TGTrack: Temporal Generative Learning for Unified Single Object Tracking

Supplementary Material

The supplementary material is organized into six sections.

- A. More details of TGTrack models.
- B. More details of benchmarks.
- C. More ablation studies.
- D. More comparisons with state-of-the-art methods.
- E. More visualizations.
- F. Broader impacts.

A. More Details of TGTrack Models

In addition to backbone capacity and input resolution, the TGTrack variants also differ in the configuration of the temporal generative decoder. As shown in Tab. 1, the decoder is aligned in both dimensionality and attention complexity with the corresponding Fast-iTPN backbone [65]. Specifically, TGTrack-T224/S224 adopt a decoder width of 384 channels with 6 attention heads, matching the lightweight Fast-iTPN-T/S backbones. TGTrack-B256/B384 increase this to 512 channels and 8 heads, consistent with the Fast-iTPN-B backbone. TGTrack-L256/L384 further expand to 768 channels with 12 heads, aligning with the Fast-iTPN-L backbone. All variants employ 8 transformer blocks to maintain consistent temporal modeling depth.

Table 1. Details of the generative decoder configurations for TGTrack model variants.

Model	Channel Width	Attention Head	Transformer Blocks
TGTrack-T224	384	6	8
TGTrack-S224	384	6	8
TGTrack-B256	512	8	8
TGTrack-B384	512	8	8
TGTrack-L256	768	12	8
TGTrack-L384	768	12	8

B. More Details of Benchmarks

In this section, we provide more detailed descriptions of the benchmarks used for evaluation.

B.1. RGB-based Tracking Benchmarks

LaSOT. LaSOT [17] is a large-scale benchmark specifically designed for long-term visual object tracking. The test set consists of 280 video sequences, each with an average length of over 2,500 frames (approximately 83 seconds). The evaluation metrics include Success Rate (AUC), Precision (P), and Normalized Precision (P_{Norm}).

LaSOT$_{ext}$. LaSOT$_{ext}$ [18] is an extended version of the LaSOT dataset, designed to offer a richer set of samples and more challenging tracking scenarios. It consists of 150 long-term video sequences. The evaluation metrics are the same as those used in LaSOT, including Success Rate (AUC), Precision (P), and Normalized Precision (P_{Norm}).

TrackingNet. TrackingNet [55] is a large-scale short-term tracking benchmark consisting of 30,643 video segments, with an average duration of 16.6 seconds per clip. All videos are sourced from real-world scenarios and cover a wide variety of object types with significant diversity and complexity. The test set includes 511 video sequences, and the evaluation metrics are the same as those used in LaSOT and LaSOT$_{ext}$. The final evaluation results are obtained by uploading the tracking outputs to the official evaluation server.

GOT-10K. GOT-10k [31] is a large-scale and diverse object tracking benchmark containing over 10,000 video sequences. The test set consists of 180 videos, and the evaluation metrics include Average Overlap (AO) and Success Rates at thresholds 0.5 and 0.75 ($SR_{0.75}$ and $SR_{0.75}$). The final evaluation results are obtained by submitting the tracking outputs to the official evaluation server.

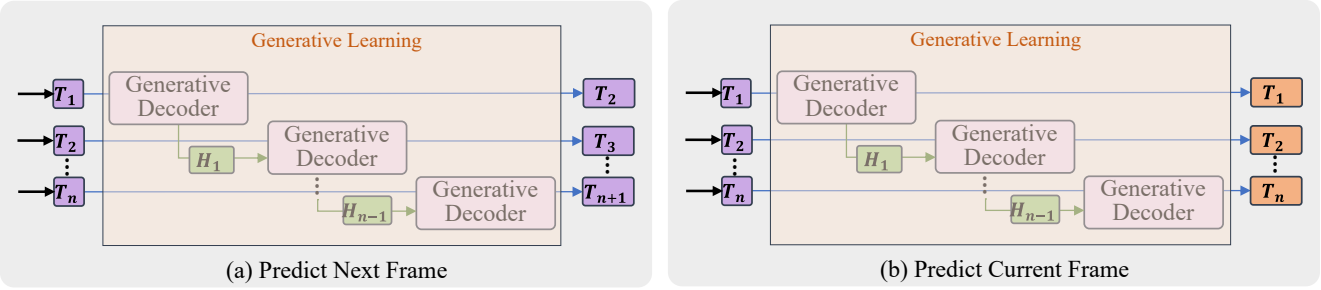
B.2. RGB-Depth Tracking Benchmarks

DepthTrack. DepthTrack [85] is a long-term tracking benchmark that combines visible light and depth data. It is divided into 150 training sequences and 50 testing sequences. Each sequence is accompanied by high-quality visible light images and corresponding depth maps, with 15 scene attributes annotated for each frame. The evaluation metrics include F-score, Precision (P), and Recall (Re), with F-score being the primary metric.

VOT-RGBD22. VOT-RGBD is a subtask in the VOT challenge, focusing on visible RGB+depth object tracking. It covers a wide range of complex scenarios to comprehensively assess the robustness of tracking algorithms. In the 2022 VOT challenge, VOT-RGBD22 [36] consists of 127 video sequences. The evaluation metrics for this dataset include a combined score of Expected Average Overlap (EAO), robustness (Rob.), and accuracy (Acc.).

B.3. RGB-Thermal Tracking Benchmarks

LasHeR. LasHeR [41] is a large-scale RGB-thermal object tracking dataset, consisting of 1,224 video sequences, with 979 sequences used for training and 245 for testing. Each sequence is composed of RGB images and their corresponding thermal infrared (TIR) images. The evaluation metrics include AUC and Precision (P) scores.



(a) Predict Next Frame (b) Predict Current Frame

Figure 1. Comparison between predicting the next frame and predicting the current frame in temporal generative learning.

Table 2. Ablation study on generative decoder width and number of attention heads.

Width & Heads	LaSOT	DepthTrack	LasHeR	VisEvent	TNL2K	Δ
512&8 (Ours)	74.6	65.5	61.7	63.9	65.4	–
384&4	74.5	65.4	61.8	63.9	65.2	**-0.06**
768&16	74.7	65.4	61.4	64.1	65.3	**-0.04**

Table 3. Comparison of TTE with alternative temporal embedding strategies.

Method	LaSOT	DepthTrack	LasHeR	VisEvent	TNL2K	Δ
TTE (Ours)	74.6	65.5	61.7	63.9	65.4	–
ILE	74.4	65.1	61.5	63.6	63.9	**-0.32**
SPE	74.1	65.3	61.8	63.7	65.1	**-0.22**

RGBT234. RGBT234 [39] is an RGB-thermal tracking benchmark consisting of 234 video sequences, with annotations that include 12 scene attributes. The evaluation metrics include Maximum Success Rate (MSR) and Maximum Precision Rate (MPR) scores.

B.4. RGB-Event Tracking Benchmarks

VisEvent. VisEvent [72] is the first large-scale RGB-event dataset for single-object tracking collected from real-world scenarios. It contains 820 video pairs, with 320 sequences used for testing. Each sequence includes visible light data and event stream data captured by an event sensor. The dataset comprises 709 short-term tracking videos and 111 long-term tracking videos, making it suitable for various tracking scenarios. The evaluation metrics include AUC and Precision (P) scores.

B.5. RGB-Language Tracking Benchmarks

TNL2K. TNL2K [71] is a benchmark designed for RGB-Language tracking. It consists of 2,000 video sequences, with 700 sequences used for testing. Each sequence is accompanied by a detailed natural language description of the target, providing information on the object's appearance, location, and behavior. The evaluation metrics include AUC and Precision (P) scores.

OTB99. OTB99 [46] is a small-scale benchmark for RGB-language tracking, derived by supplementing language annotations to the OTB100 dataset [75]. It includes 51 training video sequences and 48 testing video sequences. The evaluation metrics are the same as those used in TNL2K, including AUC and Precision (P) scores.

C. More Ablation Studies

We select TGTrack-B256 as the default variant for ablation studies due to its balanced performance in terms of accuracy and efficiency. In Sec.4.3, we have presented the ablation results of the model components, generative decoder depth, sequence length and generative loss weight. Here, we provide additional ablation results.

C.1. Clarification of PCF

PCF refers to the **P**redict **C**urrent **F**rame strategy used in temporal generative learning, as opposed to the **P**redict **N**ext **F**rame (PNF) strategy adopted in the baseline. As illustrated in Fig. 1, PCF resembles a frame reconstruction task, while our default approach (Predict Next Frame) encourages stronger temporal reasoning across frames.

C.2. Generative Decoder Hyperparameter Analysis

To further understand the behavior of the generative decoder, we conduct additional ablation studies on the decoder hyperparameters. In the main paper, we modified only the decoder depth to ensure a fair comparison with the base-

Table 4. State-of-the-art comparisons of methods utilizing large models across four large-scale RGB benchmarks. The top, second and third-best results are highlighted in **red**, blue and green, respectively.

Method	LaSOT			LaSOT$_{ext}$			TrackingNet			GOT-10k		
	AUC	P$_{Norm}$	P	AUC	P$_{Norm}$	P	AUC	P$_{Norm}$	P	AO	SR$_{0.5}$	SR$_{0.75}$
TGTrack-L384	**76.4**	**86.2**	**84.7**	**55.9**	**67.2**	**63.8**	**88.0**	**92.1**	**89.2**	**81.8**	**91.1**	**83.8**
TGTrack-L256	75.8	86.0	84.2	54.8	65.4	61.3	87.5	91.7	88.1	81.4	90.7	83.7
SUTrack-L384 [9]	75.2	84.9	83.2	53.6	64.2	60.5	87.7	91.7	88.7	81.5	89.5	83.3
ARPTrack-L384 [47]	74.2	83.4	81.7	54.2	64.4	61.2	86.6	91.1	87.4	81.5	90.6	80.5
LoRAT-L224 [49]	74.2	83.6	80.9	52.8	64.7	60.0	85.0	89.5	84.4	75.7	84.9	75.0
ODTrack-L384 [99]	74.0	84.2	82.3	53.9	65.4	61.7	86.1	91.0	86.7	78.2	87.2	77.3
ARTrackV2-L384 [1]	73.6	82.8	81.1	53.4	63.7	60.2	86.1	90.4	86.2	79.5	87.8	79.6
SUTrack-L224 [9]	73.5	83.3	80.9	54.0	65.3	61.7	86.5	90.9	86.7	81.0	90.4	82.4
ARTrack-L384 [73]	73.1	82.2	80.3	52.8	62.9	59.7	85.6	89.6	86.0	78.5	87.4	77.8
SeqTrack-L384 [7]	72.5	81.5	79.3	50.7	61.6	57.5	85.5	89.8	85.8	74.8	81.9	72.2
MixViT-L384 [12]	72.4	82.2	80.1	-	-	-	85.4	90.2	85.7	75.7	85.3	75.1
UNINEXT-L [84]	72.4	80.7	78.9	54.4	61.8	61.4	85.1	88.2	84.7	-	-	-
GRM-L320 [24]	71.4	81.2	77.9	-	-	-	84.4	88.9	84.0	-	-	-
TATrack-L384 [26]	71.1	79.1	76.1	-	-	-	85.0	89.3	84.5	-	-	-
SimTrack-L224 [5]	70.5	79.7	-	-	-	-	83.4	87.4	-	69.8	78.8	66.0
Mixformer-L320 [11]	70.1	79.9	76.3	-	-	-	83.9	88.9	83.1	-	-	-
CTTrack-L320 [62]	69.8	79.7	76.2	-	-	-	84.9	89.1	83.5	72.8	81.3	71.5
Unicorn [83]	68.5	-	-	-	-	-	83.0	86.4	82.2	-	-	-

line. Here, we additionally explore the effect of varying the embedding width and the number of attention heads.

For these experiments, we evaluate TGTrack-B256 under three configurations by scaling the decoder width and attention heads around our default setting (width = 512, heads = 8). As shown in Tab. 2, adjusting either width or number of heads results in only marginal differences in performance, while larger configurations introduce additional computational overhead. These results indicate that the generative decoder is not sensitive to these hyperparameters, further validating the robustness of our design.

C.3. Time Token Embedding Designs

We further evaluate alternative temporal embedding strategies to better contextualize our proposed Time Token Embedding (TTE). Specifically, we compare TTE with (1) Independent Learnable Embeddings (ILE) and (2) standard sinusoidal positional encodings (SPE). Tab. 3 shows that TTE outperforms both ILE and SPE, demonstrating its effectiveness in capturing temporal structure. These results confirm that encoding temporal position through a polar formulation provides richer temporal cues for the generative decoder and improves unified tracking performance.

D. More Comparison with SOTA Methods

In Sec.4 of the main manuscript, we primarily compare TGTrack with recent high-performing methods. Here, we provide additional comparisons with earlier state-of-the-art trackers to further demonstrate the effectiveness and robustness of our proposed framework, as shown in Tab. 6, Tab. 4, Tab. 7 and Tab. 5. For clearer comparison, the previous state-of-the-art trackers in Tab. 7 and Tab. 5 are grouped into two categories: **unified** multi-modal trackers and **modality-specific** trackers.

Table 5. SOTA comparisons on RGB-Language tracking.

	Method	TNL2K		OTB99	
		AUC	P	AUC	P
Ours	**TGTrack-L384**	**68.6**	**72.8**	72.9	**96.3**
	TGTrack-L256	67.7	71.7	72.5	95.1
	TGTrack-B384	66.7	70.8	70.9	94.6
	TGTrack-B256	65.4	68.7	72.5	95.2
	TGTrack-S224	64.7	67.5	69.9	92.7
	TGTrack-T224	63.0	65.3	69.8	93.3
Unified	SUTrack-L384 [9]	67.9	72.1	71.2	93.1
	SUTrack-L224 [9]	66.7	70.3	72.7	94.4
	SUTrack-B384 [9]	65.6	69.3	69.7	91.2
	SUTrack-B224 [9]	65.0	67.9	70.8	93.4
	SUTrack-T224 [9]	60.9	62.3	67.4	88.6
	SeqTrackv2-L384 [8]	62.4	66.1	71.4	93.6
	SeqTrackv2-B256 [8]	57.5	59.7	71.2	93.9
	OneTracker [27]	58.0	59.1	69.7	91.5
Modality-Specific	ChatTracker-L [63]	65.4	70.2	-	-
	DUTrack-256 [44]	64.9	70.6	70.9	93.9
	MMTrack [98]	58.6	59.4	70.5	91.8
	JointNLT [100]	56.9	58.1	65.3	85.6
	DecoupleTNL [52]	56.7	56.0	**73.8**	94.8
	TransVLT [97]	56.0	-	69.9	91.2
	CapsuleTNL [51]	-	-	71.1	92.4
	CTRTNL [45]	44.0	45.0	69.0	91.0
	TNL2K-II [71]	42.0	42.0	68.0	88.0
	SNLT [21]	27.6	41.9	66.6	80.4
	GTI [88]	-	-	58.1	73.2
	TransVG [16]	26.1	28.9	-	-
	Feng *et al.* [19]	25.0	27.0	67.0	73.0
	RTTNLD [20]	25.0	27.0	61.0	79.0
	Wang *et al.* [70]	-	-	65.8	89.1
	TNLS [46]	-	-	55.3	72.3
	OneStage-BERT [87]	19.8	-	24.6	32.2
	LBYL-BERT [87]	18.3	-	20.7	26.0

Table 6. State-of-the-art comparisons of methods utilizing base models across four large-scale RGB benchmarks. The top, second and third-best results are highlighted in **red**, blue and green, respectively.

Method	LaSOT			LaSOT$_{ext}$			TrackingNet			GOT-10k		
	AUC	P$_{Norm}$	P	AUC	P$_{Norm}$	P	AUC	P$_{Norm}$	P	AO	SR$_{0.5}$	SR$_{0.75}$
TGTrack-B384	**75.3**	**84.9**	**83.1**	**54.8**	**65.7**	**61.6**	**87.5**	**91.6**	**87.9**	79.8	88.5	80.9
TGTrack-B256	74.6	84.8	82.4	53.2	64.2	60.4	86.2	90.6	85.9	**80.8**	91.1	**81.6**
TGTrack-S224	72.9	83.0	79.8	52.4	63.4	59.6	85.3	89.7	84.5	77.2	86.8	77.6
SUTrack-B384 [9]	74.4	83.9	81.9	52.9	63.6	60.1	86.5	90.7	86.8	79.3	88.0	80.0
MambaLCT-B384 [43]	73.6	84.1	81.6	53.3	64.8	61.4	85.2	89.8	85.2	76.2	86.7	74.3
LMTrack-B384 [81]	73.2	83.4	81.0	53.6	64.7	61.5	85.7	89.9	84.7	80.1	**91.5**	79.0
SUTrack-B224 [9]	73.2	83.4	80.5	53.1	64.2	60.5	85.7	90.3	85.1	77.9	87.5	78.5
ODTrack-B384 [99]	73.2	83.2	80.6	52.4	63.9	60.1	85.1	90.1	84.9	77.0	87.9	75.1
LoRAT-B378 [49]	72.9	81.9	79.1	53.1	64.8	60.6	84.2	88.4	83.0	73.7	82.6	72.9
ARPTrack-B256 [47]	72.6	81.4	78.5	52.0	62.9	58.7	85.5	90.0	85.3	77.7	87.3	74.3
DropTrack-224 [74]	71.8	81.8	78.1	52.7	63.9	60.2	-	-	-	75.9	86.8	72.0
SeqTrack-B384 [7]	71.5	81.1	77.8	50.5	61.6	57.5	83.9	88.8	83.6	74.5	84.3	71.4
ROMTrack-384 [3]	71.4	81.4	78.2	51.3	62.4	58.6	84.1	89.0	83.7	74.2	84.3	72.4
AQATrack-256 [80]	71.4	81.9	78.6	51.2	62.2	58.9	83.8	88.6	83.1	73.8	83.2	72.1
SwinTrack [48]	71.3	-	76.5	49.1	-	55.6	84.0	-	82.8	72.4	-	67.8
OSTrack-384 [89]	71.1	81.1	77.6	50.5	61.3	57.6	83.9	88.5	83.2	73.7	83.2	70.8
VideoTrack-256 [79]	70.2	-	76.4	-	-	-	83.8	88.7	83.1	72.9	81.9	69.8
OneTracker-384 [27]	70.5	79.9	76.5	-	-	-	83.7	88.4	82.7	-	-	-
EVPTrack-224 [59]	70.4	80.9	77.2	48.7	59.5	55.1	83.5	88.3	-	73.3	83.6	70.7
GRM-B256 [24]	69.9	79.3	75.8	-	-	-	84.0	88.7	83.3	73.4	82.9	70.4
LMTrack-B256 [81]	69.8	79.2	76.3	49.0	59.6	55.8	84.2	89.0	82.8	76.3	87.1	73.9
CiteTracker-384 [42]	69.7	78.6	75.7	-	-	-	84.5	89.0	84.2	74.7	84.3	73.0
RTS [57]	69.7	76.2	73.7	-	-	-	81.6	86.0	79.4	-	-	-
MixViT-288 [12]	69.6	79.9	75.9	-	-	-	83.5	88.3	83.5	72.5	82.4	69.9
TATrack-B224 [26]	69.4	78.2	74.1	-	-	-	83.5	88.3	81.8	73.0	83.3	68.5
SimTrack-B224 [5]	69.3	78.5	-	-	-	-	82.3	86.5	-	68.6	78.9	62.4
Mixformer-22k [11]	69.2	78.7	74.7	-	-	-	83.1	88.1	81.6	70.7	80.0	67.8
AiATrack [23]	69.0	79.4	73.8	47.7	55.6	55.4	82.7	87.8	80.4	69.6	63.2	80.0
ToMP [54]	68.5	79.2	73.5	45.9	-	-	81.5	86.4	78.9	-	-	-
CTTrack-B320 [62]	67.8	77.8	74.0	-	-	-	82.5	87.1	80.3	71.3	80.7	70.3
KeepTrack [53]	67.1	77.2	70.2	48.2	-	-	-	-	-	-	-	-
STARK [82]	67.1	77.0	-	-	-	-	82.0	86.9	-	68.8	78.1	64.1
SLT [32]	66.8	75.5	-	-	-	-	82.8	87.5	81.4	67.5	76.5	60.3
SBT [78]	66.7	-	71.1	-	-	-	-	-	-	70.4	80.8	64.7
CSWinTT [61]	66.2	75.2	70.9	-	-	-	81.9	86.7	79.5	69.4	78.9	65.4
TransT [6]	64.9	73.8	69.0	-	-	-	81.4	86.7	80.3	67.1	76.8	60.9
SiamR-CNN [66]	64.8	72.2	-	-	-	-	81.2	85.4	80.0	64.9	72.8	59.7
TrDiMP [68]	63.9	-	61.4	-	-	-	78.4	83.3	73.1	68.8	80.5	59.7
AutoMatch [96]	58.3	-	59.9	-	-	-	76.0	-	72.6	65.2	76.6	54.3
DiMP [2]	56.9	65.0	56.7	39.2	47.6	45.1	74.0	80.1	68.7	61.1	71.7	49.2
Ocean [95]	56.0	65.1	56.6	-	-	-	-	-	-	61.1	72.1	47.3
SiamAttn [90]	56.0	64.8	-	-	-	-	75.2	81.7	-	-	-	-
ATOM [14]	51.5	57.6	50.5	37.6	45.9	43.0	70.3	77.1	64.8	55.6	63.4	40.2
SiamBAN [10]	51.4	59.8	-	-	-	-	-	-	-	-	-	-
SiamPRN++ [37]	49.6	56.9	49.1	34.0	41.6	39.6	73.3	80.0	69.4	51.7	61.6	32.5
MDNet [56]	39.7	46.0	37.3	27.9	34.9	31.8	60.6	70.5	56.5	29.9	30.3	9.9

Table 7. State-of-the-art comparisons on RGB-Depth, RGB-Thermal, and RGB-Event tracking. The best three results are highlighted in **red**, blue and green, respectively.

	Method	VOT-RGBD22			DepthTrack			LasHeR		RGBT234		VisEvent	
		EAO	Acc.	Rob.	F-score	Re	Pr	AUC	P	MSR	MPR	AUC	P
Ours	**TGTrack-L384**	**77.8**	83.3	93.2	**67.5**	**67.4**	**67.6**	**63.3**	**79.5**	**71.3**	**95.2**	**65.7**	**83.0**
	TGTrack-L256	77.7	83.2	93.0	66.8	66.6	67.0	62.7	78.7	71.1	94.9	65.1	82.8
	TGTrack-B384	77.5	82.9	93.2	66.0	66.1	65.9	62.8	78.9	69.5	93.1	64.7	81.6
	TGTrack-B256	77.4	82.9	93.0	65.5	65.7	65.3	61.7	77.4	69.9	93.7	63.9	81.4
	TGTrack-S224	76.6	82.2	92.4	63.4	63.3	63.5	59.4	74.5	68.5	91.4	63.1	80.6
	TGTrack-T224	76.0	81.3	91.7	63.3	63.6	63.0	58.2	73.1	67.0	90.0	62.2	80.4
Unified	STTrack [29]	77.6	82.5	**93.7**	63.3	63.4	63.2	60.3	76.0	66.7	89.8	61.9	78.6
	CSTrack [22]	77.4	83.3	92.9	65.8	66.4	65.2	60.8	75.6	70.9	94.0	65.2	82.4
	APTrack [30]	77.4	82.1	93.4	62.1	61.9	62.3	58.9	74.1	-	-	61.8	78.5
	SUTrack-L384 [9]	76.6	83.5	92.2	66.4	66.4	66.5	61.9	76.9	70.3	93.7	63.8	80.5
	SUTrack-B384 [9]	76.6	**83.9**	91.4	64.4	64.2	64.6	60.9	75.8	69.2	92.1	63.4	79.8
	SUTrack-B224 [9]	76.5	82.8	91.8	65.1	65.7	64.5	59.9	74.5	69.5	92.2	62.7	79.9
	SUTrack-L224 [9]	76.4	83.4	91.9	64.3	64.6	64.0	61.9	77.0	70.8	94.6	64.0	80.9
	SMSTracker [4]	74.8	82.2	89.7	63.6	63.1	64.1	56.0	70.3	64.5	86.9	60.4	76.3
	SeqTrackv2-L384 [8]	74.8	82.6	91.0	62.3	62.6	62.5	61.0	76.7	68.0	91.3	63.4	80.0
	SeqTrackv2-B256 [8]	74.4	81.5	91.0	63.2	63.4	62.9	55.8	70.4	64.7	88.0	61.2	78.2
	XTrack-L [64]	74.0	82.8	88.9	64.8	64.3	65.4	58.7	73.1	65.4	87.8	63.3	80.5
	SDSTrack [28]	72.8	81.2	88.3	61.9	60.9	61.4	53.1	66.5	62.5	84.8	59.7	76.7
	OneTracker [27]	72.7	81.9	87.2	60.9	60.4	60.7	53.8	67.2	64.2	85.7	60.8	76.7
	ViPT [101]	72.1	81.5	87.1	59.4	59.6	59.2	52.5	65.1	61.7	83.5	59.2	75.8
	Un-Track [76]	71.8	82.0	86.4	61.0	61.0	61.0	51.3	64.6	62.5	84.2	58.9	75.5
	SUTrack-T224 [9]	68.1	81.0	83.9	61.7	62.1	61.2	53.9	66.7	63.8	85.9	58.8	75.7
	OSTrack [89]	67.6	80.3	83.3	52.9	52.2	53.6	41.2	51.5	54.9	72.9	53.4	69.5
	ProTrack [86]	65.1	80.1	80.2	57.8	57.3	58.3	42.0	53.8	59.9	79.5	47.1	63.2
Modality-Specific	SBT-RGBD [78]	70.8	80.9	86.4	-	-	-	-	-	-	-	-	-
	DMTrack [36]	65.8	75.8	85.1	-	-	-	-	-	-	-	-	-
	DeT [85]	65.7	76.0	84.5	53.2	50.6	56.0	-	-	-	-	-	-
	SPT [102]	65.1	79.8	85.1	53.8	54.9	52.7	-	-	-	-	-	-
	STARK [82]	64.7	80.3	79.8	-	-	-	36.1	44.9	-	-	44.6	61.2
	KeepTrack [53]	60.6	75.3	79.7	-	-	-	-	-	-	-	-	-
	DRefine [35]	59.2	77.5	76.0	-	-	-	-	-	-	-	-	-
	DDiMP [34]	-	-	-	48.5	56.9	50.3	-	-	-	-	-	-
	ATCAIS [34]	55.9	76.1	73.9	47.6	45.5	50.0	-	-	-	-	-	-
	DiMP [2]	54.3	70.3	73.1	-	-	-	-	-	-	-	-	-
	ATOM [14]	50.5	59.8	68.8	-	-	-	-	-	-	-	41.2	60.8
	LTMU [13]	-	-	-	46.0	41.7	51.2	-	-	-	-	45.9	65.5
	GLGS-D [34]	-	-	-	45.3	36.9	58.4	-	-	-	-	-	-
	DAL [58]	-	-	-	42.9	36.9	51.2	-	-	-	-	-	-
	LTDSEd [33]	-	-	-	40.5	38.2	43.0	-	-	-	-	-	-
	Siam-LTD [34]	-	-	-	37.6	34.2	41.8	-	-	-	-	-	-
	SiamM-Ds [33]	-	-	-	33.6	26.4	46.3	-	-	-	-	-	-
	CA3DMS [50]	-	-	-	22.3	22.8	21.8	-	-	-	-	-	-
	TransT [6]	-	-	-	-	-	-	39.4	52.4	-	-	47.4	65.0
	APFNet [77]	-	-	-	-	-	-	36.2	50.0	57.9	82.7	-	-
	CMPP [67]	-	-	-	-	-	-	-	-	57.5	82.3	-	-
	JMMAC [93]	-	-	-	-	-	-	-	-	57.3	79.0	-	-
	MaCNet [91]	-	-	-	-	-	-	-	-	55.4	79.0	-	-
	DAFNet [25]	-	-	-	-	-	-	-	-	54.4	79.6	-	-
	mfDiMP [92]	-	-	-	-	-	-	34.3	44.7	42.8	64.6	-	-
	DAPNet [103]	-	-	-	-	-	-	31.4	43.1	-	-	-	-
	CAT [40]	-	-	-	-	-	-	31.4	45.0	56.1	80.4	-	-
	HMFT [94]	-	-	-	-	-	-	31.3	43.6	-	-	-	-
	FANet [104]	-	-	-	-	-	-	30.9	44.1	55.3	78.7	-	-
	SGT [38]	-	-	-	-	-	-	25.1	36.5	47.2	72.0	-	-
	SiamRCNN$_E$ [66]	-	-	-	-	-	-	-	-	-	-	49.9	65.9
	PrDiMP$_E$ [15]	-	-	-	-	-	-	-	-	-	-	45.3	64.4
	MDNet$_E$ [56]	-	-	-	-	-	-	-	-	-	-	42.6	66.1
	SiamCar$_E$ [82]	-	-	-	-	-	-	-	-	-	-	42.0	59.9
	VITAL$_E$ [60]	-	-	-	-	-	-	-	-	-	-	41.5	64.9
	SiamBAN$_E$ [10]	-	-	-	-	-	-	-	-	-	-	40.5	59.1
	SiamMask$_E$ [69]	-	-	-	-	-	-	-	-	-	-	36.9	56.2

E. More Visualizations

E.1. Qualitative Results on RGB Trackers

Additionally, to provide a more intuitive understanding of tracking performance across various challenging scenarios, we present qualitative results for visual comparisons. Fig. 2 shows representative tracking examples on LaSOT, which highlight the robustness of TGTrack under challenging conditions. As illustrated, TGTrack is able to accurately localize the target even in the presence of low-resolution imagery, severe occlusion, distractor objects with similar appearance, and significant target deformation. Compared to previous SOTA methods (ARTrackV2 [1], ODTrack [99], OSTrack [89]), TGTrack exhibits superior robustness and precision, maintaining stable tracking under complex and dynamic scenarios.

E.2. Qualitative Results on Unified Trackers

For the remaining modalities, including RGB-Depth, RGB-Thermal, and RGB-Event, we provide qualitative comparisons between TGTrack and three representative multi-modal unified trackers (Un-Track [76], SDSTrack [28],STTrack [29]). As shown in Fig. 3, TGTrack consistently delivers more accurate and robust tracking results across various challenging scenarios. In RGB-Depth sequences, where the presence of visually similar objects introduces strong interference and the target undergoes significant directional motion, TGTrack demonstrates stable tracking performance thanks to its temporal modeling capability. In RGB-Thermal scenarios, despite low illumination and substantial target deformation, TGTrack accurately localizes the target by leveraging the complementary advantages of RGB, thermal and temporal cues. In RGB-Event sequences, where the target moves rapidly across the camera's field of view, our model effectively adapts to these dynamic changes over time, maintaining precise tracking.

F. Broader Impacts

This work proposes a unified framework for multi-modal single object tracking with improved temporal modeling, which can benefit applications such as autonomous driving, surveillance, and assistive robotics by enhancing tracking performance under challenging scenarios. At the same time, stronger tracking capabilities may raise privacy concerns, especially if misused in surveillance systems. We encourage the responsible use of this technology, with attention to ethical deployment, data protection, and transparency. With proper safeguards, this research can contribute positively to reliable and human-centered AI systems.

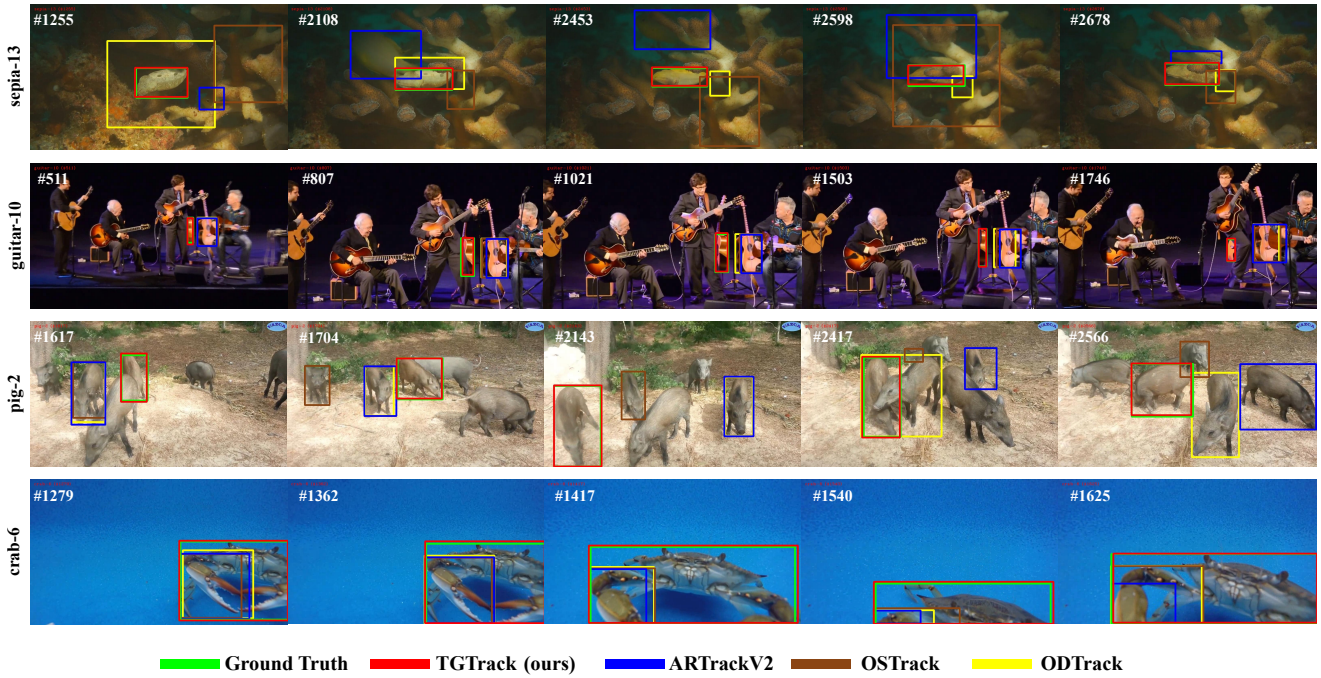


Figure 2. Qualitative comparison between TGTrack and three RGB trackers on LaSOT. The four rows illustrate representative challenging scenarios, including low-resolution imagery, severe occlusion, visually similar distractors, and significant target deformation. TGTrack effectively handles these challenges through its temporal generative learning mechanism.

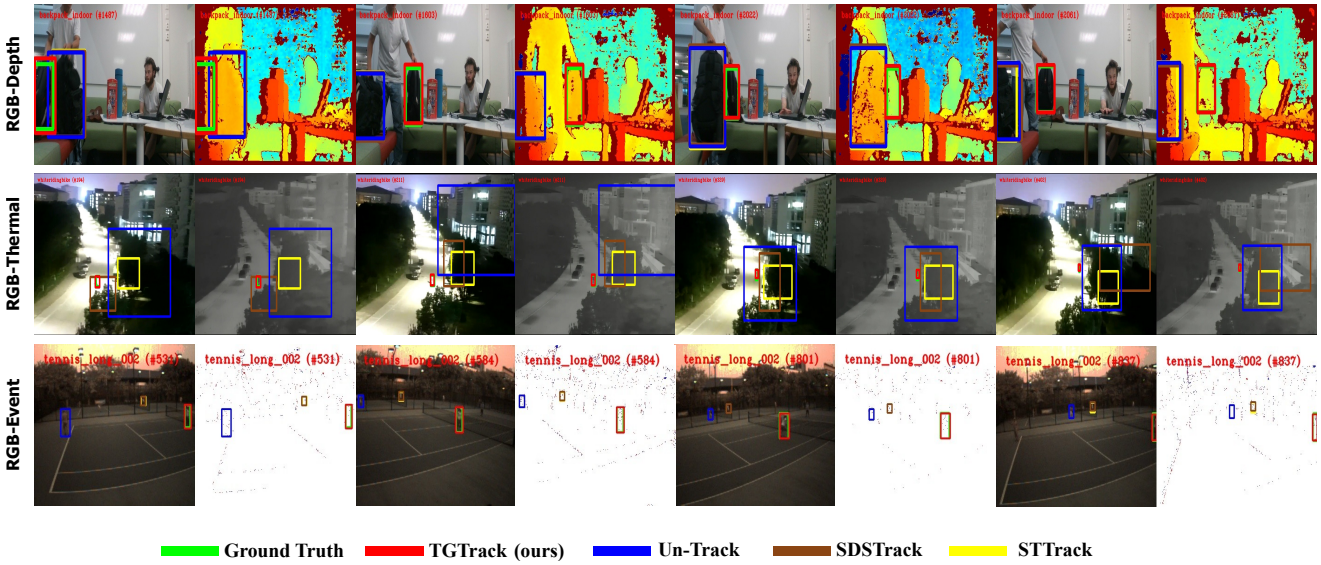


Figure 3. Qualitative comparison between TGTrack and other three multi-modal trackers on three multi-modal task. The three sequences correspond to scenarios involving similar object interference, low illumination, and fast motion. TGTrack effectively addresses these challenges through temporal generative learning.

References

[1] Yifan Bai, Zeyang Zhao, Yihong Gong, and Xing Wei. ARTrackV2: Prompting autoregressive tracker where to look and how to describe. In *CVPR*, pages 19048–19057, 2024. 3, 6

[2] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *ICCV*, 2019. 4, 5

[3] Yidong Cai, Jie Liu, Jie Tang, and Gangshan Wu. Robust object modeling for visual tracking. In *ICCV*, pages 9589–9600, 2023. 4

[4] Sixian Chan, Zedong Li, Wenhao Li, Shijian Lu, Chunhua Shen, and Xiaoqin Zhang. Smstracker: Tri-path score mask sigma fusion for multi-modal tracking. In *ICCV*, pages 4766–4775, 2025. 5

[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 *ECCV*, pages 375–392, 2022. 3, 4

[6] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *CVPR*, pages 8126–8135, 2021. 4, 5

[7] Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. Seqtrack: Sequence to sequence learning for visual object tracking. In *CVPR*, pages 14572–14581, 2023. 3, 4

[8] Xin Chen, Ben Kang, Jiawen Zhu, Dong Wang, Houwen Peng, and Huchuan Lu. Unified sequence-to-sequence learning for single- and multi-modal visual object tracking. *arXiv preprint arXiv:2304.14394*, 2024. 3, 5

[9] Xin Chen, Ben Kang, Wanting Geng, Jiawen Zhu, Yi Liu, Dong Wang, and Huchuan Lu. Sutrack: Towards simple and unified single object tracking. In *AAAI*, pages 2239–2247, 2025. 3, 4, 5

[10] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *CVPR*, 2020. 4, 5

[11] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *CVPR*, pages 13608–13618, 2022. 3, 4

[12] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. *IEEE TPAMI*, pages 0–18, 2024. 3, 4

[13] Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. High-performance long-term tracking with meta-updater. In *CVPR*, 2020. 5

[14] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ATOM: Accurate tracking by overlap maximization. In *CVPR*, 2019. 4, 5

[15] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *CVPR*, 2020. 5

[16] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. TransVG: End-to-end visual grounding with transformers. In *ICCV*, 2021. 3

[17] 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*, pages 5374–5383, 2019. 1

[18] Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Mingzhen Huang, Juehuan Liu, Yong Xu, et al. LaSOT: A high-quality large-scale single object tracking benchmark. *IJCV*, pages 439–461, 2021. 1

[19] Qi Feng, Vitaly Ablavsky, Qinxun Bai, and Stan Sclaroff. Robust visual object tracking with natural language region proposal network. *arXiv preprint arXiv:1912.02048*, 2019. 3

[20] Qi Feng, Vitaly Ablavsky, Qinxun Bai, Guorong Li, and Stan Sclaroff. Real-time visual object tracking with natural language description. In *WACV*, 2020. 3

[21] Qi Feng, Vitaly Ablavsky, Qinxun Bai, and Stan Sclaroff. Siamese natural language tracker: Tracking by natural language descriptions with siamese trackers. In *CVPR*, 2021. 3

[22] Xiaokun Feng, Dailing Zhang, Shiyu Hu, Xuchen Li, Meiqi Wu, Jing Zhang, Xiaotang Chen, and Kaiqi Huang. CSTrack: Enhancing RGB-x tracking via compact spatiotemporal features. In *ICML*, 2025. 5

[23] Shenyuan Gao, Chunluan Zhou, Chao Ma, Xinggang Wang, and Junsong Yuan. AiATrack: Attention in attention for transformer visual tracking. In *ECCV*, 2022. 4

[24] Shenyuan Gao, Chunluan Zhou, and Jun Zhang. Generalized relation modeling for transformer tracking. In *CVPR*, 2023. 3, 4

[25] Yuan Gao, Chenglong Li, Yabin Zhu, Jin Tang, Tao He, and Futian Wang. Deep adaptive fusion network for high performance RGBT tracking. In *ICCVW*, pages 1–8, 2019. 5

[26] Kaijie He, Canlong Zhang, Sheng Xie, Zhixin Li, and Zhiwen Wang. Target-aware tracking with long-term context attention. In *AAAI*, pages 773–780, 2023. 3, 4

[27] Lingyi Hong, Shilin Yan, Renrui Zhang, Wanyun Li, Xinyu Zhou, Pinxue Guo, Kaixun Jiang, Yiting Chen, Jinglun Li, Zhaoyu Chen, and Wenqiang Zhang. Onetracker: Unifying visual object tracking with foundation models and efficient tuning. In *CVPR*, pages 19079–19091, 2024. 3, 4, 5

[28] Xiaojun Hou, Jiazheng Xing, Yijie Qian, Yaowei Guo, Shuo Xin, Junhao Chen, Kai Tang, Mengmeng Wang, Zhengkai Jiang, Liang Liu, and Yong Liu. Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking. In *CVPR*, pages 26551–26561, 2024. 5, 6

[29] Xiantao Hu, Ying Tai, Xu Zhao, Chen Zhao, Zhenyu Zhang, Jun Li, Bineng Zhong, and Jian Yang. Exploiting multimodal spatial-temporal patterns for video object tracking. In *AAAI*, pages 3581–3589, 2025. 5, 6

[30] Xiantao Hu, Bineng Zhong, Qihua Liang, Liangtao Shi, Zhiyi Mo, Ying Tai, and Jian Yang. Adaptive perception for unified visual multi-modal object tracking. *IEEE TAI*, 2025. 5

[31] Lianghua Huang, Xin Zhao, and Kaiqi Huang. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE TPAMI*, pages 1562–1577, 2019. 1

[32] Minji Kim, Seungkwan Lee, Jungseul Ok, Bohyung Han, and Minsu Cho. Towards sequence-level training for visual tracking. In *ECCV*, 2022. 4

[33] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Luka Cehovin Zajc, Ondrej Drbohlav, Alan Lukezic, Amanda Berg, et al. The seventh visual object tracking VOT2019 challenge results. In *ICCVW*, 2019. 5

[34] 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 *ECCV*, 2020. 5

[35] 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*, 2021. 5

[36] 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*, 2023. 1, 5

[37] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. SiamRPN++: Evolution of siamese visual tracking with very deep networks. In *CVPR*, 2019. 4

[38] 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

[39] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. RGB-T object tracking: Benchmark and baseline. *PR*, page 106977, 2019. 2

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

[41] 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 TIP*, pages 392–404, 2021. 1

[42] Xin Li, Yuqing Huang, Zhenyu He, Yaowei Wang, Huchuan Lu, and Ming-Hsuan Yang. CiteTracker: Correlating image and text for visual tracking. In *ICCV*, pages 9974–9983, 2023. 4

[43] Xiaohai Li, Bineng Zhong, Qihua Liang, Guorong Li, Zhiyi Mo, and Shuxiang Song. Mambalct: Boosting tracking via long-term context state space model. In *AAAI*, pages 4986–4994, 2025. 4

[44] Xiaohai Li, Bineng Zhong, Qihua Liang, Zhiyi Mo, Jian Nong, and Shuxiang Song. Dynamic updates for language adaptation in visual-language tracking. In *CVPR*, pages 19165–19174, 2025. 3

[45] Yihao Li, Jun Yu, Zhongpeng Cai, and Yuwen Pan. Cross-modal target retrieval for tracking by natural language. In *CVPR*, pages 4931–4940, 2022. 3

[46] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees G. M. Snoek, and Arnold W. M. Smeulders. Tracking by natural language specification. In *CVPR*, pages 6495–6503, 2017. 2, 3

[47] Shiyi Liang, Yifan Bai, Yihong Gong, and Xing Wei. Autoregressive sequential pretraining for visual tracking. In *CVPR*, pages 7254–7264, 2025. 3, 4

[48] Liting Lin, Heng Fan, Yong Xu, and Haibin Ling. SwinTrack: A simple and strong baseline for transformer tracking. In *NeurIPS*, 2022. 4

[49] Liting Lin, Heng Fan, Zhipeng Zhang, Yaowei Wang, Yong Xu, and Haibin Ling. Tracking meets lora: Faster training, larger model, stronger performance. In *ECCV*, pages 300–318, 2024. 3, 4

[50] 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 TMM*, pages 664–677, 2018. 5

[51] Ding Ma and Xiangqian Wu. Capsule-based object tracking with natural language specification. In *ACM MM*, 2021. 3

[52] Ding Ma and Xiangqian Wu. Tracking by natural language specification with long short-term context decoupling. In *ICCV*, pages 14012–14021, 2023. 3

[53] Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Learning target candidate association to keep track of what not to track. In *ICCV*, pages 13444–13454, 2021. 4, 5

[54] Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool. Transforming model prediction for tracking. In *CVPR*, 2022. 4

[55] 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*, pages 300–317, 2018. 1

[56] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *CVPR*, 2016. 4, 5

[57] Matthieu Paul, Martin Danelljan, Christoph Mayer, and Luc Van Gool. Robust visual tracking by segmentation. In *ECCV*, 2022. 4

[58] Yanlin Qian, Song Yan, Alan Lukežič, Matej Kristan, Joni-Kristian Kämäräinen, and Jiří Matas. Dal: A deep depth-aware long-term tracker. In *ICPR*, pages 7825–7832, 2021. 5

[59] Liangtao Shi, Bineng Zhong, Qihua Liang, Ning Li, Shengping Zhang, and Xianxian Li. Explicit visual prompts for visual object tracking. In *AAAI*, pages 4838–4846, 2024. 4

[60] Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, Wangmeng Zuo, Chunhua Shen, Rynson W.H. Lau, and Ming-Hsuan Yang. VITAL: Visual tracking via adversarial learning. In *CVPR*, 2018. 5

[61] Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Transformer tracking with cyclic shifting window attention. In *CVPR*, 2022. 4

[62] Zikai Song, Run Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Compact transformer tracker with correlative masked modeling. In *AAAI*, pages 2321–2329, 2023. 3, 4

[63] Yiming Sun, Fan Yu, Shaoxiang Chen, Yu Zhang, Junwei Huang, Yang Li, Chenhui Li, and Changbo Wang. Chattracker: Enhancing visual tracking performance via chatting with multimodal large language model. In *NeurIPS*, 2024. 3

[64] Yuedong Tan, Zongwei Wu, Yuqian Fu, Zhuyun Zhou, Guolei Sun, Eduard Zamfi, Chao Ma, Danda Pani Paudel, Luc Van Gool, and Radu Timofte. Xtrack: Multimodal training boosts rgb-x video object trackers. In *ICCV*, 2025. 5

[65] Yunjie Tian, Lingxi Xie, Jihao Qiu, Jianbin Jiao, Yaowei Wang, Qi Tian, and Qixiang Ye. Fast-itpn: Integrally pretrained transformer pyramid network with token migration. *IEEE TPAMI*, pages 1–15, 2024. 1

[66] Paul Voigtlaender, Jonathon Luiten, Philip H. S. Torr, and Bastian Leibe. Siam R-CNN: Visual tracking by redetection. In *CVPR*, 2020. 4, 5

[67] Chaoqun Wang, Chunyan Xu, Zhen Cui, Ling Zhou, Tong Zhang, Xiaoya Zhang, and Jian Yang. Cross-modal pattern-propagation for RGB-T tracking. In *CVPR*, pages 7064–7073, 2020. 5

[68] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *CVPR*, 2021. 4

[69] 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

[70] Xiao Wang, Chenglong Li, Rui Yang, Tianzhu Zhang, Jin Tang, and Bin Luo. Describe and attend to track: Learning natural language guided structural representation and visual attention for object tracking. *arXiv preprint arXiv:1811.10014*, 2018. 3

[71] 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 *CVPR*, pages 13763–13773, 2021. 2, 3

[72] 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 TCYB*, pages 1997–2010, 2024. 2

[73] Xing Wei, Yifan Bai, Yongchao Zheng, Dahu Shi, and Yihong Gong. Autoregressive visual tracking. In *CVPR*, pages 9697–9706, 2023. 3

[74] Qiangqiang Wu, Tianyu Yang, Ziquan Liu, Baoyuan Wu, Ying Shan, and Antoni B Chan. DropMAE: Masked autoencoders with spatial-attention dropout for tracking tasks. In *CVPR*, pages 14561–14571, 2023. 4

[75] Yi Wu, Jongwoo Lim, and Ming Hsuan Yang. Object tracking benchmark. *IEEE TPAMI*, 2013. 2

[76] Zongwei Wu, Jilai Zheng, Xiangxuan Ren, Florin-Alexandru Vasluianu, Chao Ma, Danda Pani Paudel, Luc Van Gool, and Radu Timofte. Single-model and any-modality for video object tracking. In *CVPR*, pages 19156–19166, 2024. 5, 6

[77] Yun Xiao, Mengmeng Yang, Chenglong Li, Lei Liu, and Jin Tang. Attribute-based progressive fusion network for RGBT tracking. In *AAAI*, pages 2831–2838, 2022. 5

[78] Fei Xie, Chunyu Wang, Guangting Wang, Yue Cao, Wankou Yang, and Wenjun Zeng. Correlation-aware deep tracking. In *CVPR*, 2022. 4, 5

[79] Fei Xie, Lei Chu, Jiahao Li, Yan Lu, and Chao Ma. VideoTrack: Learning to track objects via video transformer. In *CVPR*, pages 22826–22835, 2023. 4

[80] Jinxia Xie, Bineng Zhong, Zhiyi Mo, Shengping Zhang, Liangtao Shi, Shuxiang Song, and Rongrong Ji. Autoregressive queries for adaptive tracking with spatio-temporal transformers. In *CVPR*, pages 19300–19309, 2024. 4

[81] Chenlong Xu, Bineng Zhong, Qihua Liang, Yaozong Zheng, Guorong Li, and Shuxiang Song. Less is more: Token context-aware learning for object tracking. In *AAAI*, pages 8824–8832, 2025. 4

[82] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *ICCV*, pages 10448–10457, 2021. 4, 5

[83] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *ECCV*, 2022. 3

[84] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, 2023. 3

[85] 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. 1, 5

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

[87] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *ICCV*, 2019. 3

[88] Zhengyuan Yang, Tushar Kumar, Tianlang Chen, Jingsong Su, and Jiebo Luo. Grounding-tracking-integration. *IEEE TCSVT*, 2020. 3

[89] 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, 2022. 4, 5, 6

[90] Yuechen Yu, Yilei Xiong, Weilin Huang, and Matthew R. Scott. Deformable siamese attention networks for visual object tracking. In *CVPR*, 2020. 4

[91] Hui Zhang, Lei Zhang, Li Zhuo, and Jing Zhang. Object tracking in RGB-T videos using modal-aware attention network and competitive learning. *Sensors*, page 393, 2020. 5

[92] 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 *ICCVW*, pages 0–0, 2019. 5

[93] 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 TIP*, pages 3335–3347, 2021. 5

[94] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale

benchmark and new baseline. In *CVPR*, pages 8886–8895, 2022. 5

[95] Zhipeng Zhang, Houwen Peng, Jianlong Fu, Bing Li, and Weiming Hu. Ocean: Object-aware anchor-free tracking. In *ECCV*, pages 771–787, 2020. 4

[96] Zhipeng Zhang, Yihao Liu, Xiao Wang, Bing Li, and Weiming Hu. Learn to match: Automatic matching network design for visual tracking. In *ICCV*, 2021. 4

[97] Haojie Zhao, Xiao Wang, Dong Wang, Huchuan Lu, and Xiang Ruan. Transformer vision-language tracking via proxy token guided cross-modal fusion. *PRL*, pages 10–16, 2023. 3

[98] Yaozong Zheng, Bineng Zhong, Qihua Liang, Guorong Li, Rongrong Ji, and Xianxian Li. Toward unified token learning for vision-language tracking. *IEEE TCSVT*, 34(4): 2125–2135, 2023. 3

[99] Yaozong Zheng, Bineng Zhong, Qihua Liang, Zhiyi Mo, Shengping Zhang, and Xianxian Li. Odtrack: Online dense temporal token learning for visual tracking. In *AAAI*, pages 7588–7596, 2024. 3, 4, 6

[100] Li Zhou, Zikun Zhou, Kaige Mao, and Zhenyu He. Joint visual grounding and tracking with natural language specification. In *CVPR*, pages 23151–23160, 2023. 3

[101] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *CVPR*, pages 9516–9526, 2023. 5

[102] 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 *AAAI*, pages 3870–3878, 2023. 5

[103] Yabin Zhu, Chenglong Li, Bin Luo, Jin Tang, and Xiao Wang. Dense feature aggregation and pruning for RGBT tracking. In *ACMMM*, pages 465–472, 2019. 5

[104] Yabin Zhu, Chenglong Li, Jin Tang, and Bin Luo. Quality-aware feature aggregation network for robust RGBT tracking. *IEEE TIV*, pages 121–130, 2020. 5