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

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

1. Method

1.1. Base Model

In this paper, We choose the classic one-stream RGB-based model, *e.g*., OSTrack [16], as the pre-trained model. It is composed of a ViT [4] backbone and a prediction head. The ViT backbone includes a Patch Embed Layer and multiple ViT blocks, which have been explained in detail in the preceding body sections and will not be reiterated here.

Patch Embed Layer. The input of our proposed SDSTrack consists of a pair of template frames and a pair of search frames, *i.e*., one RGB template frame $\mathbf{z}_{\text{image}}^{\text{rgb}} \in \mathbb{R}^{H_z \times W_z \times 3}$, one RGB search frame $\mathbf{x}_{\text{image}}^{\text{rgb}} \in \mathbb{R}^{H_x \times W_x \times 3}$, one X-modal template frame $\mathbf{z}_{\text{image}}^{\text{X}} \in \mathbb{R}^{H_z \times W_z \times 3}$, and one X-modal search frame $\mathbf{x}_{\text{image}}^{\text{X}} \in \mathbb{R}^{H_x \times W_x \times 3}$. They are first split and flattened into sequences of patches $\mathbf{z}_{\text{rgb}}$, $\mathbf{z}_{\text{X}} \in \mathbb{R}^{N_z \times (3P^2)}$ and $\mathbf{x}_{\text{rgb}}$, $\mathbf{x}_{\text{X}} \in \mathbb{R}^{N_x \times (3P^2)}$, where $P \times P$ is the resolution of each patch, and $N_z = \frac{H_z W_z}{P^2}$, $N_x = \frac{H_x W_x}{P^2}$ are the number of patches of template and search region respectively. Then, two trainable linear projection layers with parameter $\mathbf{E}_{\text{rgb}} \in \mathbb{R}^{(3P^2) \times D}$ and $\mathbf{E}_{\text{X}} \in \mathbb{R}^{(3P^2) \times D}$ are used to project $\mathbf{z}_{\text{rgb}}, \mathbf{x}_{\text{rgb}}$ and $\mathbf{z}_{\text{X}}, \mathbf{x}_{\text{X}}$ into D dimension latent space and the output of this projection is commonly called patch embeddings. After that, learnable $1D$ position embeddings $\mathbf{P}_{\text{z}} \in \mathbb{R}^{N_z \times D}$ and $\mathbf{P}_{\text{x}} \in \mathbb{R}^{N_x \times D}$ are added to the template patch embeddings $\hat{\mathbf{z}}_{\text{rgb}}, \hat{\mathbf{z}}_{\text{X}} \in \mathbb{R}^{N_z \times D}$ and search patch embeddings $\hat{\mathbf{x}}_{\text{rgb}}$, $\hat{\mathbf{x}}_{\text{X}} \in \mathbb{R}^{N_x \times D}$ seperately. The above processing can be represented as follows:

$$\hat{\mathbf{z}}_{\text{rgb}} = \left[\mathbf{z}_{\text{rgb}}^{1}\mathbf{E}_{\text{rgb}}; \mathbf{z}_{\text{rgb}}^{2}\mathbf{E}_{\text{rgb}}; \ldots; \mathbf{z}_{\text{rgb}}^{N_z}\mathbf{E}_{\text{rgb}}\right] + \mathbf{P}_z \quad (1)$$

$$\hat{\mathbf{z}}_{\text{X}} = \left[\mathbf{z}_{\text{X}}^{1}\mathbf{E}_{\text{X}}; \mathbf{z}_{\text{X}}^{2}\mathbf{E}_{\text{X}}; \ldots; \mathbf{z}_{\text{X}}^{N_z}\mathbf{E}_{\text{X}}\right] + \mathbf{P}_z \quad (2)$$

$$\hat{\mathbf{x}}_{\text{rgb}} = \left[\mathbf{x}_{\text{rgb}}^{1}\mathbf{E}_{\text{rgb}}; \mathbf{x}_{\text{rgb}}^{2}\mathbf{E}_{\text{rgb}}; \ldots; \mathbf{x}_{\text{rgb}}^{N_x}\mathbf{E}_{\text{rgb}}\right] + \mathbf{P}_x \quad (3)$$

$$\hat{\mathbf{x}}_{\text{X}} = \left[\mathbf{x}_{\text{X}}^{1}\mathbf{E}_{\text{X}}; \mathbf{x}_{\text{X}}^{2}\mathbf{E}_{\text{X}}; \ldots; \mathbf{x}_{\text{X}}^{N_x}\mathbf{E}_{\text{X}}\right] + \mathbf{P}_x \quad (4)$$

1.2. Complementary Masked Patch Distillation

1.2.1 Random Complementary Patch Mask (RCPM)

During training, we apply a random complementary patch masking strategy to the patch embeddings to obtain masked patch embeddings. Specifically, after obtaining the patch embeddings in the Patch Embed Layers as described in Sec. 1.1, we randomly occlude 30% of the RGB patch embeddings $\hat{\mathbf{z}}_{\text{rgb}}, \hat{\mathbf{x}}_{\text{rgb}}$. Similarly, we randomly occlude 30% of the X-modal embeddings $\hat{\mathbf{z}}_{\text{X}}, \hat{\mathbf{x}}_{\text{X}}$, but if both RGB and X modalities occlude the same positions, we remove the X-modal occlusion in the positions to ensure that at least one modality is available. Hence, we obtain masked patch embeddings $\hat{\tilde{\mathbf{z}}}_{\text{rgb}}, \hat{\tilde{\mathbf{z}}}_{\text{X}} \in \mathbb{R}^{N_z \times D}$, and $\hat{\tilde{\mathbf{x}}}_{\text{rgb}}$, $\hat{\tilde{\mathbf{x}}}_{\text{X}} \in \mathbb{R}^{N_x \times D}$. Then, we also utilize the position embeddings $\mathbf{P}_{\text{z}}$ and $\mathbf{P}_{\text{x}}$ to incorporate position information into the masked patch embeddings. During training, the process of the masked data is the same as that of the clean data in the model.

2. Experiments

2.1. Robustness performance

To comprehensively analyze the robustness of our SDSTrack, we compared its performance with previous methods on various challenging attributes on the LasHeR [7] and VisEvent [11] test sets.

LasHeR. The attribute-based performance results of our method on the LasHeR [7] test set are presented in Tab. 1. Our SDSTrack achieves state-of-the-art performance in the majority of attributes. Specifically, in sequences involving

[*]Corresponding authors.

	SGT++ [5]	CAT [6]	FANet [18]	APFNet [12]	STARKS50 [13]	TransT [1]	OSTrack [16]	ProTrack [15]	ViPT [17]	**SDSTrack** **(ours)**
NO	0.530/0.325	0.654/0.430	0.597/0.405	0.667/0.467	0.610/0.485	0.700/0.522	0.728/0.562	0.754/0.580	0.876/0.684	**0.876/0.708**
PO	0.341/0.240	0.418/0.295	0.415/0.292	0.473/0.345	0.426/0.344	0.494/0.373	0.487/0.393	0.505/0.396	0.624/0.503	**0.634/0.506**
TO	0.307/0.211	0.361/0.260	0.341/0.250	0.417/0.314	0.386/0.308	0.433/0.327	0.403/0.329	0.439/0.342	0.576/0.461	**0.604/0.478**
HO	0.147/0.167	0.226/0.234	0.167/0.227	0.271/0.277	0.330/0.340	0.346/0.338	0.299/0.318	0.402/0.386	0.473/0.438	**0.589/0.527**
MB	0.324/0.207	0.398/0.266	0.400/0.260	0.459/0.328	0.427/0.337	0.477/0.350	0.465/0.370	0.524/0.395	0.573/0.459	**0.599/0.475**
LI	0.296/0.205	0.315/0.226	0.330/0.235	0.418/0.308	0.296/0.252	0.338/0.266	0.331/0.282	0.424/0.334	0.498/0.412	**0.541/0.438**
HI	0.422/0.241	0.525/0.357	0.527/0.355	0.604/0.412	0.443/0.342	0.496/0.344	0.527/0.407	0.595/0.444	0.679/0.542	**0.694/0.551**
AIV	0.179/0.143	0.226/0.190	0.188/0.184	0.321/0.262	0.203/0.208	0.224/0.230	0.234/0.222	0.304/0.267	0.375/0.350	**0.543/0.482**
LR	0.373/0.216	0.424/0.252	0.432/0.260	0.461/0.294	0.376/0.267	0.450/0.297	0.435/0.312	0.462/0.321	0.564/0.416	**0.575/0.425**
DEF	0.356/0.274	0.383/0.306	0.331/0.282	0.458/0.368	0.432/0.363	0.542/0.433	0.479/0.406	0.519/0.428	0.674/0.557	**0.695/0.563**
BC	0.318/0.237	0.398/0.298	0.402/0.295	0.449/0.337	0.427/0.343	0.471/0.361	0.480/0.387	0.498/0.388	**0.649/0.518**	0.644/0.513
SA	0.346/0.246	0.374/0.265	0.391/0.282	0.428/0.317	0.399/0.329	0.438/0.346	0.445/0.370	0.451/0.363	0.574/**0.465**	**0.574**/0.463
CM	0.364/0.238	0.419/0.294	0.420/0.293	0.477/0.351	0.439/0.349	0.479/0.357	0.486/0.389	0.541/0.416	0.621/0.500	**0.636/0.507**
TC	0.327/0.224	0.370/0.262	0.374/0.264	0.431/0.316	0.393/0.315	0.457/0.341	0.439/0.352	0.458/0.358	0.573/0.460	**0.577/0.462**
FL	0.325/0.189	0.387/0.226	0.353/0.207	0.376/0.279	0.399/0.322	0.453/0.335	0.438/0.353	0.520/0.386	0.591/0.465	**0.598/0.465**
OV	0.217/0.245	0.260/0.230	0.247/0.236	0.364/0.342	0.528/0.464	0.623/0.514	0.747/0.639	0.548/0.458	**0.762/0.650**	0.700/0.606
FM	0.330/0.237	0.399/0.291	0.389/0.285	0.451/0.339	0.433/0.357	0.501/0.386	0.491/0.403	0.520/0.414	0.631/0.514	**0.656/0.528**
SV	0.364/0.250	0.444/0.307	0.441/0.307	0.498/0.360	0.452/0.364	0.521/0.393	0.521/0.418	0.545/0.425	0.650/0.525	**0.664/0.530**
ARC	0.281/0.216	0.325/0.244	0.317/0.239	0.405/0.310	0.406/0.343	0.490/0.382	0.463/0.387	0.475/0.391	0.593/0.495	**0.611/0.501**
ALL	0.365/0.251	0.450/0.314	0.441/0.309	0.500/0.362	0.449/0.361	0.524/0.394	0.515/0.412	0.538/0.420	0.651/0.525	**0.665/0.531**

Table 1. **Attribute performance** on the LasHeR [7] test set.

	ATOM(EF) [3]	MDNet(MF) [8]	VITAL(MF) [10]	LTMU(EF) [2]	TransT [1]	STARKS50 [13]	OSTrack [16]	ProTrack [15]	ViPT [17]	**SDSTrack** **(ours)**
Camera Motion	0.557/0.385	0.589/0.406	0.602/0.415	0.640/0.452	0.644/0.474	0.649/0.479	0.658/0.511	0.665/0.505	0.711/0.559	**0.737/0.576**
Rotation	0.452/0.351	0.421/0.330	0.434/0.334	0.574/0.442	0.525/0.428	0.546/0.453	0.595/0.487	0.574/0.476	**0.678/0.551**	0.648/0.527
Deformation	0.298/0.222	0.156/0.127	0.155/0.129	0.351/0.253	0.325/0.224	0.346/0.264	0.362/0.264	0.378/0.269	**0.468/0.359**	0.462/0.354
Full Occlusion	0.401/0.258	0.478/0.277	0.482/0.279	0.486/0.320	0.529/0.376	0.480/0.331	0.543/0.399	0.547/0.380	0.566/0.427	**0.594/0.443**
Low Illumination	0.518/0.358	0.601/0.389	0.581/0.375	0.599/0.421	0.618/0.450	0.587/0.431	0.623/0.478	0.627/0.472	0.721/0.565	**0.743/0.571**
Out-of-View	0.423/0.282	0.406/0.260	0.427/0.281	0.525/0.360	0.501/0.381	0.511/0.383	0.521/0.407	0.568/0.425	0.543/0.425	**0.566/0.433**
Partial Occlusion	0.438/0.285	0.548/0.335	0.536/0.324	0.530/0.362	0.543/0.386	0.498/0.352	0.589/0.439	0.532/0.380	**0.672/0.512**	0.664/0.500
Viewpoint Change	0.556/0.386	0.583/0.385	0.557/0.357	0.634/0.457	0.611/0.462	0.636/0.483	0.654/0.534	0.676/0.508	0.758/0.605	**0.784/0.627**
Scale Variation	0.559/0.369	0.596/0.345	0.581/0.327	0.611/0.429	0.597/0.437	0.526/0.385	0.611/0.507	0.570/0.424	0.729/0.572	**0.735/0.574**
Background Clutter	0.551/0.369	0.642/0.407	0.626/0.397	0.599/0.417	0.626/0.455	0.585/0.421	0.673/0.572	0.596/0.440	0.733/0.573	0.**740/0.573**
Motion Blur	0.504/0.360	0.456/0.321	0.462/0.332	0.558/0.410	0.554/0.424	0.560/0.428	0.592/0.466	0.491/0.376	0.637/0.504	**0.638/0.510**
Aspect Ration Change	0.502/0.346	0.517/0.321	0.512/0.320	0.581/0.416	0.592/0.434	0.547/0.399	0.640/0.495	0.575/0.430	0.692/0.548	0.**725/0.568**
Fast Motion	0.569/0.398	0.570/0.368	0.544/0.357	0.607/0.439	0.577/0.423	0.566/0.424	0.656/0.506	0.567/0.432	0.749/0.585	**0.753/0.589**
No Motion	0.589/0.429	0.598/0.426	0.635/0.427	0.707/0.506	0.625/0.490	0.618/0.466	0.688/0.556	0.671/0.507	0.707/0.583	**0.725/0.589**
Illumination Variation	0.580/0.404	0.672/0.453	0.651/0.432	0.638/0.453	0.607/0.454	0.592/0.448	0.642/0.500	0.611/0.465	0.744/0.588	**0.767/0.595**
Over Exposure	0.600/0.408	0.688/0.445	0.661/0.434	0.619/0.424	0.548/0.392	0.562/0.410	0.604/0.450	0.593/0.433	0.734/0.546	**0.752/0.560**
Background Object Motion	0.550/0.367	0.634/0.401	0.618/0.388	0.611/0.423	0.612/0.439	0.574/0.412	0.659/0.497	0.588/0.433	0.727/0.561	**0.729/0.561**
ALL	0.608/0.412	0.661/0.426	0.649/415	0.655/0.459	0.650/0.474	0.612/0.446	0.695/0.534	0.632/0.471	0.758/0.592	**0.767/0.597**

Table 2. **Attribute performance** on the VisEvent [11] test set.

occlusion, such as Partial Occlusion (PO), Total Occlusion (TO), and Hyaline Occlusion (HO), our method achieves the best results, indicating its effectiveness in accurately tracking targets even when they are partially or entirely occluded. Notably, it shows a precision improvement of 11.6% and a success improvement of 8.9% in Hyaline Occlusion (HO). Regarding sequences related to illumination, such as Low Illumination (LI), High Illumination (HI), and Abrupt Illumination Variation (AIV), our method demonstrates the best performance. Particularly, it achieves a precision improvement of 16.8% and a success improvement of 13.2% in Abrupt Illumination Variation (AIV), showing a strong ability to adapt to variations in external lighting conditions. In sequences involving similarity interference, such as Similar Appearance (SA) and Thermal Crossover (TC), our method achieves superior results, indicating its effectiveness in distinguishing similar objects. Furthermore, our method excels in handling motion interference, including Motion Blur (MB), Camera Moving (CM), and Fast Motion (FM), outperforming other methods. For example, it obtains a precision improvement of 2.6% and a success improvement of 1.6% in Motion Blur (MB), suggesting its capability to cope with camera or target movement interference. Additionally, our SDSTrack exhibits superior performance in attributes such as Low Resolution (LR), Deformation (DEF), Frame Lost (FL), Scale Variation (SV), *etc.*, demonstrating its robustness.

VisEvent. We also evaluate the attribute-based performance of our SDSTrack on the VisEvent [11] test set. The results are shown in Tab. 2. Our method achieves state-of-the-art performance in the majority of attributes. Notably, our method outperforms other methods in sequences involving motion interference, such as Camera Motion, Motion Blur, Fast Motion, and Background Object Motion. Specifically, it shows a precision rate of 73.7% and a success rate of 57.6% in Camera Motion, a precision rate of 75.3%, and a success rate of 58.9% in Fast Motion, indicating the effective utilization of multimodal information, thereby enhancing tracking robustness. Regarding sequences related to illumination, such as Low Illumination, Illumination Variation, and Over Exposure, our method achieves the best results. For example, our SDSTrack obtains a precision rate of 74.3% and a success rate of 57.1% in Low Illumination, and a precision rate of 76.7% and a success rate of 59.5% in Illumination Variation. Furthermore, our SDSTrack exhibits superior performance in Full Occlusion, Out-of-View, Viewpoint Change, Scale Variation, Background Clutter, Aspect Ratio Change, and other attributes, demonstrating improved robustness.

Overall, the results obtained on both the LasHeR [7] and VisEvent [11] test sets indicate the strong performance and robustness of our SDSTrack.

2.2. Supplementary Ablation Studies

Effect of different mask strategies. We conduct a series of exploration experiments to investigate the strategies of applying masking to multimodal embeddings in the Random Complementary Patch Mask (RCPM) approach. The results of these experiments are presented in Tab. 3. Take the strategy "10%, 10%" as an example, we randomly occlude 10% of the RGB patch embeddings $(\hat{\mathbf{z}}_{\text{rgb}}, \hat{\mathbf{x}}_{\text{rgb}})$ and 10% of the X-modal embeddings $(\hat{\mathbf{z}}_{\text{X}}, \hat{\mathbf{x}}_{\text{X}})$. However, if both the RGB and X modalities occlude many positions, we remove the X-modal occlusion in the positions to ensure that at least one modality is valid. On the contrary, the strategy "30%, 30%†" (or "50%, 50%†") means that if both RGB and X occlude many positions, we still preserve the X-modal occlusion. The experimental results suggest that lower occlusion rates, such as 10% and 20%, can only marginally improve robustness, particularly in scenarios involving RGB dropping, resulting in poor overall performance. On the other hand, higher occlusion rates, such as 40% and 50%, effectively enhance the model's robustness, especially when RGB data is missing. These strategies reduce dependence on the RGB modality but obtain moderate overall performance due to a reduction in the amount of effective semantic information. Comparing the strategies "30%, 30%" and "30%, 30%†" as well as "50%, 50%" and "50%, 50%†", where † denotes that we will not remove the X-modal occlusion if RGB and X occlude the same positions, we find that it is necessary to retain the effectiveness of at least one modality. Therefore, we choose an occlusion rate of 30% and ensure information availability from at least one modality. This strategy not only yields the best performance in various challenging scenarios, such as RGB dropping, complete occlusion of the target, motion interference, and abrupt changes in illumination, but also demonstrates the best overall performance.

strategies	w/o RGB		Total Occlusion		Motion Blur		Abrupt IV		overall	
	Pre	Suc	Pre	Suc	Pre	Suc	Pre	Suc	Pre	Suc
10%, 10%	0.502	0.406	0.459	0.586	0.456	0.580	0.436	0.488	0.649	0.516
20%, 20%	0.481	0.394	0.466	0.593	0.469	0.596	0.428	0.486	0.658	0.525
30%, 30%†	0.500	0.405	0.460	0.584	0.466	0.589	0.443	0.492	0.655	0.523
40%, 40%	0.533	0.428	0.467	0.591	0.467	0.593	0.470	0.525	0.654	0.521
50%, 50%	0.534	0.432	0.463	0.586	0.464	0.585	0.456	0.499	0.655	0.523
50%, 50%†	0.534	0.428	0.454	0.578	0.449	0.571	0.464	0.519	0.646	0.514
30%, 30%	**0.538**	**0.432**	**0.478**	**0.604**	**0.475**	**0.599**	**0.482**	**0.543**	**0.665**	**0.531**

Table 3. **Ablation studies on the different random complementary patch mask strategies** on the LasHeR [7] test set. † denotes that we will not remove the X-modal occlusion if RGB and X occlude the same positions.

fusion strategies	Pr	Re	F-score
1 gap	0.575	0.571	0.573
2 even gaps	0.605	0.590	0.598
3 even gaps	0.610	0.596	0.603
4 even gaps	0.616	0.597	0.606
6 even gaps	0.600	0.598	0.599
12 even gaps	0.573	0.575	0.574
4 uneven gaps	**0.619**	**0.609**	**0.614**

Table 4. **Ablation studies on the effect of strategies for reusing ViT blocks as fusion stages** on the DepthTrack [14] test set.

Effect of different fusion strategies. To investigate the effect of different strategies for reusing ViT blocks as multimodal fusion stages, we designed several comparative strategies. The ViT structure is not explicitly staged, so we try to perform multimodal fusion by reusing ViT blocks in different gaps. For example, we reuse only the 11th ViT block as the fusion stage (1 gap) or reuse the 5th and 11th blocks as multiple fusion stages (2 even gaps). We also explore other strategies following a similar pattern. As shown in Tab. 4, the results indicate that solely reusing the 11th block as the fusion stage (1 gap) fails to achieve satisfactory fusion, resulting in an F-score of only 57.3%. By reusing the 5th and 11th blocks (2 even gaps), we enable multi-

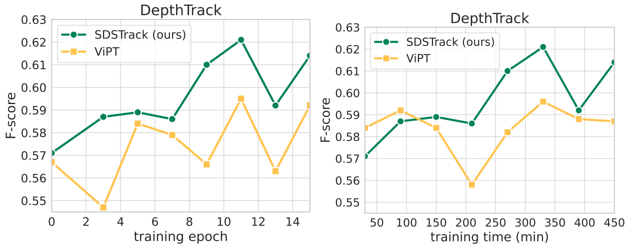

Figure 1. **Comparison of training speed and overall model performance among various trackers** on the DepthTrack [14] test set. The left figure illustrates the F-scores of trackers at different epochs, while the right figure demonstrates the F-scores of trackers at different training durations.

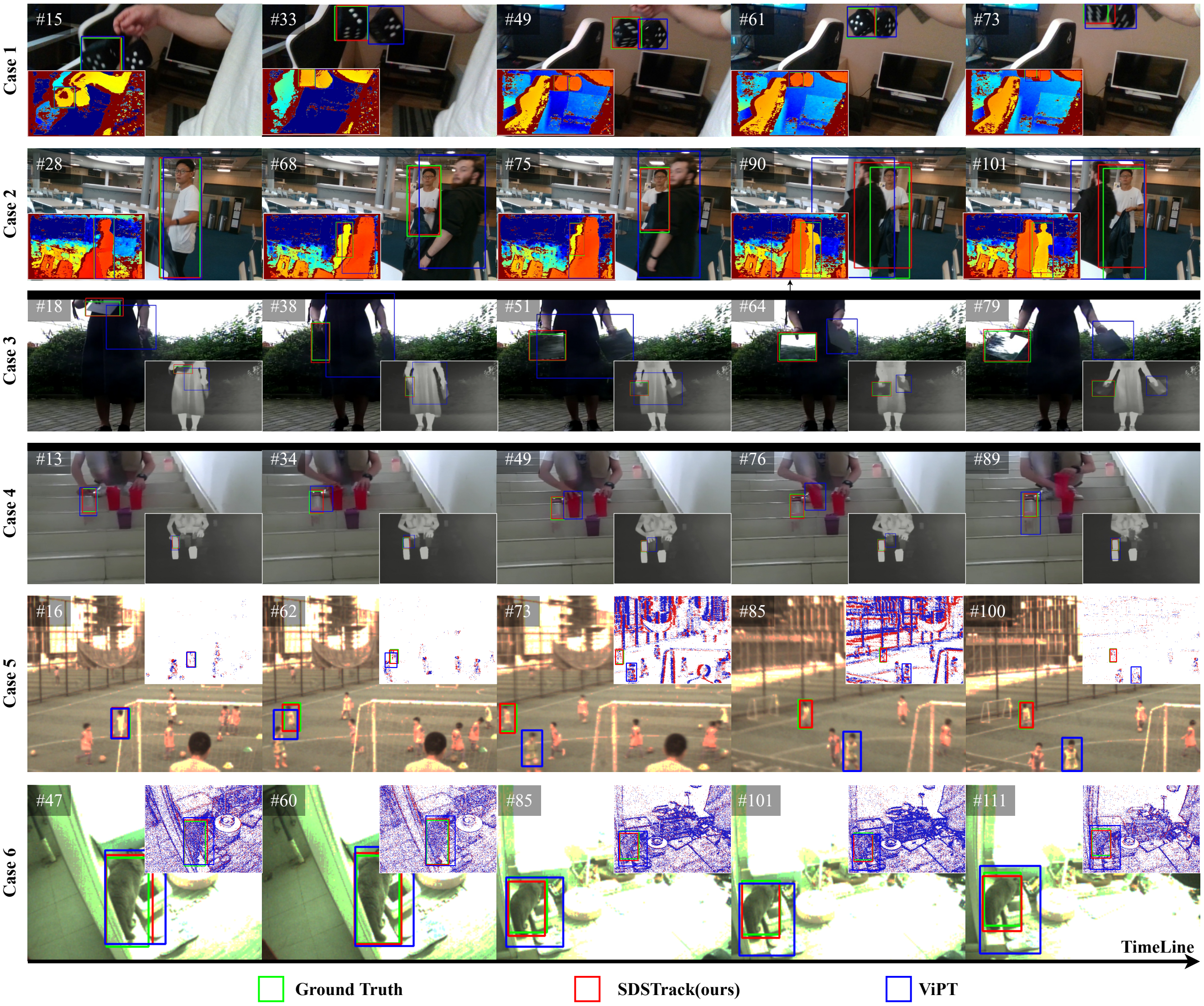

Figure 2. **Visualization of tracking results.** Case 1 and Case 2 represent RGB-D tracking, Case 3 and Case 4 denote RGB-T tracking, and Case 5 and Case 6 signify RGB-E tracking. The small images in the corners of each figure show the X-modal images.

modal fusion across multiple semantic levels, leading to improved fusion performance. Refining the gaps by reusing the 3rd, 7th, and 11th blocks (3 even gaps) or the 2nd, 6th, 9th, and 11th blocks (4 even gaps), we achieve even better outcomes with F-scores exceeding 60%. However, continuously refining gaps, like "6 even gaps" and "12 even gaps", increases the computational burden without significant performance improvement. In this paper, inspired by [9], we choose the strategy "4 uneven gaps", *i.e.*, reusing the 3rd, 6th, 9th, and 11th blocks as fusion stages, effectively utilizing mid-level features compared to the strategy "4 even gaps" and obtaining the best performance.

Training speed and tracking performance. In order to evaluate the training speed and performance of our method compared to existing methods, we conduct a comparative analysis. We ensure that the sampling number per epoch and data preprocessing is consistent across all methods, thereby attributing any differences in training speed and model performance solely to variations in model design. As shown in Fig. 1, the left figure demonstrates the performance of different methods in terms of F-score at each training epoch, and the right figure compares the methods under the same training time. The left figure demonstrates that our method, SDSTrack, outperforms the previous state-of-the-art (SOTA) method, ViPT [17], in terms of F-score in each training epoch, consistently reaching the new SOTA level. Furthermore, the right figure shows that despite the potentially slower training speed caused by our self-distillation strategy (as mentioned in the limitations section), our SDSTrack requires only minimal training time to surpass the performance of previous SOTA methods. This advantage persists throughout the subsequent training process. In summary, our method efficiently fine-tunes pre-trained models on small-scale multimodal datasets with fewer epochs and less training time.

2.3. Visualization

We present the tracking results of different modalities in Fig. 2. In scenes where target occlusion occurs, as observed in Case 2, the previous SOTA method, ViPT [17], is susceptible to interference from objects in front of the target, leading to tracking failures. In contrast, our SDSTrack effectively mitigates the impact of target occlusion, resulting in more robust tracking. In scenes with a high presence of similar objects, as shown in Case 1, Case 4, and Case 5, our method demonstrates its ability to resist the influence of these objects, ensuring accurate tracking. Furthermore, in scenarios with poor image quality, as observed in Case 3 and Case 6, our method successfully overcomes the influence of the target itself or external light changes by fusing multimodal images. This enables us to accurately track the target even under challenging conditions. Our approach fully leverages the information provided by multimodal images, reducing reliance on a specific modality, particularly RGB images, and thereby enabling more accurate and robust tracking.

References

[1] 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. 2

[2] 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. 2

[3] 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. 2

[4] 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. 1

[5] Chenglong Li, Nan Zhao, Yijuan Lu, Chengli Zhu, and Jin Tang. Weighted sparse representation regularized graph learning for rgb-t object tracking. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1856–1864, 2017. 2

[6] Chenglong Li, Lei Liu, Andong Lu, Qing Ji, and Jin Tang. Challenge-aware rgbt tracking. In *European Conference on Computer Vision*, pages 222–237. Springer, 2020. 2

[7] 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. 1, 2, 3

[8] 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. 2

[9] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021. 4

[10] 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. 2

[11] Xiao Wang, Jianing Li, Lin Zhu, Zhipeng Zhang, Zhe Chen, Xin Li, Yaowei Wang, Yonghong Tian, and Feng Wu. Visevent: Reliable object tracking via collaboration of frame and event flows. *IEEE Transactions on Cybernetics*, 2023. 1, 2, 3

[12] 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. 2

[13] 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. 2

[14] 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

[15] 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

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

[17] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9516–9526, 2023. 2, 4, 5

[18] 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. 2