OneTracker: Unifying Visual Object Tracking with Foundation Models and Efficient Tuning

Lingyi Hong[1*] Shilin Yan[1*] Renrui Zhang[2,3] Wanyun Li[1] Xinyu Zhou[1] Pinxue Guo[4]
Kaixun Jiang[4] Yiting Chen[1] Jinglun Li[4] Zhaoyu Chen[4] Wenqiang Zhang[1,5†]
[1] Shanghai Key Lab of Intelligent Information Processing,
School of Computer Science, Fudan University, Shanghai, China
[2] CUHK MMLab, [3] Shanghai Artificial Intelligence Laboratory
[4] Shanghai Engineering Research Center of AI & Robotics,
Academy for Engineering & Technology, Fudan University, Shanghai, China
[5] Engineering Research Center of AI & Robotics, Ministry of Education,
Academy for Engineering & Technology, Fudan University, Shanghai, China

Abstract

Visual object tracking aims to localize the target object of each frame based on its initial appearance in the first frame. Depending on the input modility, tracking tasks can be divided into RGB tracking and RGB+X (e.g. RGB+N, and RGB+D) tracking. Despite the different input modalities, the core aspect of tracking is the temporal matching. Based on this common ground, we present a general framework to unify various tracking tasks, termed as ***OneTracker****. OneTracker first performs a large-scale pre-training on a RGB tracker called Foundation Tracker. This pretraining phase equips the Foundation Tracker with a stable ability to estimate the location of the target object. Then we regard other modality information as prompt and build Prompt Tracker upon Foundation Tracker. Through freezing the Foundation Tracker and only adjusting some additional trainable parameters, Prompt Tracker inhibits the strong localization ability from Foundation Tracker and achieves parameter-efficient finetuning on downstream RGB+X tracking tasks. To evaluate the effectiveness of our general framework OneTracker, which is consisted of Foundation Tracker and Prompt Tracker, we conduct extensive experiments on 6 popular tracking tasks across 11 benchmarks and our OneTracker outperforms other models and achieves state-of-the-art performance.*

1. Introduction

Object tracking [4, 13, 56, 91, 119, 120] is a foundation visual task, that involves localizing a target object in each video frame based on the initial bounding box in the first frame. It has various applications, such as self-driving [10, 36, 117], visual surveillance [81, 94], and video compression [46]. In addition to the conventional RGB tracking (Figure 1 (a)), there are various downstream tracking tasks that incorporate additional information and boost performance, including RGB+N, RGB+M, and RGB+D/T/E tracking. In RGB+N tracking [28, 62, 89, 104], the natural linguistic descriptions of target are additionally provided to exclude the interference of similar objects and enhance the localization. In RGB+M tracking [24, 41, 72, 95, 103], the masks of the target in the first frame are offered instead of bounding boxes. In RGB+D/T/E tracking, the depth, thermal, and event maps are utilized as an extra to handle with the vulnerability of RGB trackers to complex scenarios and improve the robustness. The goal of all these downstream tasks is to localize the target with the assistance of multimodal information. Thus, we unify them as a whole, terming them as RGB+X tracking (Figure 1 (b)). Despite the diversity of tracking tasks, the core objective remains the same: localizing the target in the search frame given its initial appearance, similar to the underlying principles of human attention mechanisms [8]. Cognitive scientists have discovered that the human vision system builds the correspondence or motion [2] on the temporal dimension [48] to determine the object's position in the current frame [27], regardless of the form of additional modalities in various tracking tasks.

Currently, there is a prevailing trend where models are designed and trained for specific tasks using data from certain domains, offering convenience and yielding competitive results on individual tasks. However, this design philosophy presents certain challenges. (1) Independent mod-

*Equal Contribution
†Corresponding Author

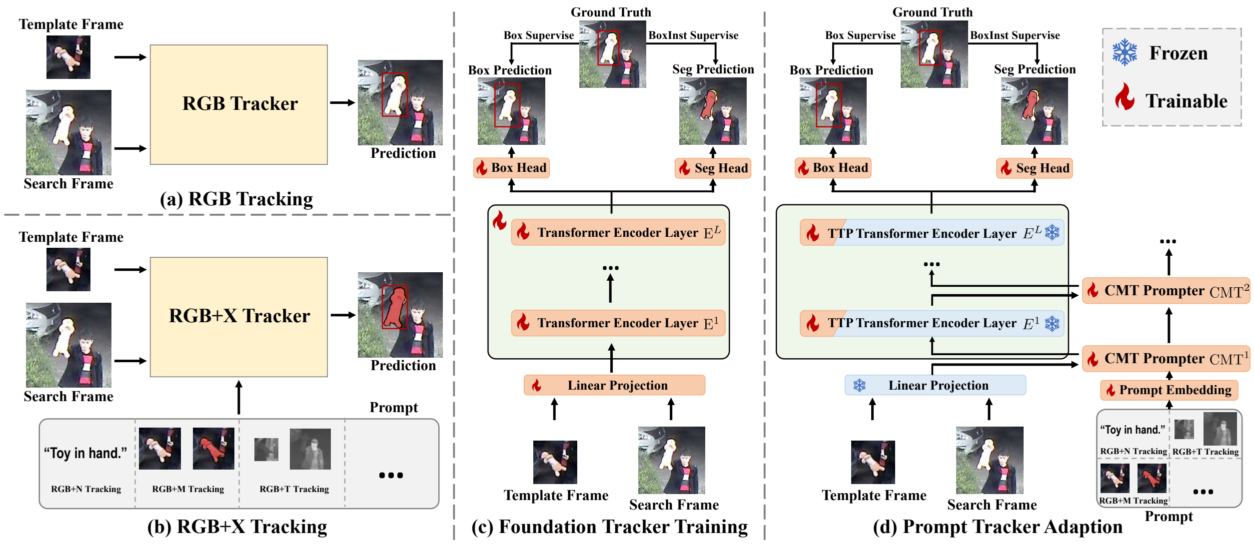


Figure 1. **(a) The definition of RGB tracking. (b) The definition of RGB+X tracking. (c) Overview of Foundation Tracker training. (d) The parameter-efficient finetuning of Prompt Tracker.**

els require customized architectures, resulting in complex training procedures and redundant parameters. (2) For certain tracking tasks, the limited availability of large-scale data severely restricts performance potential. (3) The separate design approach falls short of accurately simulating human attention mechanisms, which are crucial in tracking. Although some previous works [1, 66, 85, 90, 99, 100] have made attempts to address these problems in a unified model, they still exhibit certain limitations. [1, 100] are not specifically designed for tracking tasks, resulting in sub-optimal performance on tracking benchmarks. [66, 85, 90, 99] only consider RGB images as input. [106, 121] attempt to utilize the multi-modal information, but their applicability is limited to RGB+D/T/E tracking tasks. Moreover, these models fail to capture the unified temporal attention mechanisms observed in human tracking.

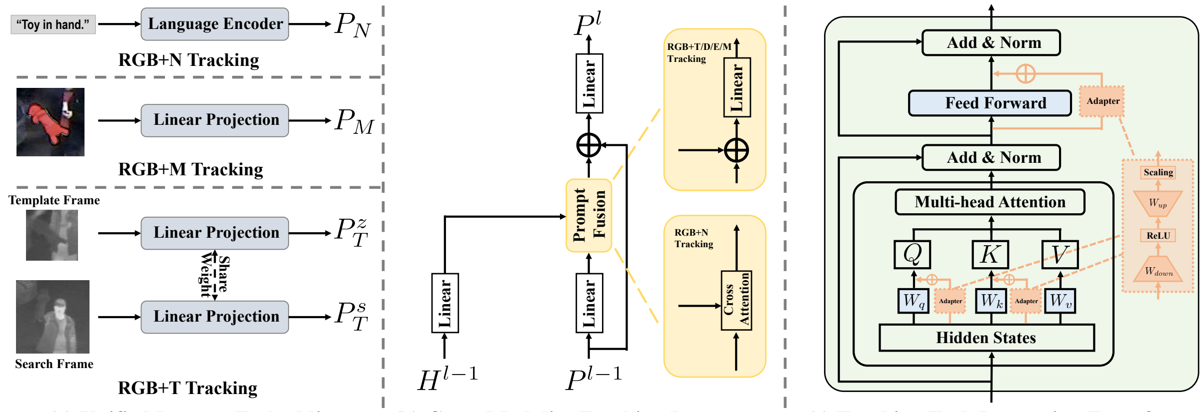
To address these challenges, we propose **OneTracker**, a general framework to unify RGB and RGB+X tracking within a consistent format. OneTracker firstly presents a Foundation Tracker for RGB tracking tasks, and then adapts it to RGB+X with parameter-efficient strategy. In detail, we pretrain a Foundation Tracker [35] on several RGB tracking datasets [28, 45, 67] (Figure 1 (c)). After the large-scale pretraining, Foundation Tracker possesses strong localization capabilities, allowing it to accurately locate the target object in the search frame based on its appearance in the template frame. Then, we proceed to finetune Foundation Tracker on specific downstream RGB+X tracking tasks, referred as Prompt Tracker (Figure 1 (d)). In contrast to leveraging an extra parallel module to fuse multimodal information, we propose Cross Modality Tracking Prompters (CMT Prompter) to introduce multimodal features in a prompt-tuning manner. CMT Prompters learn semantic understanding of multimodal information and integrate it with RGB images. Furthermore, to enhance the adaptation to downstream task, we replace the vanilla Transformer layers with Tracking Task Perception Transformer (TTP Transformer) layers. Because the linear layer in Transformer contains most of the knowledge of specific tasks, we only introduce few trainable parameters into each linear layer of Transformer to bridge the RGB tracking and RGB+X tracking. By leveraging CMT Prompter and TTP Transformer layer, Prompt Tracker inherits the strong localization ability from Foundation Tracker and achieves competitive performance on downstream RGB+X tracking tasks after a quick finetuning on a small number of parameters. Given the minimal number of additional parameters, Prompt Tracker maintains a similar speed to Foundation Tracker.

Overall, our contributions are summaried as follows:

- We present a unified tracking architecture, termed as OneTracker, which is consisted of Foundation Tracker and Prompt Tracker, to tackle various forms of tracking tasks, i.e. both RGB tracking and RGB+N/M/D/T/E tracking.
- We propose a Foundation Tracker trained on several RGB tracking datasets, which owns strong ability to accurately localize target objects in search frame.
- To better adapt Foundation Tracker to downstream RGB+X tracking tasks efficiently, we propose CMT Prompter and TTP Transformer layer, enhancing the model's ability to incorporate additional modalities, termed as Prompt Tracker.
- OneTracker achieves state-of-the-art performance on 11 benchmarks from 6 tracking tasks.

2. Related Works

Large-scale Pretraining Vision Models. Large-scale pretraining models, or foundation models [7], have emerged as powerful models which are trained on broad data and



(a) Unified Prompt Embedding **(b) Cross Modality Tracking Prompters** **(c) Tracking Task Perception Transformer**

Figure 2. **(a) Unified Prompt Embedding structure. (b) Cross Modality Tracking (CMT) Prompters. (c) Tracking Task Perception (TTP) Transformer layers.**

can be adapted to various downstream tasks. These models, initially popularized in Natural Language Processing (NLP) by [23, 64, 75, 77], have extended their influence to multiple domains. In the realm of computer vision, [111] is the first to extend ViT [25] to 2 billion parameters. [3, 39, 97] learn representations from images corrupted by masking. [30, 76, 80, 87, 96] explore the vision-language training strategy to align visual and text feature in a unified space. These large-scale pretraining vision models have shown their exceptional transferability across various downstream tasks. Inspired by the success of large-scale pretraining strategies, we propose Foundation Tracker, which is trained on a combination of diverse tracking datasets and demonstrates strong temporal matching capabilities.

Parameter-Efficient Transfer Learning. Parameter-efficient transfer learning (PETL) is introduced to serve as a lightweight alternative to address this limitation, which involves freezing the pretrained language model and adding a small number of extra trainable parameters to achieve quick adaptation to downstream tasks while maintaining parameter efficiency [43, 44, 54, 61, 63, 73]. PETL also demonstrates its high efficiency in computer vision fields. VPT [47] inserts additional parameters to the input sequence before encoder. Diverse kinds of adapter [11, 14, 33, 116] are proposed to adjust ViT to downstream tasks. ProTrack [106] and ViPT [121] attempt to introduce the prompting concept into trakcing area, while they just focus on RGB+D/T/E tracking. The question of how to transfer large-scale pretraining tracker to other tracking tasks, such as RGB+M and RGB+N tracking, remains unanswered. In this work, we propose Prompt Tracker based on Foundation Tracker. We introduce CMT Prompter and TTP Transformer layer to perform the parameter-efficient finetuing on RGB+X tracking tasks.

Visual Tracking. Visual object tracking is a fundamental task, including RGB tracking and RGB+X tracking. RGB tracking [29, 68] utilizes raw RGB images for object tracking. Because only levaraging pure RGB image is prone to some complex scenarios, RGB+X tracking is proposed for robust tracking by incorporating multimodal information. RGB+T/E/D tracking [93, 102, 112, 113, 115, 122] take advantage of thermal or event flows or depth maps. RGB+N tracking [28, 62, 89] fuses language description with RGB images, and RGB+M tracking [24, 37, 40–42, 72, 95] provides the mask of target in the first frame. Despite promising performance of task-specific [13, 18, 91, 109] or multi-task [1, 66, 85, 90, 99, 100] trackers, these models can not simulate the human temporal matching mechanism well and lack the ability to handle multi-modal tracking tasks. In this work, we propose a general manner to unify RGB and RGB+X tracking tasks.

3. Methodology

In this work, we propose OneTracker, consisting of Foundation Tracker and Prompt Tracker, to implement a unified framework for tracking tasks. The overall structure of Foundation Tracker and Prompt Tracker is in Figure 1 (c) and (d). We will illustrate the unification of tracking tasks (Sec. 3.1), the structure of Foundation Tracker (Sec. 3.2), the structure of Prompt Tracker (Sec. 3.3), and details of training and finetuning (Sec. 3.4).

3.1. Tracking Unification

The core aspect of tracking involves estimating the position of moving objects in each video frame based on its initial appearance. Depending on the different input format, tracking tasks can be divided into two main categories: RGB Tracking and RGB+X Tracking (Figure 1 (a) and (b)).

RGB Tracking. RGB tracking is a extensively studied tracking task, focusing on tracking objects using RGB image information. Given a video sequence with the bounding box of target object in the first frame, the formula of RGB tracking is like,

$$B = \mathrm{FT}(I, B_0; \theta), \tag{1}$$

where I, B_0, B denote the RGB frames of a video, the initial box prediction, the box predictions in the subsequent frames. FT is the Foundation Tracker with parameter θ.

RGB+X Tracking. We introduce an unified format that encompasses RGB+X tracking and RGB tracking. The formula for RGB+X tracking can be expressed as:

$$B = \mathrm{PT}(I, B_0, X; \theta^{'}), \tag{2}$$

where X is the additional information input of specific RGB+X tracking task and PT is the Prompt Tracker with parameter $\theta^{'}$. X varies depending on the RGB+X tracking task. For RGB+N tracking, X_N is the language description. For RGB+M tracking, X_M is the mask of the target in the first frame and B_0 is not provided. For RGB+D/T/E tracking, X_D, X_T, and X_E correspond to the depth, thermal, and event map of each frame. To further illustrate our framework, we can rewrite the Equation 2 as follows:

$$\begin{aligned} B &= \mathrm{PT}(I, B_0, X; \mathrm{TTP}(\theta)) \\ &= \mathrm{FT}(\mathrm{CMT}(I, B_0, X); \mathrm{TTP}(\theta)), \end{aligned} \tag{3}$$

where TTP denotes the replacement of the vanilla Transformer layers with our Tracking Task Perception (TTP) Transformer layers and CMT is the Cross Modality Tracking Prompters (CMT Prompters). Through our specific design, we succeed in unifying RGB tracking and RGB+X tracking in a general format.

3.2. Foundation Tracker

The structure of Foundation Tracker (Figure 1 (c)) is similar to ViT [25] with several transformer encoder layers, which are responsible for processing the input frames and capturing their spatial and temporal dependencies. To begin, the template frame I_z and search frame I_s are taken as input to Foundation Tracker. The RGB frames I_z and I_s are divided into patches and flattened into 1D tokens $H^z_{RGB} \in \mathbb{R}^{N_z \times D}$ and $H^s_{RGB} \in \mathbb{R}^{N_s \times D}$, where N_z and N_s denote the token number of template and search frame, and D is the dimension of tokens. Then the tokens are concatenated into $H^0 = H^0_{RGB} = [H^z_{RGB}, H^s_{RGB}]$ and fed into L-layer transformer encoder layers. The forward process of the transformer encoder layers can be written as:

$$H^l = \mathrm{E}^l(H^{l-1}), l = \{1, 2, ..., L\}, \tag{4}$$

where E^l is the l-th transformer encoder layers, and H^{l-1} is the input to E^l. The structure of transformer encoder layer is the same as the vanilla transformer layer [83]. We extract features and build the temporal matching between template and search frame. Finally, a box head is leveraged to convert the temporal correlation from transformer encoder into localization coordinates. Moreover, an extra segmentation head is leveraged to generate the mask prediction, whose structure is the same as [108]. We train Foundation Tracker on several RGB tracking benchmarks, including LaSOT [28], TrackingNet [67], and GOT-10K [45]. The segmentation head is optimized in a box-supervised manner [82] because of the absence of mask ground truth in tracking datasets. Similar to the foundation model in NLP, after the large-scale pretraining, our Foundation Tracker obtains the strong ability of temporal matching and transferability to downstream RGB+X tracking tasks.

3.3. Prompt Tracker

Different from previous RGB+X works, which add an additional module to fuse mutlimodal features, we regard the multimodal information as a kind of prompt and provide Foundation Tracker with complementarity in a prompt-tuning manner, termed as Prompt Tracker (Figure 1 (d)). To enable efficient adaptation to downstream tasks, we propose the Cross Modality Tracking Prompters (CMT Prompters) and the Tracking Task Perception Transformer (TTP Transformer) layers.

Unified Prompt Embedding. With the general definition of RGB+X tracking, the Prompt Tracker leverages a unified prompt embedding module (Figure 2 (a)) to transform different modality downstream information into tokens $P^0 = P_X$. The choice of prompt embedding strategy depends on the specific downstream task's modality. To deal with language description in RGB+N tracking, we adopt BERT [23] as a language encoder to extract the linguistic feature $P_N \in \mathbb{R}^{L \times D}$ with a sequence length of L. For RGB+M tracking, a patch embed layer is utilized to project the mask of the target object into patches and flatten them into 1D tokens $P_M \in \mathbb{R}^{N_z \times D}$. The size of P_M is the same as H^z_{RGB}. For RGB+T tracking, the corresponding multimodal maps of the template and search frame are fed into a patch embed layer and then flattened into 1D tokens $P_T \in \mathbb{R}^{N_z \times D}$ and $P_T \in \mathbb{R}^{N_s \times D}$. The prompt embedding of RGB+D and RGB+E tracking follows the same procedure as RGB+T tracking. Through the unified prompt embedding module, we effectively map the multimodal information into a cohesive token representation.

Cross Modality Tracking Prompters. After unified prompt embedding, we propose the Cross Modality Tracking Prompters (CMT Prompters) to fuse the extra information. Although a few works have attempted to insert some trainable parameters into pretrained models to bridge the upstream and downstream tasks, how to integrate the cross-modal information for tracking tasks is more challenging. CMT Prompters are designed to extract the semantic representations of multimodal information and provide Foundation Tracker with complementarity. As depicted in Figure 2 (b), CMT Prompters consist of multiple linear layers and a prompt fusion module, which can be written as:

$$P^{l+1} = \mathrm{CMT}^l(H^l, P^l), l = \{0, 1, ..., L-1\}, \tag{5}$$

where CMT^{l-1} denotes the l-th CMT Prompter, and P^l is the output of CMT^l. The prompt P^l is added to the original matching results H^l in a form of residuals:

$$H^l = H^l + P^{l+1}, l = \{0, 1, ..., L-1\}. \tag{6}$$

CMT Prompters take the matching results H^l from l-th transformer encoder layer and the prompt P^l as input.

RGB Tracking														
		TransT [12]	STARK [98]	MixFormer [17]	OSTrack [110]	AiATrack [34]	SimTrack [9]	GRM [35]	UniTrack [90]	UTT [66]	Unicorn [99]	OmniTracker [85]	UNINEXT [100]	**One Tracker**
LaSOT [28]	AUC(↑)	64.9	66.4	69.2	69.1	69.0	69.3	69.9	35.1	64.6	65.3	69.1	69.2	**70.5**
	P_{Norm}(↑)	73.8	76.3	78.7	78.7	79.4	78.5	79.3	-	-	73.1	77.3	77.1	**79.9**
	P(↑)	69.0	71.2	74.7	75.2	73.8	74.0	75.8	32.6	67.2	68.7	75.4	75.5	**76.5**
TrackingNet [67]	AUC(↑)	81.4	81.3	83.1	83.1	82.7	82.3	**84.0**	-	79.7	79.0	83.4	83.2	83.7
	P_{Norm}(↑)	86.7	86.1	88.1	87.8	87.8	86.5	**88.7**	-	-	82.0	86.7	86.9	88.4
	P(↑)	80.3	78.1	81.6	82.0	80.4	-	**83.3**	-	77.0	77.4	82.3	**83.3**	82.7

RGB+N Tracking														
		TNLS-III [62]	RTTNLD [31]	SiamRPN [55]	VITAL [79]	MDNet [69]	ATOM [21]	DiMP [5]	PrDIMP [22]	SiamRPN++ [56]	TNL2K-2 [89]	SNLT [32]	JointNLT [118]	**One Tracker**
OTB99 [62]	AUC(↑)	55.0	61.0	61.2	65.2	64.6	67.6	67.3	68.3	65.8	68.0	66.6	65.3	**69.7**
	P(↑)	72.0	79.0	75.8	84.2	82.8	82.4	81.9	83.0	79.7	88.0	80.4	85.6	**91.5**
TNL2K [89]	AUC(↑)	-	25.0	-	-	-	-	-	-	-	42.0	27.6	56.9	**58.0**
	P(↑)	-	27.0	-	-	-	-	-	-	-	42.0	41.9	58.1	**59.1**

RGB+D Tracking														
		ATOM [20]	LTDSEd [49]	DRefine [51]	keep_track [52]	LTMU_B [50]	DiMP [6]	DDiMP [50]	DeT [101]	OSTrack [109]	SPT [123]	ProTrack [105]	ViPT [121]	**One Tracker**
DepthTrack [102]	F-score(↑)	-	40.5	-	-	46.0	-	48.5	53.2	52.9	53.8	57.8	59.4	**60.9**
	R(↑)	-	38.2	-	-	41.7	-	46.9	50.6	52.2	54.9	57.3	59.6	**60.4**
	P(↑)	-	43.0	-	-	51.2	-	50.3	56.0	53.6	52.7	58.3	59.2	**60.7**
VOT RGBD2022 [52]	EAO(↑)	50.5	-	59.2	60.6	-	54.3	-	65.7	67.6	65.1	65.1	72.1	**72.7**
	Accuracy(↑)	69.8	-	77.5	75.3	-	70.3	-	76.0	80.3	79.8	80.1	81.5	**81.9**
	Robustness(↑)	68.8	-	76.0	79.7	-	73.1	-	84.5	83.3	85.1	80.2	87.1	**87.2**

RGB+T Tracking														
		SGT++ [57]	DAPNet [124]	HMFT [115]	FANet [125]	mfDiMP [114]	STARKS50 [98]	CAT [59]	APFNet [92]	OSTrack [110]	TransT [12]	ProTrack [105]	ViPT [121]	**One Tracker**
LasHeR [60]	PR(↑)	36.5	43.1	43.6	44.1	44.7	44.9	45.0	50.0	51.5	52.4	53.8	65.1	**67.2**
	SR(↑)	25.1	30.9	31.3	31.4	31.4	34.3	36.1	36.2	39.4	41.2	42.0	52.5	**53.8**
RGBT234 [58]	MPR(↑)	64.6	72.0	79.6	72.9	78.7	79.0	80.4	79.0	82.3	82.7	79.5	83.5	**85.7**
	MSR(↑)	42.8	47.2	54.4	54.9	55.3	55.4	56.1	57.3	57.5	57.9	59.9	61.7	**64.2**

RGB+E Tracking														
		MetaTracker [71]	ATOM [20]	STARKS50 [98]	ProTrack [105]	PrDIMP50 [22]	VITAL [79]	TransT [12]	LTMU [19]	SiamRCNN [84]	MDNet [69]	OSTrack [110]	ViPT [121]	**One Tracker**
VisEvent [88]	MPR(↑)	49.1	60.8	61.2	63.2	64.4	64.9	65.0	65.5	65.9	66.1	69.5	75.8	**76.7**
	MSR(↑)	29.8	41.2	44.6	47.1	45.3	-	47.4	45.9	49.9	-	53.4	59.2	**60.8**

RGB+M Tracking														
		STM [70]	CFBI [107]	AOT [108]	STCN [16]	XMem [15]	SiamMask [86]	Siam R-CNN [84]	UniTrack [90]	Unicorn [99]	TarVIS [1]	OmniTracker [85]	UNINEXT [100]	**One Tracker**
DAVIS16 [72]	$\mathcal{J}\&\mathcal{F}$(↑)	89.3	89.4	91.1	91.6	92.0	69.8	-	-	87.4	-	88.5	-	**88.9**
	$\mathcal{J}$(↑)	88.7	88.3	90.1	90.8	90.7	71.7	-	-	86.5	-	87.3	-	**88.1**
	$\mathcal{F}$(↑)	89.9	90.5	92.1	92.5	93.2	67.8	-	-	88.2	-	89.7	-	**89.7**
DAVIS17 [74]	$\mathcal{J}\&\mathcal{F}$(↑)	81.8	81.9	84.9	85.4	86.2	56.4	70.6	-	69.2	82.0	71.0	81.8	**82.5**
	$\mathcal{J}$(↑)	79.2	79.1	82.3	82.2	82.9	54.3	66.1	58.4	65.2	78.7	66.8	77.7	**79.4**
	$\mathcal{F}$(↑)	84.3	84.6	87.5	88.6	89.5	58.5	75.0	-	73.2	87.0	75.2	85.8	**85.6**

Table 1. **Overall performance on RGB tracking and RGB+X tracking.**

Firstly, the H^l and P^l are mapped to lower-dimensional latent space using a linear layer, respectively. Subsequently, a prompt fusion module is employed to integrate the modalities. For RGB+N tracking, cross-attention is utilized to merge the linguistic feature and temporal correlation. For other RGB+X tracking, P^l and H^l are added and merged by a linear layer. Finally, another linear layer projects the fused feature to the original dimension. The structure and format of the CMT Prompter remain consistent across different RGB+X tracking tasks. By leveraging CMT Prompter, we achieve the integration between RGB images and multimodal information with high efficiency through prompt-tuning techniques.

Tracking Task Perception Transformer. Although CMT Prompter effectively complements auxiliary modalities as prompts, the Prompt Tracker lacks specialization for certain downstream tasks. For example, Foundation Tracker excels at localizing targets based on RGB images, while the lack of perception of linguistic features may result in sub-optimal performance in RGB+N tracking tasks. Given the Foundation Tracker parametrazed by θ, θ may not be the optimal weights for downstream tasks. Suppose the best weights on downstream tasks are $\theta^{'}$, the purpose of full finetuning is to learn difference $\Delta\theta$ between θ and $\theta^{'}$ and update Foundation Tracker to $\theta + \Delta\theta$. The drawbacks of full finetuning are that we must learn a different set of parameters $\Delta\theta$, whose dimension $|\Delta\theta|$ is equal to $|\theta|$ for each different downstream task, and lack of large-scale data in

Method	# Params	LaSOT			DepthTrack			LasHeR		VisEvent		OTB		DAVIS17		
		AUC	P_{norm}	P	F	R	P	PR	SR	PR	SR	AUC	P	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Foundation Tracker	-	70.5	79.9	76.5	55.9	55.6	55.7	53.3	42.1	70.1	53.6	67.3	88.9	42.7	37.4	48.1
Full Finetune	99.83M	-	-	-	57.2	56.9	57.1	65.4	52.5	75.6	59.8	68.5	89.6	77.8	75.4	80.2
Prompt Tracker	2.8M	-	-	-	**60.9**	**60.4**	**60.7**	**67.2**	**53.8**	**76.7**	**60.8**	**69.7**	**91.5**	**82.5**	**79.4**	**85.6**
w/o CMT Prompters	2.55M	-	-	-	56.5	55.4	56.7	60.7	47.1	74.0	54.5	68.7	89.9	80.4	78.6	82.2
w/o TTP TransFormer	0.25M	-	-	-	59.2	58.8	59.1	65.6	52.3	75.3	59.0	69.3	90.8	81.7	79.3	84.0

Table 2. **Ablation study on the Prompt Tracker.** # Params denotes the number of trainable parameters.

specific downstream tasks may lead to suboptimal finetuning performance and result in catastrophic forgetting.

Thus, to bridge the gap between RGB tracking and downstream RGB+X tracking tasks, we propose Tracking Task Perception Transformer (TTP Transformer) by adding some adapters to Foundation Tracker. Where to insert trainable parameters is a crucial question. The linear layer in transformer encoder contains the knowledge of specific tasks, especially the linear layer in Feed Forward Network (FFN) [38]. As shown in Figure 2 (c), we insert trainable adapters with a small number of parameters to the linear projection operation in vanilla transformer encoder layers, i.e. the query/key/value projection matrixes and the output layers in FFN to enable the efficient adaption. The structure of the adapter follows a similar approach with [38]. For a pretrained linear layer with weight matrix $W \in \mathbb{R}^{d\times k}$, the formula can be written as:

$$h = Wx, \tag{7}$$

where $h \in \mathbb{R}^{d\times t}$ and $x \in \mathbb{R}^{k\times t}$ denote the output and input. k and d are the dimension of h and x. t is the token number of x. With an adapter, the process becomes:

$$h = Wx + \Delta Wx = Wx + s \cdot W_{up}\text{ReLU}(W_{down}x), \tag{8}$$

where $W_{down} \in \mathbb{R}^{d\times r}$ and $W_{up} \in \mathbb{R}^{r\times d}$ is two mapping matrixes, ReLU is relu operation, s is the constant scaling factor, and rank $r \ll min(d,k)$. During finetuning, we freeze the W and only update W_{down} and W_{up}. Through optimizing W_{down} and W_{up}, we achieve the highly efficient learning of $\Delta\theta$. The TTP Transformer layers bridge the RGB tracking and RGB+X tracking while maintaining the temporal matching knowledge in Foundation Tracker.

3.4. Training and Inference

Training. The whole training process of OneTracker consists of two stages: Foundation Tracker pretraining and Prompt Tracker finetuning. In the first pretraining stage, we pretrain our Foundation Tracker on a combination of several large-scale RGB tracking datasets, including LaSOT [28], TrackingNet [67], and GOT-10K [45], which is the same as previous trackers [18, 35, 110]. Following [18, 35, 110], we adopt the weighted focal loss [53] for classification, l_1 loss and generalized IoU loss [78] for bounding box regression. Because there is no mask annotations in tracking datasets, we leverage BoxInst [82] to supervise the segmentation head of Foundation Tracker. The total loss function can be formulated as:

$$L_{stage1} = L_{cls} + \lambda_{iou}L_{iou} + \lambda_{L_1}L_1 + \lambda_{L_{mask}}L_{mask}^{boxinst} \tag{9}$$

In the second stage, we finetune our Foundation Tracker on RGB+X downstream tracking datasets. We freeze the parameters of Foundation Tracker and only train the CMT Prompters and adapters in TTP Transformer layers. For RGB+ N/D/T/E tracking, the loss function L_{stage2} is equal to L_{stage1}. For RGB+M tracking, due to the available mask annotations, we drop the BoxInst auxiliary loss and utilize the mask annotations to optimize the segmentation head.

Inference. Because of the slight difference in the input format of several tracking tasks, we adopt different inference manners. For RGB tracking and RGB+N tracking, Hanning window penalty is utilized to leverage the positional prior following previous works [18, 35, 110]. For RGB+D/T/E tracking, the multimodal map is also cropped by using Hanning window penalty. For RGB+M tracking, the first frame with the mask annotation and the previous frame with the predicted mask are fed into the Prompt Tracker to perform online target matching without cropping. Due to the specific design of OneTracker, we can apply them to several tracking tasks without any modification to the structure of models.

4. Experiments

4.1. Implementation Details

OneTracker is built on the encoder of ViT-B [26], which includes 12 sequential transformer layers. The box head and segmentation head follow the structure in [18, 35, 109] and [108], respectively. For RGB+N tracking tasks, we adopt BERT [23] as text encoder. During the first pretraining stage, Foundation Tracker is optimized with AdamW optimizer [65] for 300 epochs. The initial learning rate is 4×10^{-5} for the ViT backbone and 4×10^{-4} for the heads. It decays by a factor of 10 after 240 epochs. For the finetuning of Prompt Tracker, we freeze the parameters of Foundation Tracker and only adapt the CMT Prompter and TTP Transformer layers. We finetune Prompt Tracker for 60 epochs on the corresponding training data of each downstream tasks with an initial learning rate of 4×10^{-5} by using AdamW optimizer, and learning rate decreased by 10 after 48 epochs. We set λ_{iou} as 2, λ_{L_1} as 5, $\lambda_{L_{mask}}$ as 1, and r as 16. More details are in supplementary materials.

Number	# Params	DepthTrack			LasHeR		VisEvent		OTB		DAVIS17		
		F	R	P	PR	SR	PR	SR	AUC	P	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
0	-	56.5	58.8	59.1	60.7	47.1	74.0	54.5	68.7	89.9	80.4	78.6	82.2
1	0.02M	57.6	57.2	57.3	61.5	48.7	75.7	59.2	68.9	90.0	**82.5**	**79.4**	**85.6**
2	0.04M	58.4	58.1	58.2	63.2	50.1	76.0	59.5	69.2	90.4	75.4	60.7	77.5
4	0.08M	59.1	58.9	59.3	64.1	51.0	76.1	59.7	69.3	90.4	67.1	73.3	73.5
6	0.12M	59.5	59.3	59.4	65.7	52.5	76.4	60.3	69.5	91.0	58.7	52.5	64.9
12	0.25M	**60.9**	**60.4**	**60.7**	**67.2**	**53.8**	**76.7**	**60.8**	**69.7**	**91.5**	48.3	44.5	52.1

Table 3. **Ablation study on the number of CMT Prompters.** # Params denotes the number of trainable parameters. In this experiment, we just count the number of parameters in CMT Prompters.

Task	CMT	TTP	DepthTrack			LasHeR		VisEvent		OTB		DAVIS17		
			F	R	P	PR	SR	PR	SR	AUC	P	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
			53.9	53.2	53.4	59.7	48.5	73.5	56.8	68.4	89.6	58.7	51.1	66.3
	✓		55.4	54.4	54.9	62.8	50.2	74.3	57.7	68.9	90.3	65.3	60.2	70.4
		✓	57.9	57.0	57.2	64.8	52.0	74.8	58.4	69.2	90.5	70.2	66.8	73.6
	✓	✓	58.6	58.1	57.9	66.8	53.2	74.9	58.8	69.6	91.1	76.4	73.5	79.3
✓	✓	✓	**60.9**	**60.4**	**60.7**	**67.2**	**53.8**	**76.7**	**60.8**	**69.7**	**91.5**	**82.5**	**79.4**	**85.6**

Table 4. **Ablation study on the training strategy.**

4.2. Benchmark Results on 6 Tracking Tasks

We evaluate our OneTracker on 6 tracking tasks, including RGB tracking and RGB+X Tracking. We compare the results with task-specific counterparts in Table 1.

RGB Tracking. To show the strong temporal matching ability of our OneTracker, we compare our Foundation Tracker on widely-used RGB tracking benchmarks: LaSOT [28] and TrackingNet [67]. Area under the success curve (AUC), normalized precision (P_{Norm}), and precision (P) are adopted as metrics. Our model achieves 70.5 AUC and 69.7 AUC on LaSOT and TrackingNet, respectively, outperforming all other trackers. UniTrack, UTT, Unicorn, and OmniTracker are designed for multiple RGB tracking tasks, and UNINEXT is good at instance perception. Our Foundation Tracker surpasses these models at least 1.3 AUC on LaSOT. GRM is one of the strongest tracking-specific model, and Foundation Tracker also outperforms it 0.6 AUC on LaSOT. Moreover, our Foundation Tracker can generate the mask of the target object due to an extra segmentation head, which is not possible with other trackers.

RGB+N Tracking. Following previous works [118], we conduct experiments on OTB99 and TNL2K. Our Prompt Tracker superpasses all existing RGB+N trackers at least 1.7 AUC and 2.5 precision on OTB99, although Prompt Tracker is not specifically designed for RGB+N tracking.

RGB+D/T/E Tracking. Following [121], we evalute our Prompt Tracker on DepthTrack [101] and VOT-RGBD2022 [52] for RGB+D tracking, LasHeR [60] and RGBT234 [58] for RGB+T tracking, and VisEvent [88] for RGB+E tracking. Our Prompt Trakcer greatly exceeds all other trackers in terms of performance. Despite the fact that ViPT [121] also adopts a similar prompt-tuning structure, our superior results demonstrate the effectiveness of CMT Prompters and TTP Transformer layers.

RGB+M Tracking. We choose DAVIS16 [72] and DAVIS17 [74] for RGB+M tracking. DAVIS16 is a single-object benchmark with 20 evaluation splits and DAVIS17 is the multi-object expansion of DAVIS16. Region similarity $\mathcal{J}$, contour accuracy $\mathcal{F}$, and their averaged score $\mathcal{J}\&\mathcal{F}$ are adopted as metrics. Our Prompt Tracker achieves the best performance on its multi-task counterparts [1, 100] and other unified tracking models [84–86, 90, 99] by a large margin on both DAVIS16 and DAVIS17, despite its less training data and training cost. There still exists a small gap between Promp Tracker and specific models with memory mechanisms [15, 16, 108], but our Prompt Tracker only relies on the first and previous frame, enabling it to handle videos with any length.

4.3. Ablation Study.

Foundation Tracker and Prompt Tracker. To verify the strong temporal matching ability of Foundation Tracker and the effectiveness of parameter-efficient finetuning, we evaluate the performance on RGB and RGB+X tracking in Table 2. In the first row, we solely feed the RGB image into Foundation Tracker, and the results demonstrate its strong ability to track based on visual information alone. Then, we conduct full finetuning of the Foundation Tracker on downstream RGB+X tracking datasets, as well as parameter-efficient finetuning using our proposed CMT Prompters and TTP Transformer layers. The integration of multi-modal information boosts the localization accuracy, while our Prompt Tracker achieves better performance while only adjusting 2.8M parameters, which highlights the effectiveness of our Foundation Tracker and Prompt Tracker. Furthermore, we perform an ablation study on the CMT Prompters and TTP Transformer layers. The results in the fourth and fifth rows illustrate the impact of these components in en-

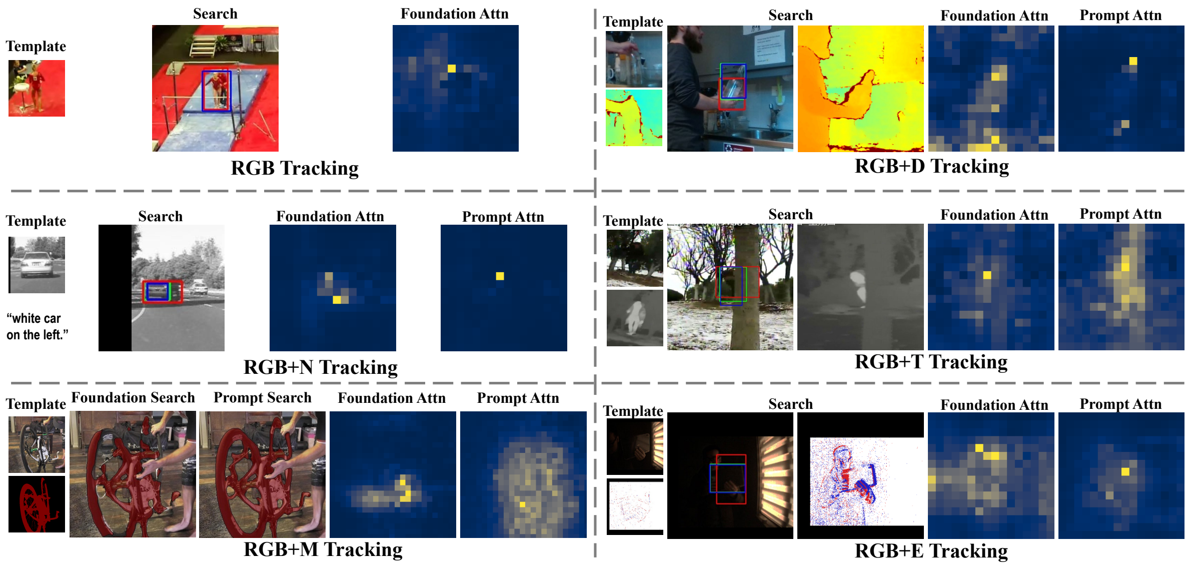


Figure 3. **Visualization results.** The **blue**, **red**, and **green** bounding boxes denote ground truth, Foundation Tracker, and Prompt Tracker. Foundation Attn and Prompt Attn denotes the attention map of Foundation Tracker and Prompt Tracker.

hancing the tracking performance.

CMT Prompter Layers. We explore the impact on the performance of different numbers of CMT Prompter layers in Table 3. We insert CMT Prompters at different positions, each 1, 2, 3, 6, 12 transformer blocks. A value of 0 for CMT Prompter layer denotes that we add the embedding of multimodal information to RGB image embeddings directly. As the number of CMT Prompter layers increases, the performance of Prompt Tracker improves, suggesting the effectiveness of our CMT Prompters. However, interestingly, the performance on RGB+M tracking shows the opposite trend, with a significant drop in performance as the number of CMT Prompter layers increases. This observation demonstrates that the mask embedding is effective only in capturing superficial features and does not provide substantial benefits when incorporated deeply into the model.

Training Strategy. We analyze the training strategy of our Prompt Tracker on RGB+X tracking tasks in Table 4. We investigate different approaches to jointly training the Prompt Tracker on multiple RGB+X datasets. From the first to the fourth row, we jointly train Prompt Tracker on the combination of several RGB+X datasets. In the first row, we train the Prompt Tracker by only separating the embedding layers for different modalities.. In the second and third rows, we separate the CMT Prompters and the TTP Transformer layers for different modalities, respectively. In the fourth row, we separate both the CMT Prompters and TTP Transformer layers. By progressively separating the training of each module, we observe continuous improvement in performance. This phenomenon can be attributed to the limited amount of training data available for each downstream task, making it hard to train the Prompt Tracker jointly for all RGB+X tracking tasks. In the fifth row, Prompt Tracker is trained on corresponding data for specific tasks, achieving better performance.

Temporal Matching Attention Visualization. We visualize the temporal matching attention map of Foundation Tracker and Prompt Tracker in Figure 3. The Foundation Tracker, after undergoing extensive large-scale training, demonstrates its strong temporal matching ability. It effectively captures the temporal dependencies and provides accurate predictions for the target object. With the parameter-efficient finetuning on downstream RGB+X tracking datasets, the Prompt Tracker further improves the tracking performance. By leveraging multimodal information and refining predictions, Prompt Tracker achieves more precise and accurate results on certain datasets. These visualizations demonstrate the effectiveness of our OneTracker in establishing temporal correspondences.

5. Conclusion

We propose a general framework, OneTracker, to unify several RGB tracking and RGB+X tracking tasks. OneTracker involves pretraining a Foundation Tracker on RGB tracking datasets and adapting it to downstream RGB+X tracking tasks using prompt-tuning techniques. By leveraging the strengths of pretraining and parameter-efficient finetuning mechanisms, our framework achieves state-of-the-art results in various tracking scenarios. Superior performance on 11 benchmarks of 6 tasks demonstrates the effectiveness and powerful generation ability of OneTracker.

Acknowledgments: This work was supported by National Natural Science Foundation of China (No.62072112), Scientific and Technological Innovation Action Plan of Shanghai Science and Technology Committee (No.22511102202).

References

[1] Ali Athar, Alexander Hermans, Jonathon Luiten, Deva Ramanan, and Bastian Leibe. Tarvis: A unified approach for target-based video segmentation. *arXiv preprint arXiv:2301.02657*, 2023. 2, 3, 5, 7

[2] Armin Bahl, Georg Ammer, Tabea Schilling, and Alexander Borst. Object tracking in motion-blind flies. *Nature neuroscience*, 16(6):730–738, 2013. 1

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3

[4] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*, pages 850–865. Springer, 2016. 1

[5] 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. 5

[6] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *ICCV*, 2019. 5

[7] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 2

[8] György Buzsáki and Rodolfo Llinás. Space and time in the brain. *Science*, 358(6362):482–485, 2017. 1

[9] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 375–392. Springer, 2022. 5

[10] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE international conference on computer vision*, pages 2722–2730, 2015. 1

[11] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *arXiv preprint arXiv:2205.13535*, 2022. 3

[12] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8126–8135, 2021. 5

[13] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6668–6677, 2020. 1, 3

[14] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 3

[15] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 640–658. Springer, 2022. 5, 7

[16] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 5, 7

[17] 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. 5

[18] 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. 3, 6

[19] 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. 5

[20] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ATOM: Accurate tracking by overlap maximization. In *CVPR*, 2019. 5

[21] 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. 5

[22] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7183–7192, 2020. 5

[23] 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. 3, 4, 6

[24] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. *arXiv preprint arXiv:2302.01872*, 2023. 1, 3

[25] 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. 3, 4

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold,

Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 6

[27] John Duncan. Selective attention and the organization of visual information. *Journal of experimental psychology: General*, 113(4):501, 1984. 1

[28] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 1, 2, 3, 4, 5, 6, 7

[29] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. LaSOT: A high-quality benchmark for large-scale single object tracking. In *CVPR*, 2019. 3

[30] Rongyao Fang, Shilin Yan, Zhaoyang Huang, Jingqiu Zhou, Hao Tian, Jifeng Dai, and Hongsheng Li. Instructseq: Unifying vision tasks with instruction-conditioned multi-modal sequence generation. *arXiv preprint arXiv:2311.18835*, 2023. 3

[31] Qi Feng, Vitaly Ablavsky, Qinxun Bai, Guorong Li, and Stan Sclaroff. Real-time visual object tracking with natural language description. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 700–709, 2020. 5

[32] Qi Feng, Vitaly Ablavsky, Qinxun Bai, and Stan Sclaroff. Siamese natural language tracker: Tracking by natural language descriptions with siamese trackers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5851–5860, 2021. 5

[33] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 3

[34] Shenyuan Gao, Chunluan Zhou, Chao Ma, Xinggang Wang, and Junsong Yuan. Aiatrack: Attention in attention for transformer visual tracking. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 146–164. Springer, 2022. 5

[35] Shenyuan Gao, Chunluan Zhou, and Jun Zhang. Generalized relation modeling for transformer tracking. *arXiv preprint arXiv:2303.16580*, 2023. 2, 5, 6

[36] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 1

[37] Pinxue Guo, Wei Zhang, Xiaoqiang Li, and Wenqiang Zhang. Adaptive online mutual learning bi-decoders for video object segmentation. *IEEE Transactions on Image Processing*, 31:7063–7077, 2022. 3

[38] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021. 6

[39] 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

[40] Lingyi Hong, Wei Zhang, Liangyu Chen, Wenqiang Zhang, and Jianping Fan. Adaptive selection of reference frames for video object segmentation. *IEEE Transactions on Image Processing*, 31:1057–1071, 2021. 3

[41] Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. Lvos: A benchmark for long-term video object segmentation. *arXiv preprint arXiv:2211.10181*, 2022. 1

[42] Lingyi Hong, Wei Zhang, Shuyong Gao, Hong Lu, and WenQiang Zhang. Simulflow: Simultaneously extracting feature and identifying target for unsupervised video object segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7481–7490, 2023. 3

[43] 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

[44] 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

[45] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2019. 2, 4, 6

[46] Laurent Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE transactions on image processing*, 13(10):1304–1318, 2004. 1

[47] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 709–727. Springer, 2022. 3

[48] Christin Kosse and Denis Burdakov. Natural hypothalamic circuit dynamics underlying object memorization. *Nature communications*, 10(1):2505, 2019. 1

[49] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Luka Čehovin Zajc, Ondrej Drbohlav, Alan Lukezic, Amanda Berg, et al. The seventh visual object tracking vot2019 challenge results. In *ICCVW*, pages 0–0, 2019. 5

[50] 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 *ECCVW*, pages 547–601. Springer, 2020. 5

[51] 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 *ICCVW*, pages 2711–2738, 2021. 5

[52] 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 *ECCVW*, pages 431–460. Springer, 2023. 5, 7

[53] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018. 6

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

[55] 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. 5

[56] 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. 1, 5

[57] Chenglong Li, Nan Zhao, Yijuan Lu, Chengli Zhu, and Jin Tang. Weighted sparse representation regularized graph learning for RGB-T object tracking. In *ACMMM*, pages 1856–1864, 2017. 5

[58] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. RGB-T object tracking: Benchmark and baseline. *Pattern Recognition*, 96:106977, 2019. 5, 7

[59] Chenglong Li, Lei Liu, Andong Lu, Qing Ji, and Jin Tang. Challenge-aware RGBT tracking. In *ECCV*, pages 222–237. Springer, 2020. 5

[60] 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. 5, 7

[61] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 3

[62] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. Tracking by natural language specification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6495–6503, 2017. 1, 3, 5

[63] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021. 3

[64] 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. 3

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

[66] Fan Ma, Mike Zheng Shou, Linchao Zhu, Haoqi Fan, Yilei Xu, Yi Yang, and Zhicheng Yan. Unified transformer tracker for object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8781–8790, 2022. 2, 3, 5

[67] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, pages 300–317, 2018. 2, 4, 5, 6, 7

[68] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. TrackingNet: A large-scale dataset and benchmark for object tracking in the wild. In *ECCV*, 2018. 3

[69] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4293–4302, 2016. 5

[70] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019. 5

[71] Eunbyung Park and Alexander C Berg. Meta-tracker: Fast and robust online adaptation for visual object trackers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 569–585, 2018. 5

[72] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 1, 3, 5, 7

[73] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*, 2020. 3

[74] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 5, 7

[75] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[76] 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

[77] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 3

[78] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 6

[79] Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, Wangmeng Zuo, Chunhua Shen, Rynson WH Lau, and Ming-Hsuan Yang. Vital: Visual tracking via adversarial learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8990–8999, 2018. 5

[80] Weijie Su, Xizhou Zhu, Chenxin Tao, Lewei Lu, Bin Li, Gao Huang, Yu Qiao, Xiaogang Wang, Jie Zhou, and Jifeng Dai. Towards all-in-one pre-training via maximizing multi-modal mutual information. *arXiv preprint arXiv:2211.09807*, 2022. 3

[81] Ying-Li Tian, Max Lu, and Arun Hampapur. Robust and efficient foreground analysis for real-time video surveillance. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 1182–1187. IEEE, 2005. 1

[82] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5443–5452, 2021. 4, 6

[83] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[84] 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. 5, 7

[85] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Xiyang Dai, Lu Yuan, and Yu-Gang Jiang. Omnitracker: Unifying object tracking by tracking-with-detection. *arXiv preprint arXiv:2303.12079*, 2023. 2, 3, 5

[86] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H. S. Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, 2019. 5, 7

[87] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 3

[88] 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. *arXiv preprint arXiv:2108.05015*, 2021. 5, 7

[89] Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13763–13773, 2021. 1, 3, 5

[90] Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip Torr, and Luca Bertinetto. Do different tracking tasks require different appearance models? *Advances in Neural Information Processing Systems*, 34:726–738, 2021. 2, 3, 5, 7

[91] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2411–2418, 2013. 1, 3

[92] Yun Xiao, Mengmeng Yang, Chenglong Li, Lei Liu, and Jin Tang. Attribute-based progressive fusion network for RGBT tracking. In *AAAI*, 2022. 5

[93] 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

[94] Junliang Xing, Haizhou Ai, and Shihong Lao. Multiple human tracking based on multi-view upper-body detection and discriminative learning. In *2010 20th International Conference on Pattern Recognition*, pages 1698–1701. IEEE, 2010. 1

[95] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 585–601, 2018. 1, 3

[96] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022. 3

[97] Hongwei Xue, Peng Gao, Hongyang Li, Yu Qiao, Hao Sun, Houqiang Li, and Jiebo Luo. Stare at what you see: Masked image modeling without reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22732–22741, 2023. 3

[98] 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. 5

[99] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI*, pages 733–751. Springer, 2022. 2, 3, 5, 7

[100] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. *arXiv preprint arXiv:2303.06674*, 2023. 2, 3, 5, 7

[101] Song Yan, Jinyu Yang, Jani Käpylä, Feng Zheng, Aleš Leonardis, and Joni-Kristian Kämäräinen. Depthtrack: Unveiling the power of RGBD tracking. In *ICCV*, pages 10725–10733, 2021. 5, 7

[102] 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. 3, 5

[103] Shilin Yan, Xiaohao Xu, Lingyi Hong, Wenchao Chen, Wenqiang Zhang, and Wei Zhang. Panovos: Bridging non-panoramic and panoramic views with transformer for video segmentation. *arXiv preprint arXiv:2309.12303*, 2023. 1

[104] Shilin Yan, Renrui Zhang, Ziyu Guo, Wenchao Chen, Wei Zhang, Hongyang Li, Yu Qiao, Zhongjiang He, and Peng Gao. Referred by multi-modality: A unified temporal transformer for video object segmentation. *arXiv preprint arXiv:2305.16318*, 2023. 1

[105] Jinyu Yang, Zhe Li, Feng Zheng, Ales Leonardis, and Jingkuan Song. Prompting for multi-modal tracking. In *ACMMM*, pages 3492–3500, 2022. 5

[106] 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

[107] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V*, pages 332–348. Springer, 2020. 5

[108] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34: 2491–2502, 2021. 4, 5, 6, 7

[109] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, pages 341–357. Springer, 2022. 3, 5, 6

[110] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 341–357. Springer, 2022. 5, 6

[111] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022. 3

[112] Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, and Bo Dong. Object tracking by jointly exploiting frame and event domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13043–13052, 2021. 3

[113] Jiqing Zhang, Bo Dong, Haiwei Zhang, Jianchuan Ding, Felix Heide, Baocai Yin, and Xin Yang. Spiking transformers for event-based single object tracking. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 8801–8810, 2022. 3

[114] 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. 5

[115] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8886–8895, 2022. 3, 5

[116] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 3

[117] Ziyu Zhang, Sanja Fidler, and Raquel Urtasun. Instance-level segmentation for autonomous driving with deep densely connected mrfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 669–677, 2016. 1

[118] Li Zhou, Zikun Zhou, Kaige Mao, and Zhenyu He. Joint visual grounding and tracking with natural language specification. *arXiv preprint arXiv:2303.12027*, 2023. 5, 7

[119] Xinyu Zhou, Pinxue Guo, Lingyi Hong, Jinglun Li, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. Reading relevant feature from global representation memory for visual object tracking. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[120] Zechu Zhou, Xinyu Zhou, Zhaoyu Chen, Pinxue Guo, Qian-Yu Liu, and Wenqiang Zhang. Memory network with pixel-level spatio-temporal learning for visual object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1

[121] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. *arXiv preprint arXiv:2303.10826*, 2023. 2, 3, 5, 7

[122] 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. *arXiv preprint arXiv:2208.09787*, 2022. 3

[123] 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. *AAAI*, 2023. 5

[124] Yabin Zhu, Chenglong Li, Bin Luo, Jin Tang, and Xiao Wang. Dense feature aggregation and pruning for rgbt tracking. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 465–472, 2019. 5

[125] 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. 5