Online Dense Point Tracking with Streaming Memory

Supplementary Material

1. Implementation Details

Network Details. We adopt the classical optical flow estimation network RAFT [10] as our backbone. Besides, following the modification of DOT [5] to the architecture of RAFT, we also use a stride 1 in the first convolutional layer of the image encoder and predict the visibility mask by an additional mask decoder. The channel number of different features is set as follows: D of encoded feature from the image encoder is 256, D_k of memory key is 128, D_v of memory value is 256, and D_s of sensory memory is 128. In addition, the bilinear kernel $b(\cdot)$ used in our visibility-guided splatting has the following formulation:

$$b(\Delta) = \max(0, 1 - |\Delta_x|) \cdot \max(0, 1 - |\Delta_y|), \quad (1)$$

where $\Delta = (x_1, y_1) + \mathbf{f}^N_{1\to t}\left[(x_1, y_1)\right] - (x_t, y_t)$.

Loss Function. The loss function of SPOT consists of the l_1 loss [10] for the predicted flows and binary cross-entropy loss for the visibility prediction. Specifically, we use exponentially increasing weights for predictions from different GRU iterations. Given ground-truth optical flow $\mathbf{f}^{gt}_{1\to t}$ and visibility mask $\mathbf{v}^{gt}_{1\to t}$, our loss function is defined as:

$$\sum_{i=1}^{N} 0.8^{N-i}\left[\lambda ||\mathbf{f}^{gt}_{1\to t} - \mathbf{f}^{i}_{1\to t}||_1 + \mathrm{BCE}(\mathbf{v}^{gt}_{1\to t}, \mathbf{v}^{i}_{1\to t})\right],$$

where λ is set to 1000 empirically.

Training Details. We employ FlashAttention-2 [3] for fast attention computation within the memory reading module. During training, we use Adam optimizer with one-cycle [9] learning rate on eight NVIDIA H100 GPUs. The learning rate is 1e-4. The batch size is 24 for the first training stage (*i.e.*, 500k steps on optical flow dataset Kubric-CVO [12]) and 8 for the second one (*i.e.*, 100k steps on point tracking dataset Kubric-MOVi-F [4]).

Evaluation Details. We set the iteration number N of flow decoding to 16 by default, as 16 iterations already achieve the peak accuracy on real-world videos from DAVIS [8]. In contrast, we set it to 32 on Kinetics [2], RGB-Stacking [6], and RoboTAP [11]. Because we find that more iterations of GRU can improve the accuracy of these three datasets. These may be due to the different motion characteristics (*e.g.*, faster motion of human actions and weak texture of robotic scenes) of these three datasets.

2. More Quantitative Analysis

Online Setup under Original Window Size. We additionally give a setup without modifying the window size of existing offline models: when processing frame t, we provide the model with all prior frames. This process is repeated for each frame sequentially, ensuring predictions are based solely on past frames. This gives an upper bound of performance for reference though with extremely slow speed. We use ‡ to denote this setup. Due to the extremely slow inference speed of this setup, we only provide the result on DAVIS (First) in Tab. 1: SPOT still beats other online versions by a large margin, with a much faster inference speed. It still supports our contributions of strong performance and superior efficiency.

Table 1. Point tracking results on DAVIS (First). ‡ represents evaluated in an online fashion under the original window size with extremely slow speed.

Online	AJ ↑	$< \delta^x_{avg}$ ↑	OA ↑
TAPIR‡	56.7	70.2	85.7
CoTracker2‡	55.9	68.7	83.7
SpatialTracker‡	57.3	70.6	85.0
DOT‡	57.3	69.7	85.2
Online TAPIR	56.2	69.3	84.6
Ours	**61.5**	**75.0**	**88.9**

Further Discussion on Forward Splatting and Backward Warping. The main idea of SPOT is breaking down long-term tracking into two simpler steps, i.e., long-range propagation with flow and similarity-based short-range retrieval. SPOT instantiates propagation with splatting and retrieval with attention. We can also first use attention to algin $\mathbf{F}_t$ to memory frames, then backward warp them to frame 1, i.e., a reversed pipeline of SPOT. Though there are many-to-one mapping and empty regions in splatting, we tackle them with visibility mask and inpainting. However, there are also problems with backward warping, i.e., warping error foreground for occluded region, warping out-of-frame empty regions. These problems are caused by underlying physical motion. Therefore, there is no preference for which long-range propagation method to choose. The important point is the framework of two-step breakdown. Here, we further provide an additional ablation experiment: we warp the features of memory frames to the first frame directly, without the stage of attention for short-range retrieval. As shown in Tab. 2, though the ablated model achieve similar results as our SPOT on short videos, it fails to generalize to longer videos (CVO Extended) due to error accumulation.

Computational Complexity. SPOT only maintains 3 frames memory, the computational complexity will not increase after time step t being larger than 3. In addition, we provide curve of GPU memory and speed w.r.t video reso-

Table 2. Alation results on CVO dataset. We ablate the attention block of SPOT here.

Method	CVO (Final)	CVO (Extended)
	EPE ↓ (all / vis / occ)	EPE ↓ (all / vis / occ)
SPOT	**1.17 / 0.67 / 3.49**	**6.42 / 3.86 / 9.98**
Attention Ablation	1.18 / 0.69 / 3.50	395.35 / 379.44 / 428.48

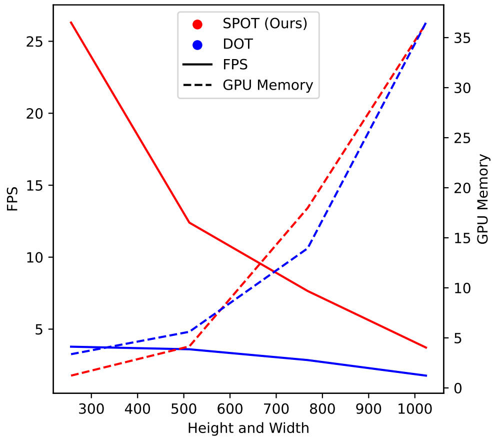


Figure 1. GPU memory and inference FPS v.s. input resolution.

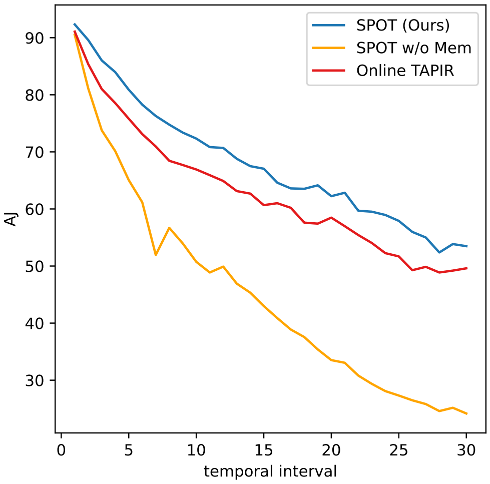


Figure 2. AJ of our SPOT, SPOT w/o Memory, and Online TAPIR across varying temporal intervals on DAVIS (First).

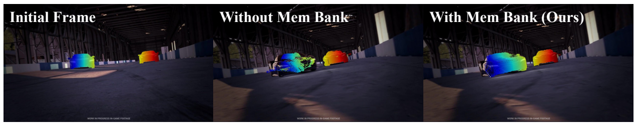


Figure 3. Qualitative ablation of memory bank on challenging real-world video.

lution in Fig. 1. SPOT runs much faster than DOT across varying resolutions up to 1024x1024, while consumes similar memory. The main computational overhead still lies in RAFT. So, we believe recent progress in efficient and high-resolution optical flow estimation can greatly and immediately benefits SPOT due to the unified architecture design.

Compared to Recent Methods. Track-On [1] and DELTA [7] are very recent works that focus on online sparse tracking and offline dense tracking, respectively. However, SPOT tackles online dense tracking. Here, we provide an efficiency comparison with them under our setting. On H100, SPOT tracks 512x512 videos (around 262k points) at **12.4** FPS with **4.15**GB GPU memory. But Track-On can only tracks up to 21k points at **1.23** FPS with **77.2**GB GPU memory; DELTA runs at **0.19** FPS with **49.58**GB GPU memory. Considering the significant latency and memory consumption, both are not suitable for online dense point tracking we consider here.

Accuracy vs. Temporal Interval. We repurpose TAP-Vid to evaluate performance across varying temporal intervals explicitly. Specifically, we employ the DAVIS (First) split here and evaluate the performance for each temporal interval (up to 30 here) without averaging the results across the temporal intervals. Fig. 2 shows that AJ will degrade gradually as the temporal interval increases, while SPOT generally outperforms Online TAPIR [11] on all temporal intervals. Besides, AJ of SPOT w/o memory degrades rapidly, illustrating the important role of memory module. Though length of our memory bank is only 3, SPOT propagates information from first frame to memory frames directly, greatly alleviating the information degradation and helping handling long videos.

3. More Qualitative Results

Qualitative Analysis of Memory Bank. We provide qualitative ablation of memory bank on video with large appearance variations in Fig. 3. As shown in Fig. 3, removing the memory bank leads to the failure of tracking due to appearance changes. Our SPOT with memory bank can successfully track the points of cars and overcome the challenge of appearance variations.

Qualitative Result on Long Occlusion. SPOT introduces visibility mask during splatting (Eq. 7 of main paper). And splatted feature of occluded region will be all zeros and no

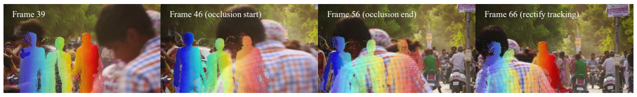


Figure 4. Qualitative result on long occlusion.

effective information can be read out through attention. So SPOT 'degrades' to pairwise method in such extreme case. We provide a such case in Fig. 4, where right woman (red color) is totally occluded by man for 10 frames (longer than 3 frames memory). Once the woman reappears, SPOT successfully locates the woman and recovers from occlusions.

Qualitative Results on CVO. We provide more qualitative comparison results with the previous state-of-the-art method for dense tracking, DOT [5], on long-range optical flow benchmark CVO in Fig. 5. The areas where our SPOT achieves substantial improvements are highlighted with bounding boxes. Please zoom in for more details. DOT typically fails to estimate the motion of small objects, occluded objects, and objects with weak textures. By contrast, our SPOT successfully estimates the motion for these hard cases.

Qualitative Results on Real-world Videos. We also provide more qualitative results with the previous state-of-the-art method for online tracking, Online TAPIR [11], on real-world videos from DAVIS in Fig. 6. Fig. 6 shows that our SPOT has superior performance on real-world videos. Please zoom in for more details.

Failure Cases. We provide some failure cases in Fig. 7. Our SPOT fails to track the bike after extreme long occlusion, i.e., more than 20 occluded frames in the first case. Besides, SPOT cannot distinguish texture-less objects properly, especially there are four similar fast-moving ducks in the second case. Finally, for fast motion shown in the third and fourth cases, SPOT may lose the track of thin object or even the whole object.

References

[1] Görkay Aydemir, Xiongyi Cai, Weidi Xie, and Fatma Güney. Track-on: Transformer-based online point tracking with memory. *arXiv preprint arXiv:2501.18487*, 2025. 2

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[3] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. 2023. 1

[4] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3749–3761, 2022. 1

[5] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. Dense optical tracking: connecting the dots. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197, 2024. 1, 3

[6] Alex X Lee, Coline Manon Devin, Yuxiang Zhou, Thomas Lampe, Konstantinos Bousmalis, Jost Tobias Springenberg, Arunkumar Byravan, Abbas Abdolmaleki, Nimrod Gileadi, David Khosid, et al. Beyond pick-and-place: Tackling robotic stacking of diverse shapes. In *5th Annual Conference on Robot Learning*, 2021. 1

[7] Tuan Duc Ngo, Peiye Zhuang, Chuang Gan, Evangelos Kalogerakis, Sergey Tulyakov, Hsin-Ying Lee, and Chaoyang Wang. Delta: Dense efficient long-range 3d tracking for any video. *arXiv preprint arXiv:2410.24211*, 2024. 2

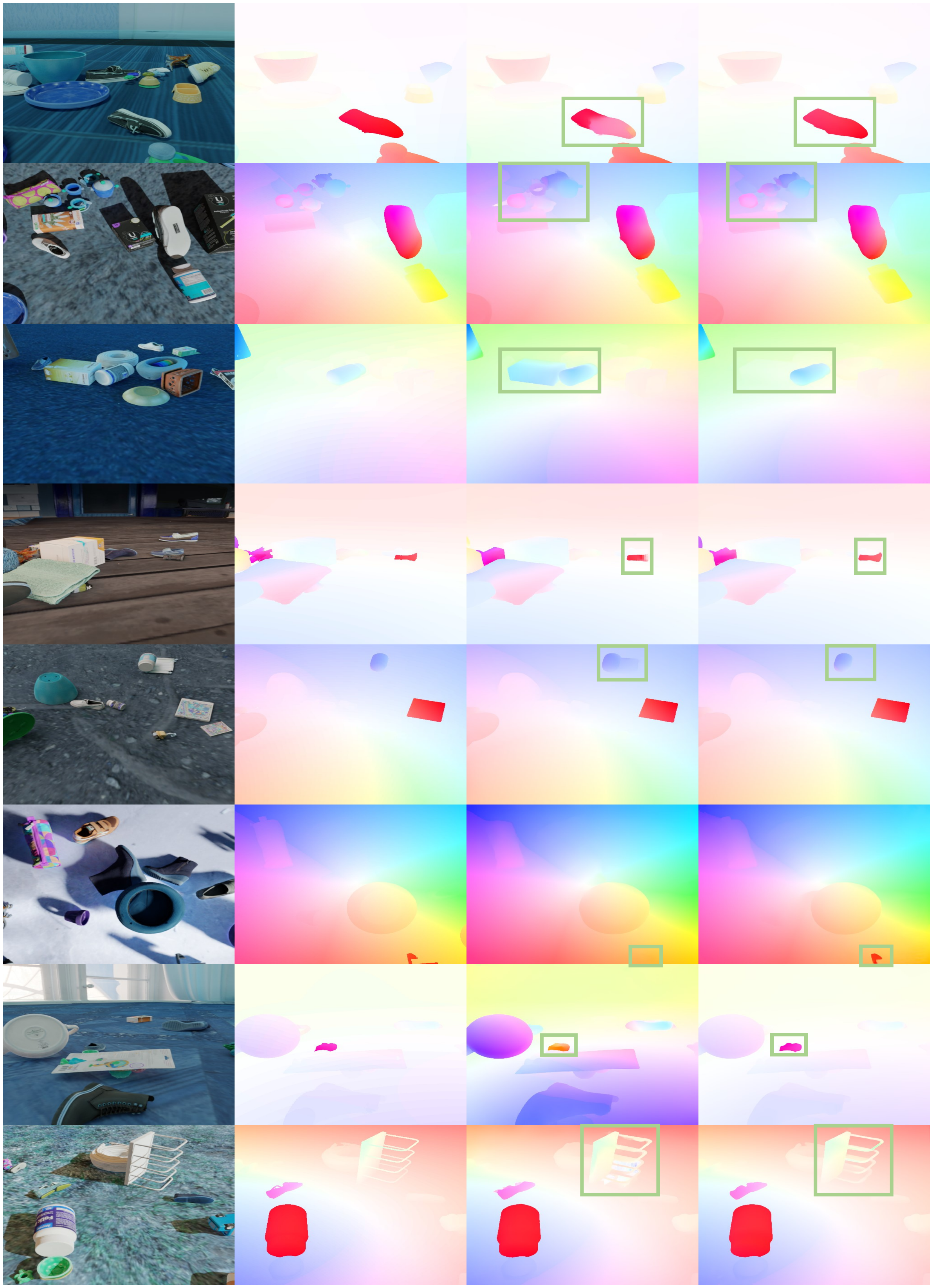
[8] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 1

[9] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, pages 369–386. SPIE, 2019. 1

[10] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 1

[11] Mel Vecerik, Carl Doersch, Yi Yang, Todor Davchev, Yusuf Aytar, Guangyao Zhou, Raia Hadsell, Lourdes Agapito, and Jon Scholz. Robotap: Tracking arbitrary points for few-shot visual imitation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5397–5403. IEEE, 2024. 1, 2, 3, 5

[12] Guangyang Wu, Xiaohong Liu, Kunming Luo, Xi Liu, Qingqing Zheng, Shuaicheng Liu, Xinyang Jiang, Guangtao Zhai, and Wenyi Wang. Accflow: Backward accumulation for long-range optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12119–12128, 2023. 1



(a) Reference Frame (b) Ground Truth (c) DOT (d) **SPOT**

Figure 5. More qualitative comparison on the CVO (Extended). Notable areas are marked by a bounding box. Please zoom in for details.

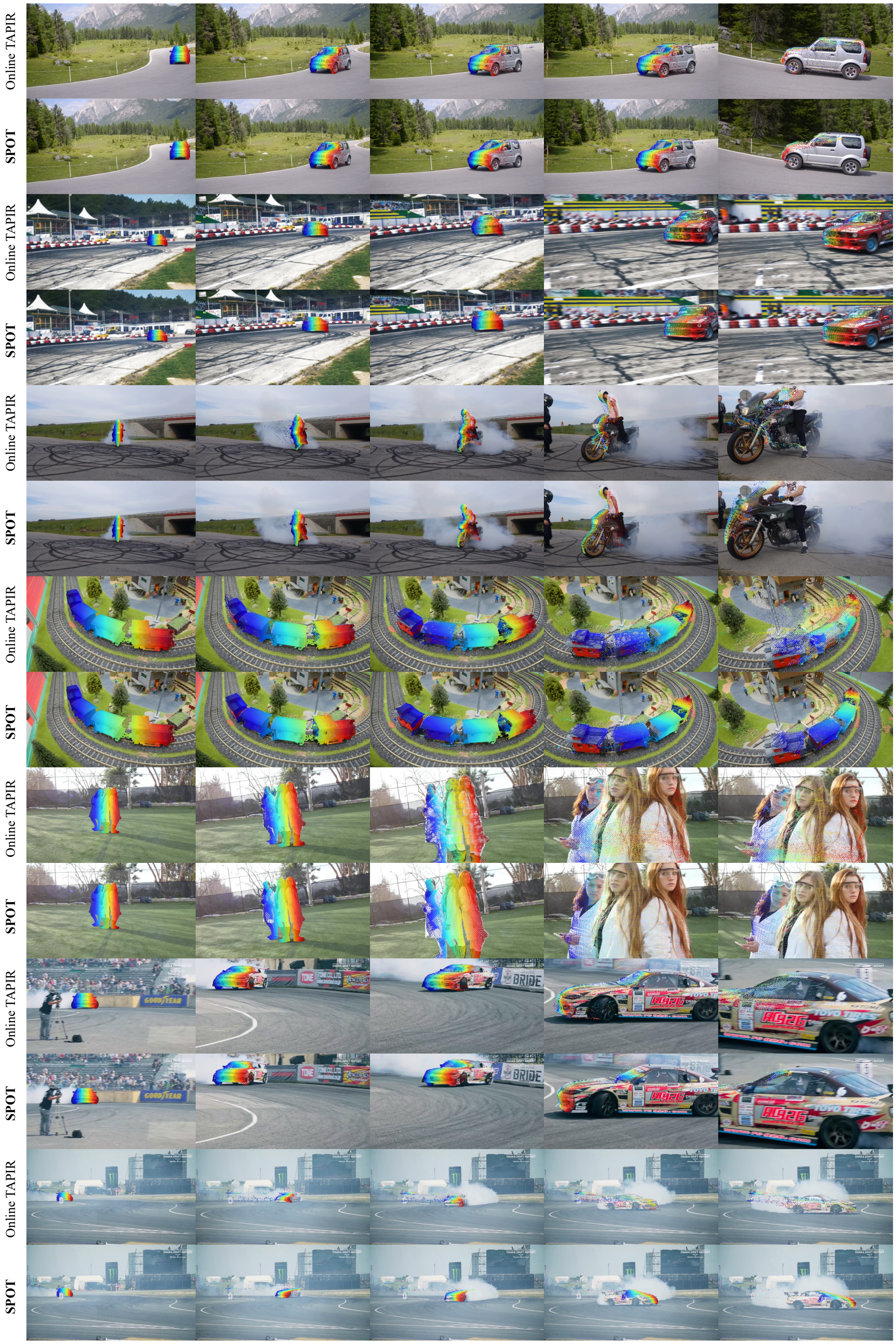


Figure 6. More qualitative comparison on DAVIS. For each sequence, we show tracking results of Online TAPIR [11] and SPOT. Only foreground points of the first frame are visualized, each point is displayed with a different color and overlayed with white stripes if occluded. Please zoom in for details.

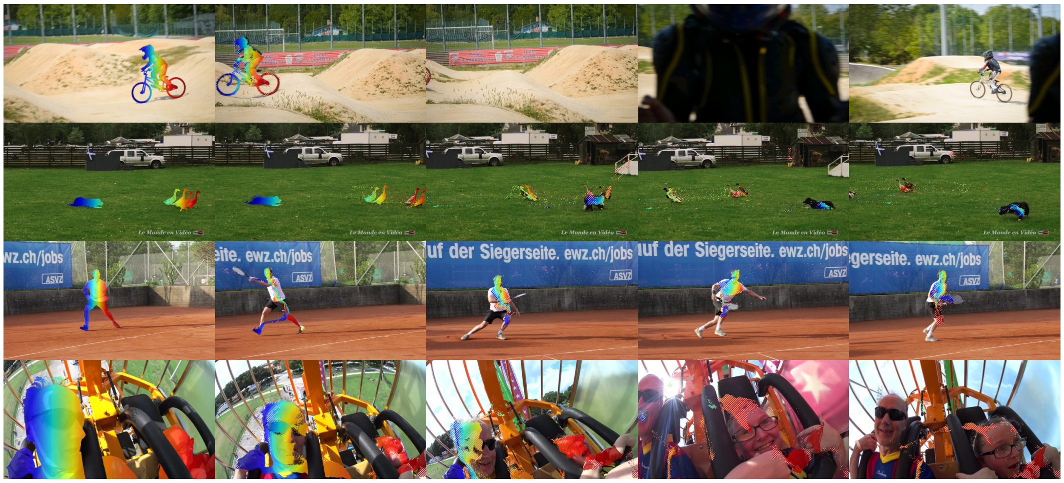


Figure 7. Failure cases on DAVIS. Please zoom in for details.