Fast Spatial Tracking with Visual Geometry Transformer Supplementary Material

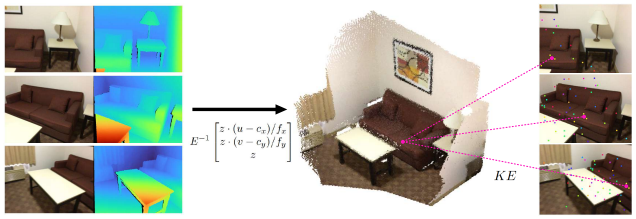
A. Training Data Generation

We train our tracker on nine datasets: Kubric, Dynamic Replica, PointOdyssey, ScanNet, ARKitscenes, nuScenes, Argoverse 2, DynPose100K++, and WildSDG. This section details the data generation pipelines for each dataset.

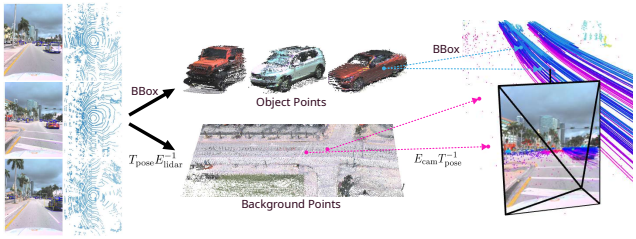
Synthetic. For *Kubric* dataset, we follow the standard data generation pipeline [5] and produce approximately 100,000 sequences, each containing 24 frames. Ground truth 2D and 3D annotations are produced using the rendered depth and object instance information provided by the simulator. Unlike CoTracker3 [7], track generation is integrated directly into our dataloader, providing higher diversity during training. *Dynamic Replica* and *PointOdyssey* both provide 3D trajectories derived from mesh vertices in a global coordinate system. For training, we transform these world tracks into each camera's coordinate system using the provided extrinsic parameters.

Indoor. *ScanNet* and *ARKitScenes* contain dense depth and camera poses but do not include tracking annotations. However, since they represent static indoor environments, 3D point tracks can be generated via reconstruction (Fig. 1a). Specifically, we first project all valid pixels into the world coordinate system using depths and camera poses. We then sample 3D points from the reconstructed point cloud and project them into each frame's coordinate system to obtain the 3D camera tracks and 2D tracks. Visibility is determined by comparing the projected depth with the depth map at the projected pixel location.

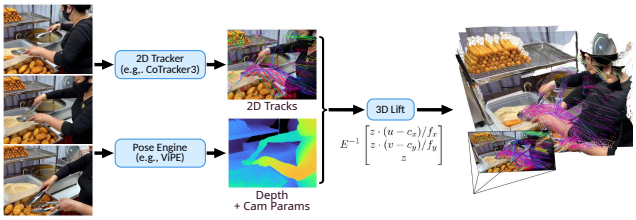
Outdoor Driving. *NuScenes* and *Argoverse 2* provide sparse LiDAR point clouds with 3D bounding box annotations. Due to the presence of dynamic objects, we create background and objec tracks separately (Fig. 1b). For background, object points are first removed using the bounding boxes, then the remaining points transformed into a global coordinate system using vehicle poses. Background tracks are then sampled and projected back to each frame. For foreground objects, the object point clouds are extracted and aligned in each object's local coordinate system. Tracks are sampled then placed into each frame using the object's bounding box. Since no dense depth map is available, accurate visibility computation is not feasible and we omit visibility supervision for these datasets. Our approach is similar to DriveTrack [1], but differs in that we generate tracks for the entire scene rather than a single object, and we do not interpolate sparse LiDAR depth for visibility estimation.



(a) Indoor datasets (ScanNet [3] and ARKitScenes [4])



(b) Outdoor driving datasets (nuScenes [2] and Argoverse 2 [9])



(c) ViPE datasets [6] (DynPose100K++ and Wild-SDG-1M)

Figure 1. Training data generation pipeline for the three types of datasets that lack tracking annotations.

ViPE. ViPE [6] is a recent reconstruction framework that jointly optimizes depth and camera parameters, and is used to generate pseudo ground truth geometry for the *DynPose100K++* (YouTube videos) and *Wild-SDG-1M* (AI-generated) datasets. Due to the lack of annotations, we follow CoTracker3 and employ an off-the-shelf 2D tracker [7] to generate pseudo ground truth 2D tracks (Fig. 1c). These 2D tracks are subsequently lifted to 3D using depths and camera parameters estimated by ViPE.

B. Qualitative Results

In figure Fig. 2, we show additional qualitative comparisons with DELTA [8], the SOTA 3D tracker that also predicts 3D tracks in a feed-forward fashion. Overall, our method can deal with complex motions and produce more consistent and smoother tracks in both indoor and outdoor scenarios.

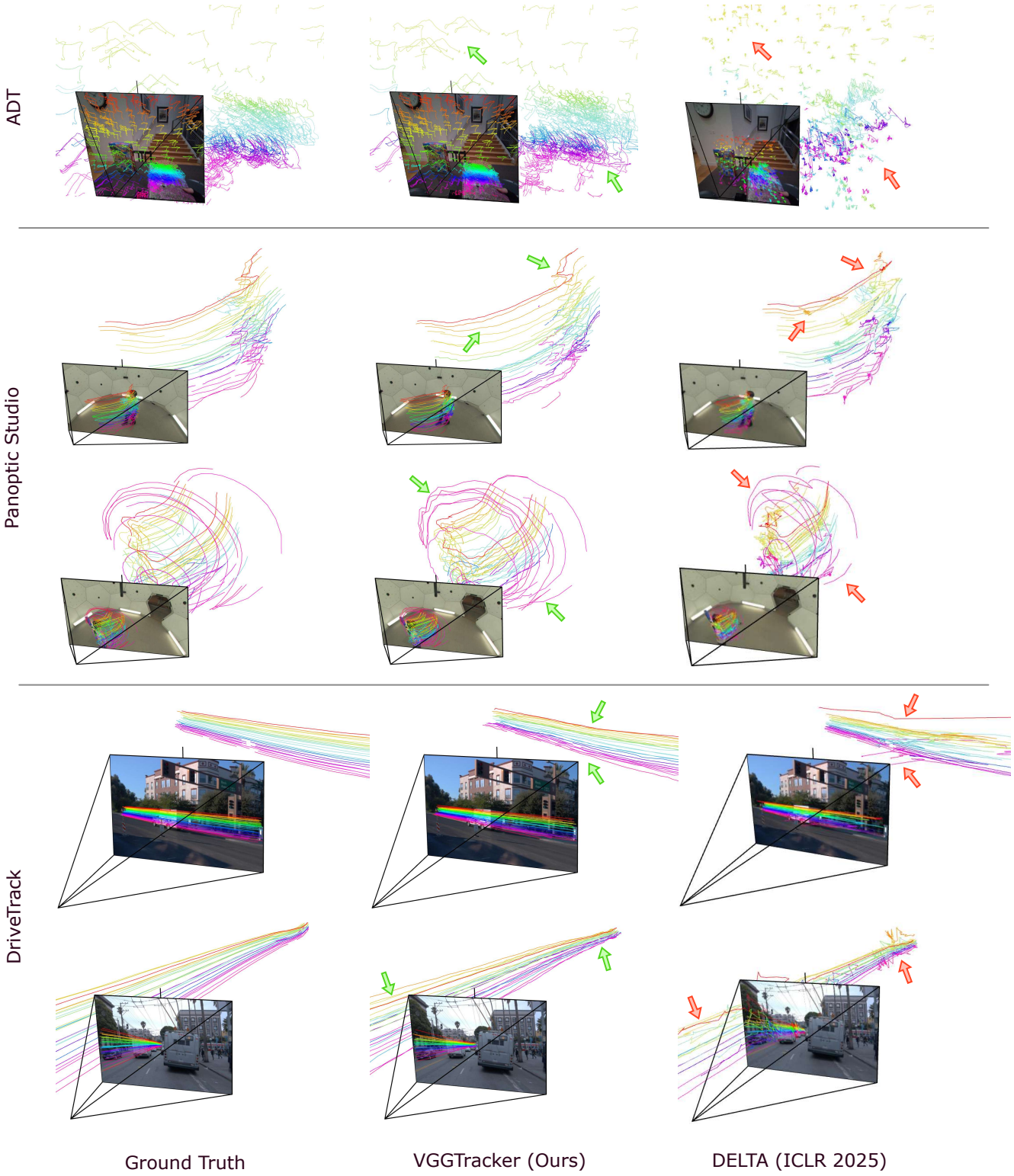


Figure 2. Qualitative comparisons with DELTA [8] on the TAPVid-3D benchmark. 3D tracks are aligned with the ground truth to have the correct scale and shift. 2D tracks are visualized by projecting the 3D camera tracks using the known intrinsics provided by the benchmark. To highlight object motion in DriveTrack, the tracks are also motion corrected using the ego vehicle pose.

References

[1] Arjun Balasingam, Joseph Chandler, Chenning Li, Zhoutong Zhang, and Hari Balakrishnan. DriveTrack: A benchmark for long-range point tracking in real-world videos. In *CVPR*, pages 22488–22497, 2024. 1

[2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 1

[3] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017. 1

[4] Afshin Dehghan, Gilad Baruch, Zhuoyuan Chen, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, and Elad Shulman. ARKitScenes: A diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. In *NeurIPS*, 2021. 1

[5] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A scalable dataset generator. In *CVPR*, pages 3749–3761, 2022. 1

[6] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025. 1

[7] Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. In *ICCV*, pages 6013–6022, 2025. 1

[8] Tuan Duc Ngo, Peiye Zhuang, Evangelos Kalogerakis, Chuang Gan, Sergey Tulyakov, Hsin-Ying Lee, and Chaoyang Wang. DELTA: Dense efficient long-range 3D tracking for any video. In *ICLR*, 2025. 1, 2

[9] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *NeurIPS*, 2021. 1