	0.552	0.448	0.545	0.421

Table 7. **Ablation studies on the effect of various components on robustness** on the LasHeR [35] testing set. SMA, RCPM, and SD denote symmetric multimodal adaptation, random complementary patch mask, and self-distillation learning.

Method	DepthTrack [58]			LasHeR [35]		VisEvent [51]		FPS
	Pr	Re	F-score	Pre	Suc	Pre	Suc	
CLGS_D [26]	0.584	0.369	0.453	-	-	-	-	7.27
DDiMP [26]	0.503	0.469	0.485	-	-	-	-	4.77
SPT [72]	0.527	0.549	0.538	-	-	-	-	**25**
CAT [34]	-	-	-	0.450	0.314	-	-	20
DAFNet [14]	-	-	-	0.620	0.458	-	-	21
SiamRCNN(EF) [47]	-	-	-	-	-	0.661	0.499	4.7
LTMU(EF) [8]	-	-	-	-	-	0.655	0.459	13
ViPT [71]	0.592	0.596	0.594	0.651	0.525	0.758	0.592	24.78
Ours	**0.619**	**0.609**	**0.614**	**0.665**	**0.531**	**0.767**	**0.597**	20.86

Table 8. **Comparison of inference speed**. The FPS of previous methods is obtained from respective or related papers, except for ViPT [71] tested in an environment identical to our SDSTrack.

Visualization results. We visualize multimodal tracking in Fig. 8. The results show that SDSTrack fully utilizes multimodal information to handle challenging situations, enabling more accurate and robust tracking.

5. Conclusion

This paper introduces a novel symmetric approach called SDSTrack for multimodal visual object tracking. This approach aims to achieve parameter-efficient fine-tuning on limited multimodal data and enhance the model's robustness in extreme scenarios. Specifically, we utilize lightweight adapter-based tuning to symmetrically fine-tune the pre-trained RGB-based tracker, transitioning the feature extraction capability from the RGB to the X domain and achieving multimodal fusion effectively. Furthermore, we devise a random complementary patch mask strategy to obtain masked data. The masked and clean data flows are self-distilled to enhance the model's robustness in extreme scenarios. Our SDSTrack's effectiveness is demonstrated through extensive experiments on multiple benchmarks.

Limitations. Although improving accuracy and robustness in extreme scenarios, SDSTrack introduces more computational complexity during training, primarily arising from self-distillation learning. This problem will be a point we need to explore further in the future.

Acknowledgment. This work was supported by NSFC 62088101 Autonomous Intelligent Unmanned Systems, and Zhejiang University Advanced Perception on Robotics and Intelligent Learning Lab.

References

[1] Reza Azad, Nika Khosravi, Mohammad Dehghanmanshadi, Julien Cohen-Adad, and Dorit Merhof. Medical image segmentation on mri images with missing modalities: A review. *arXiv preprint arXiv:2203.06217*, 2022. 1

[2] Lizhi Bai, Jun Yang, Chunqi Tian, Yaoru Sun, Maoyu Mao, Yanjun Xu, and Weirong Xu. Dcanet: Differential convolution attention network for rgb-d semantic segmentation. *arXiv preprint arXiv:2210.06747*, 2022.

[3] Francesco Barbato, Elena Camuffo, Simone Milani, and Pietro Zanuttigh. Continual road-scene semantic segmentation via feature-aligned symmetric multi-modal network. *arXiv preprint arXiv:2308.04702*, 2023. 1

[4] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6182–6191, 2019. 1, 2

[5] Boyu Chen, Peixia Li, Lei Bai, Lei Qiao, Qiuhong Shen, Bo Li, Weihao Gan, Wei Wu, and Wanli Ouyang. Backbone is all your need: A simplified architecture for visual object tracking. In *European Conference on Computer Vision*, pages 375–392. Springer, 2022.

[6] Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. Seqtrack: Sequence to sequence learning for visual object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14572–14581, 2023.

[7] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13608–13618, 2022. 1

[8] Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. High-performance long-term tracking with meta-updater. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6298–6307, 2020. 8

[9] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4660–4669, 2019. 1, 6

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[12] Zheng Fang, Sifan Zhou, Yubo Cui, and Sebastian Scherer. 3d-siamrpn: An end-to-end learning method for real-time 3d single object tracking using raw point cloud. *IEEE Sensors Journal*, 21(4):4995–5011, 2020. 2

[13] Shenyuan Gao, Chunluan Zhou, Chao Ma, Xinggang Wang, and Junsong Yuan. Aiatrack: Attention in attention for transformer visual tracking. In *European Conference on Computer Vision*, pages 146–164. Springer, 2022. 1

[14] Yuan Gao, Chenglong Li, Yabin Zhu, Jin Tang, Tao He, and Futian Wang. Deep adaptive fusion network for high performance rgbt tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 6, 8

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 5

[17] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[18] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):583–596, 2014. 2

[19] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 3

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[21] Le Hui, Lingpeng Wang, Mingmei Cheng, Jin Xie, and Jian Yang. 3d siamese voxel-to-bev tracker for sparse point clouds. *Advances in Neural Information Processing Systems*, 34:28714–28727, 2021. 2

[22] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 4

[23] Zhengkai Jiang, Peng Gao, Chaoxu Guo, Qian Zhang, Shiming Xiang, and Chunhong Pan. Video object detection with locally-weighted deformable neighbors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8529–8536, 2019. 1

[24] Zhengkai Jiang, Yu Liu, Ceyuan Yang, Jihao Liu, Peng Gao, Qian Zhang, Shiming Xiang, and Chunhong Pan. Learning where to focus for efficient video object detection. In *European Conference on Computer Vision*, pages 18–34. Springer, 2020. 1

[25] Zhengkai Jiang, Yuxi Li, Ceyuan Yang, Peng Gao, Yabiao Wang, Ying Tai, and Chengjie Wang. Prototypical contrast adaptation for domain adaptive semantic segmentation. In *European Conference on Computer Vision*, pages 36–54. Springer, 2022. 1

[26] Matej Kristan, Aleš Leonardis, Jiří Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Martin Danelljan, Luka Čehovin Zajc, Alan Lukežič, Ondrej 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*, pages 547–601. Springer, 2020. 6, 8

[27] Matej Kristan, Jiří Matas, Aleš Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Hyung Jin Chang, Martin Danelljan, Luka Cehovin, Alan Lukežič, et al. The ninth visual object tracking vot2021 challenge results. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2711–2738, 2021. 6

[28] Matej Kristan, Aleš Leonardis, Jiří Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Hyung Jin Chang, Martin Danelljan, Luka Čehovin Zajc, Alan Lukežič, et al. The tenth visual object tracking vot2022 challenge results. In *European Conference on Computer Vision*, pages 431–460. Springer, 2022. 6

[29] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 3

[30] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8971–8980, 2018. 2

[31] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4282–4291, 2019. 2

[32] Chenglong Li, Nan Zhao, Yijuan Lu, Chengli Zhu, and Jin Tang. Weighted sparse representation regularized graph learning for rgb-t object tracking. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1856–1864, 2017. 6

[33] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: Benchmark and baseline. *Pattern Recognition*, 96:106977, 2019. 6, 8

[34] Chenglong Li, Lei Liu, Andong Lu, Qing Ji, and Jin Tang. Challenge-aware rgbt tracking. In *European Conference on Computer Vision*, pages 222–237. Springer, 2020. 6, 8

[35] Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, Jin Tang, and Dengdi Sun. Lasher: A large-scale high-diversity benchmark for rgbt tracking. *IEEE Transactions on Image Processing*, 31:392–404, 2021. 6, 7, 8

[36] L Lin, H Fan, Y Xu, and H Ling. Swintrack: A simple and strong baseline for transformer tracking. arxiv 2021. *arXiv preprint arXiv:2112.00995*, 2021. 1

[37] Hong Liu, Dong Wei, Donghuan Lu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. M3ae: Multimodal representation learning for brain tumor segmentation with missing modalities. *arXiv preprint arXiv:2303.05302*, 2023. 1

[38] Ye Liu, Xiao-Yuan Jing, Jianhui Nie, Hao Gao, Jun Liu, and Guo-Ping Jiang. Context-aware three-dimensional mean-shift with occlusion handling for robust object tracking in rgb-d videos. *IEEE Transactions on Multimedia*, 21(3):664–677, 2018. 6

[39] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 5

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[41] Yang Luo, Mingtao Dong, Xiqing Guo, and Jin Yu. Rgb-t tracking based on mixed attention. *arXiv preprint arXiv:2304.04264*, 2023. 2, 3

[42] Yanlin Qian, Song Yan, Alan Lukežič, Matej Kristan, Joni-Kristian Kämäräinen, and Jiří Matas. Dal: A deep depth-aware long-term tracker. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7825–7832. IEEE, 2021. 6

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[44] Ukcheol Shin, Kyunghyun Lee, and In So Kweon. Complementary random masking for rgb-thermal semantic segmentation. *arXiv preprint arXiv:2303.17386*, 2023. 1, 5

[45] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022. 4

[46] Zhengzheng Tu, Qishun Wang, Hongshun Wang, Kunpeng Wang, and Chenglong Li. Erasure-based interaction network for rgbt video object detection and a unified benchmark. *arXiv preprint arXiv:2308.01630*, 2023. 1

[47] Paul Voigtlaender, Jonathon Luiten, Philip HS Torr, and Bastian Leibe. Siam r-cnn: Visual tracking by re-detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6578–6588, 2020. 8

[48] Chaoqun Wang, Chunyan Xu, Zhen Cui, Ling Zhou, Tong Zhang, Xiaoya Zhang, and Jian Yang. Cross-modal pattern-propagation for rgb-t tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7064–7073, 2020. 6

[49] Mengmeng Wang, Daobilige Su, Lei Shi, Yong Liu, and Jaime Valls Miro. Real-time 3d human tracking for mobile robots with multisensors. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5081–5087. IEEE, 2017. 2

[50] Mengmeng Wang, Yong Liu, Daobilige Su, Yufan Liao, Lei Shi, Jinhong Xu, and Jaime Valls Miro. Accurate and real-time 3-d tracking for the following robots by fusing vision and ultrasonar information. *IEEE/ASME Transactions On Mechatronics*, 23(3):997–1006, 2018. 2

[51] Xiao Wang, Jianing Li, Lin Zhu, Zhipeng Zhang, Zhe Chen, Xin Li, Yaowei Wang, Yonghong Tian, and Feng Wu. Visevent: Reliable object tracking via collaboration of frame and

event flows. *IEEE Transactions on Cybernetics*, 2023. 6, 7, 8

[52] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 5

[53] Xing Wei, Yifan Bai, Yongchao Zheng, Dahu Shi, and Yihong Gong. Autoregressive visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9697–9706, 2023. 1

[54] Yun Xiao, Mengmeng Yang, Chenglong Li, Lei Liu, and Jin Tang. Attribute-based progressive fusion network for rgbt tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2831–2838, 2022. 3, 6

[55] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 5

[56] Jiazheng Xing, Mengmeng Wang, Xiaojun Hou, Guang Dai, Jingdong Wang, and Yong Liu. Multimodal adaptation of clip for few-shot action recognition. *arXiv preprint arXiv:2308.01532*, 2023. 4

[57] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10448–10457, 2021. 1, 2, 6

[58] Song Yan, Jinyu Yang, Jani Käpylä, Feng Zheng, Aleš Leonardis, and Joni-Kristian Kämäräinen. Depthtrack: Unveiling the power of rgbd tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10725–10733, 2021. 1, 2, 6, 7, 8

[59] Jinyu Yang, Zhe Li, Feng Zheng, Ales Leonardis, and Jingkuan Song. Prompting for multi-modal tracking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3492–3500, 2022. 2, 3, 6

[60] Zhengyuan Yang, Tushar Kumar, Tianlang Chen, Jingsong Su, and Jiebo Luo. Grounding-tracking-integration. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(9):3433–3443, 2020. 2

[61] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European Conference on Computer Vision*, pages 341–357. Springer, 2022. 1, 2, 4, 5, 6

[62] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 3

[63] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021. 3

[64] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 2

[65] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1147, 2023. 2

[66] Lichao Zhang, Martin Danelljan, Abel Gonzalez-Garcia, Joost Van De Weijer, and Fahad Shahbaz Khan. Multi-modal fusion for end-to-end rgb-t tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 6

[67] Pengyu Zhang, Jie Zhao, Chunjuan Bo, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Jointly modeling motion and appearance cues for robust rgb-t tracking. *IEEE Transactions on Image Processing*, 30:3335–3347, 2021. 6

[68] Yang Zhang, Yang Yang, Chenyun Xiong, Guodong Sun, and Yanwen Guo. Attention-based dual supervised decoder for rgbd semantic segmentation. *arXiv preprint arXiv:2201.01427*, 2022. 1

[69] Pengyao Zhao, Quanli Liu, Wei Wang, and Qiang Guo. Tsdm: Tracking by siamrpn++ with a depth-refiner and a mask-generator. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 670–676. IEEE, 2021. 6

[70] Li Zhou, Zikun Zhou, Kaige Mao, and Zhenyu He. Joint visual grounding and tracking with natural language specification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23151–23160, 2023. 2

[71] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9516–9526, 2023. 1, 2, 3, 5, 6, 7, 8

[72] Xue-Feng Zhu, Tianyang Xu, Zhangyong Tang, Zucheng Wu, Haodong Liu, Xiao Yang, Xiao-Jun Wu, and Josef Kittler. Rgbd1k: A large-scale dataset and benchmark for rgb-d object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3870–3878, 2023. 2, 6, 8

[73] Yabin Zhu, Chenglong Li, Jin Tang, and Bin Luo. Quality-aware feature aggregation network for robust rgbt tracking. *IEEE Transactions on Intelligent Vehicles*, 6(1):121–130, 2020. 6

[74] Zhiyu Zhu, Junhui Hou, and Dapeng Oliver Wu. Cross-modal orthogonal high-rank augmentation for rgb-event transformer-trackers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22045–22055, 2023. 2, 3

[75] Zhuangwei Zhuang, Rong Li, Kui Jia, Qicheng Wang, Yuanqing Li, and Mingkui Tan. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16280–16290, 2021. 1