

DVGT: Driving Visual Geometry Transformer

Supplementary Material

16-Frame, 8-View



Ego View / 360° View



Figure 1. Video demonstration of our DVGT’s reconstruction of scene geometry from images on the validation set.

Appendix

A. Implementation Details

Architecture. We utilize a ViT-L model pretrained by DINOv3 [9] as the image encoder. The subsequent geometry transformer is composed of $L = 24$ attention blocks, bringing the model size to approximately 1.7 billion parameters. Each block consists of three specialized layers: an intra-view local attention layer, a cross-view spatial attention layer, and a cross-frame temporal attention layer. Following the ViT-L configuration [9], each attention layer is set to a feature dimension of 1024 with 16 heads. To enhance training stability, we incorporate QKNorm [5] and Layer-Scale [11] (initialized at 0.01) into each layer. For dense prediction, we follow the design in [15] and feed tokens from the 4th, 11th, 17th, and 23rd blocks into a DPT [8] decoder for upsampling.

Training. We train our model on a mixture of public datasets using the AdamW [7] optimizer for 160K iterations. We employ a cosine learning rate scheduler with a peak learning rate of 10^{-4} and a linear warmup of 8K iterations. To ensure training stability and efficiency, we apply gradient norm clipping with a threshold of 1.0 and leverage bfloat16 precision alongside gradient checkpointing. The

training process takes about six days on 64 H20 GPUs.

During training, we construct each batch (batch size of 1) by first sampling a dataset based on its weight, followed by a random scene. From this scene, we dynamically sample views (from 2 to the maximum available) and frames to yield a total of 48 images per batch. Input images are isotropically resized to a long-edge resolution of 518 pixels. We then apply a central crop on the shorter edge to a random size between 144 and 224 pixels (ensuring divisibility by the 16-pixel patch size), resulting in aspect ratios between 1.5 and 3.3. Following VGGT [13], we apply aggressive per-frame augmentations—including color jittering, Gaussian blur, and grayscale conversion—to improve robustness against varying lighting conditions.

B. Evaluation Metrics

We evaluate our method’s performance on two primary tasks: 3D point map reconstruction and ego-pose estimation. For 3D point map reconstruction, we adopt two sets of metrics. First, following prior works [6, 13, 14], we measure overall geometric quality using **Accuracy** and **Completeness**. Formally, let $\mathcal{P} = \{\mathbf{p}_i\}$ and $\mathcal{G} = \{\mathbf{g}_j\}$ denote the sets of valid points from the predicted point map \mathbf{P}_{pred} and the ground truth \mathbf{P}_{gt} , respectively. These metrics are

Table 1. Detailed statistics of the datasets used in our experiments. All temporal statistics are reported at a 2Hz sampling rate.

Dataset	Train Scenes	Test Scenes	Min Frames	Max Frames	Avg Frames	Avg Aspect Ratio	Num of Views
nuScenes	700	150	32	41	40	1.77	6
KITTI	138	13	2	1033	62	3.31	2
OpenScene	19376	2026	1	41	38	1.77	8
Waymo	798	202	34	40	40	1.77	5
DDAD	150	50	10	20	17	1.59	6

calculated as:

$$\begin{aligned}\text{Accuracy} &= \frac{1}{|\mathcal{P}|} \sum_{\mathbf{p} \in \mathcal{P}} \min_{\mathbf{g} \in \mathcal{G}} \|\mathbf{p} - \mathbf{g}\|_2, \\ \text{Completeness} &= \frac{1}{|\mathcal{G}|} \sum_{\mathbf{g} \in \mathcal{G}} \min_{\mathbf{p} \in \mathcal{P}} \|\mathbf{g} - \mathbf{p}\|_2,\end{aligned}\quad (1)$$

where $\|\cdot\|_2$ denotes the L_2 distance, quantifying the proximity between the prediction and the underlying scene geometry.

Second, we evaluate the ray depth accuracy using the **Absolute Relative error (Abs Rel)** and **threshold accuracy** ($\delta < 1.25$). Here, ray depth refers to the distance from a 3D point to the ego-vehicle center of its corresponding frame, which measures the local structure of the 3D point map reconstruction. Let Ω be the set of valid pixels with available ground truth, and u index a pixel within Ω . Given the predicted ray depth \mathbf{D}_{pred} and ground truth \mathbf{D}_{gt} , the metrics are defined as:

$$\begin{aligned}\text{Abs Rel} &= \frac{1}{|\Omega|} \sum_{u \in \Omega} \frac{|\mathbf{D}_{pred}(u) - \mathbf{D}_{gt}(u)|}{\mathbf{D}_{gt}(u)}, \\ \text{Acc}_\delta &= \frac{1}{|\Omega|} \sum_{u \in \Omega} \mathbb{I} \left(\max \left(\frac{\mathbf{D}_{pred}(u)}{\mathbf{D}_{gt}(u)}, \frac{\mathbf{D}_{gt}(u)}{\mathbf{D}_{pred}(u)} \right) < \delta \right),\end{aligned}\quad (2)$$

where $\mathbb{I}(\cdot)$ is the indicator function that equals 1 if the condition holds and 0 otherwise.

For ego-pose estimation, following [12], we report the **Area Under the Curve (AUC)** of the pose accuracy at a 30° threshold. Specifically, we first compute the Relative Rotation Accuracy (RRA) and Relative Translation Accuracy (RTA) for all frame pairs. RRA measures the geodesic distance between the predicted and ground truth rotation matrices, while RTA evaluates the angular deviation between the translation vectors. The accuracy at a specific threshold τ is defined as the percentage of pairs satisfying $\max(\text{RRA}, \text{RTA}) < \tau$. The final AUC@30 is obtained by integrating this accuracy over the threshold range:

$$\text{AUC@30}^\circ = \frac{1}{30} \int_0^{30} \text{Acc}(\tau) d\tau, \quad (3)$$

where $\text{Acc}(\tau)$ represents the fraction of camera pairs with both angular errors smaller than τ .

C. Video Demonstration

Figure 1 shows a sampled image from the video demo that demonstrates our model’s reconstruction of 3D scene geom-

etry on the validation set. Given multi-frame, multi-view images as input, **DVGT** generates dense point maps which recover scene geometry with high fidelity and consistency, validating the effectiveness of our method.

D. Dataset Details

We utilize five diverse datasets for training and evaluation: Waymo [10], nuScenes [1], OpenScene [2], DDAD [4], and KITTI [3]. Detailed statistics for each dataset are summarized in Table 1. All data sequences are downsampled to a temporal frequency of 2Hz, and the frame counts reported in the table are calculated based on this sampling rate.

During the training phase, the sampling ratio across datasets is set as *nuScenes* : *KITTI* : *OpenScene* : *Waymo* : *DDAD* = 6 : 5 : 77 : 6 : 6. For each training iteration, we first select a dataset according to these weights and sample a batch from it. To ensure robust feature learning across various sensor configurations, we implement a dynamic sampling strategy:

- A target aspect ratio is randomly sampled from the range $[1.5, 3.3]$.
- The number of camera views is randomly selected from $[2, 8]$.
- Given a hardware constraint of 48 images per GPU per iteration, we determine the maximum possible sequence length T_{max} .
- A specific frame number is then sampled from $[2, T_{max}]$, and the final batch size is calculated to saturate the GPU memory efficiency.

E. Author Contributions

- Sicheng Zuo and Zixun Xie implemented the model, constructed the dataset, designed and conducted the experiments, and wrote the manuscript.
- Wenzhao Zheng proposed this project, designed the approach, supervised the experiments, and revised the manuscript.
- Shaoqing Xu participated in the method discussion, provided computing resources, and assisted in writing the manuscript.
- Fang Li, Shengyin Jiang, Long Chen, and Zhi-Xin Yang provided GPU and infrastructure support.
- Jiwen Lu supervised the overall project.

References

- [1] 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*, pages 11621–11631, 2020. 2
- [2] OpenScene Contributors. Openscene: The largest up-to-date 3d occupancy prediction benchmark in autonomous driving. In *CVPR*, pages 18–22, 2023. 2
- [3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 32(11):1231–1237, 2013. 2
- [4] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, pages 2485–2494, 2020. 2
- [5] Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *EMNLP*, pages 4246–4253, 2020. 1
- [6] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 1
- [7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1
- [8] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pages 12179–12188, 2021. 1
- [9] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 1
- [10] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, pages 2446–2454, 2020. 2
- [11] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *ICCV*, pages 32–42, 2021. 1
- [12] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *ICCV*, pages 9773–9783, 2023. 2
- [13] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, pages 5294–5306, 2025. 1
- [14] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3 : Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025. 1
- [15] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xianggang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *NeurIPS*, 37:21875–21911, 2024. 1