DVGT: Driving Visual Geometry Transformer

Sicheng Zuo[1,*] Zixun Xie[1,2,4,*] Wenzhao Zheng[1,*,✉] Shaoqing Xu[2,3,†]
Fang Li[2,3] Shengyin Jiang[2] Long Chen[2] Zhi-Xin Yang[3] Jiwen Lu[1]

[1]Tsinghua University [2]Xiaomi EV [3]University of Macau [4]Peking University

Project Page: https://wzzheng.net/DVGT
Large Driving Models: https://github.com/wzzheng/LDM

Figure 1. **DVGT is a large visual geometry transformer specifically designed for autonomous driving**. It accepts a sequence of unposed multi-view images and predicts a metric-scaled global 3D point map in the ego-centric coordinate system and the ego poses for each frame, which outperforms the other geometry prediction models without post-alignment with external sensors.

Abstract

Perceiving and reconstructing 3D scene geometry from visual inputs is crucial for autonomous driving. However, there still lacks a driving-targeted dense geometry perception model that can adapt to different scenarios and camera configurations. To bridge this gap, we propose a Driving Visual Geometry Transformer (DVGT), which reconstructs a global dense 3D point map from a sequence of unposed multi-view visual inputs. We first extract visual features for each image using a DINO backbone, and employ alternating intra-view local attention, cross-view spatial attention, and cross-frame temporal attention to infer geometric relations across images. We then use multiple heads to decode a global point map in the ego coordinate of the first frame and the ego poses for each frame. Unlike conventional methods that rely on precise camera parameters, DVGT is free of explicit 3D geometric priors, enabling flexible processing of arbitrary camera configurations. DVGT directly predicts metric-scaled geometry from image sequences, eliminating the need for post-alignment with external sensors. Trained on a large mixture of driving datasets including nuScenes, OpenScene, Waymo, KITTI, and DDAD, DVGT significantly outperforms existing models on various scenarios. Code is available at https://github.com/wzzheng/DVGT.

1. Introduction

Vision-centric autonomous driving has been widely explored due to its economic advantages and human-like sensing capabilities [5, 6, 11, 12, 14, 18, 26, 44, 56–60]. Its cornerstone is the accurate perception of 3D scene geometry from visual inputs, where recent works have shown promising results with considerable potential [14, 16, 26, 33, 48].

Most existing methods focus on single-frame depth prediction [2, 8, 22, 27, 47, 53] or 3D occupancy predic-

*Equal contributions. †Project leader. ✉Corresponding author.

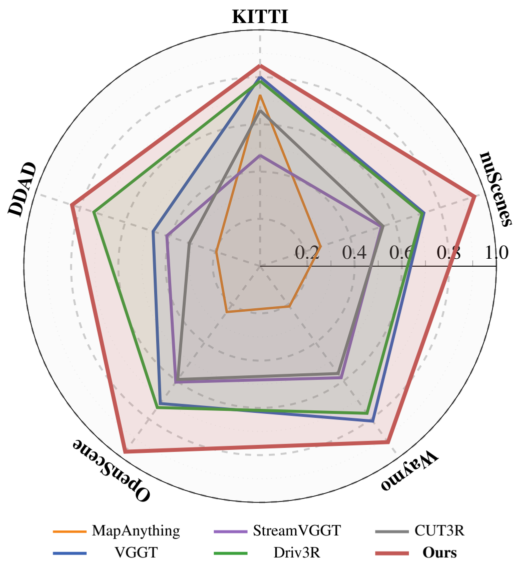

Figure 2. **Quantitative comparison of 3D scene reconstruction.** Our method (red) demonstrates superior accuracy ($\delta < 1.25$ for ray depth estimation) across all evaluated datasets.

tion [14–16, 33, 48, 50, 54, 55], which require ground-truth poses for temporal fusion to achieve global 3D scene geometry understanding. They usually rely on strong geometric priors and employ explicit 2D-to-3D projection to obtain 3D geometry [13, 14, 16, 23, 25, 26, 36]. This design choice tightly couples the model design to specific sensor configurations. Also, most existing models are trained on limited data with a single camera configuration, hindering the adaptability and scalability across different vehicles and scenarios [24, 41]. Although recent general visual geometry models [37, 45, 49, 61] have shown strong 3D reconstruction performance, there still lacks a dense visual geometry model specifically designed for autonomous driving that can adapt to different scenarios and camera configurations.

To bridge this gap, we propose a Driving Visual Geometry Transformer (DVGT) to generate global dense 3D point maps in an end-to-end manner, offering a continuous and high-fidelity geometric representation of the scene, as shown in Figure 1. Given a sequence of unposed multi-view visual inputs, we first extract visual features for each image with a vision foundation model and then employ alternating intra-view local attention, cross-view spatial attention, and cross-frame temporal attention to infer geometric relations across images. We add multiple heads to jointly predict the ego pose and the global 3D point map at the metric scale. Since the multi-view images are captured by fixed surrounding cameras in the ego car, we directly generate the global point map in the ego car coordinate of the first frame and predict the ego pose for each frame instead of the camera pose for each image. Our model does not contain any spatial inductive bias and thus has more flexibility to adapt to different camera configurations. For training, we construct a large-scale and diverse mixture of driving datasets, including nuScenes, OpenScene, Waymo, KITTI, and DDAD [1, 3, 7, 9, 31], and generate dense geometric pseudo ground truths by aligning general-purpose monocular depth models [40] with projected sparse LiDAR depth. Extensive experiments show that our DVGT demonstrates superior performance over both general-purpose and driving-specific methods for 3D scene geometry reconstruction on various driving scenarios, as shown in Figure 2. DVGT directly predicts metric-scaled global 3D point maps without the need for post-alignment with external sensors.

2. Related Work

Geometry Perception for Autonomous Driving. Early methods explored depth prediction from multi-view images [2, 8, 10, 47] to capture scene geometry. Yet, these methods often provide a 2.5-D representation and struggle with producing a unified 3D scene. To achieve a complete understanding of the scene geometry, TPVFormer [14] proposed 3D occupancy prediction, which uses dense voxels to describe the fine-grained geometry. Subsequent works have followed this direction, including voxel-based methods [33, 43, 48, 55, 62] and object-centric methods [16, 17, 29, 32, 36, 63, 64]. However, discrete occupancy grids introduce additional quantization errors (typically around 0.5m) [33, 43], which pose a challenge to represent geometry details accurately. In contrast, our DVGT predicts dense, continuous 3D point maps, enabling a fine-grained and complete representation of the scene geometry.

Generalization Across Camera Configurations. The ability of driving perception models to generalize across different camera configurations is crucial for practical deployment. Conventional methods [13, 14, 16, 26] rely on explicit 2D-to-3D geometric projection to constrain the interaction between 2D images and 3D representations. This creates a strong dependency on camera-specific priors, which fundamentally limits their ability to generalize across different camera setups [41]. To mitigate this, recent work [24] attempts cross-sensor generalization by projecting images into a unified virtual camera space. However, this approach remains dependent on geometric projection, and its performance degrades as camera parameters vary widely. In contrast, our proposed DVGT is a universal model that is adaptable to different camera configurations, which enables unified scaling on data from diverse sources.

General Visual Geometry Models. Recent general visual geometry models have shown significant progress, which reconstruct 3D point maps from images [19, 20, 35, 37, 38, 42, 45, 49, 61]. This paradigm, pioneered by DUSt3R [42] for image pairs and advanced by following-up works [20, 35, 38], has been extended to multi-view reconstruction across various scenes by recent models [37, 45]. However, most models can only recover scene geometry

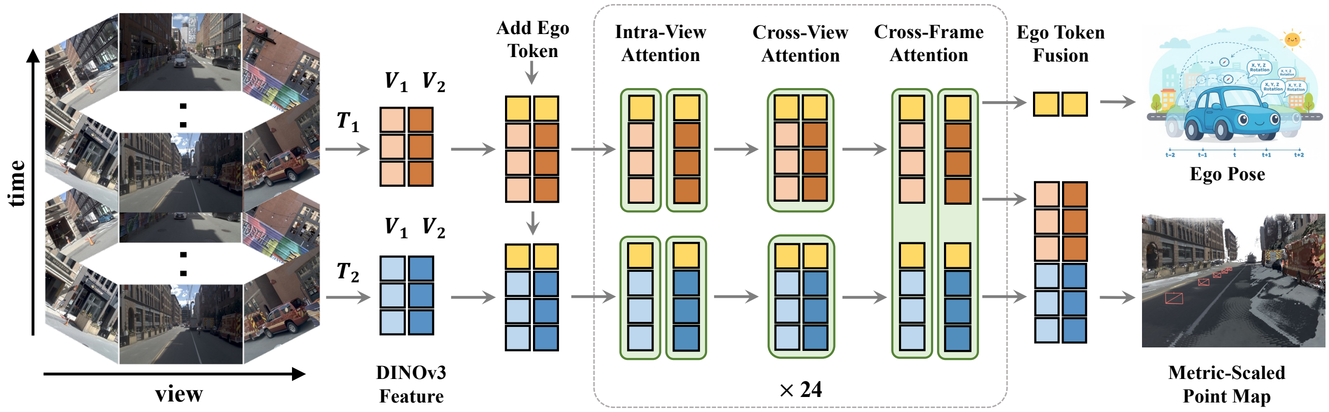

Figure 3. **Framework of our DVGT for metric-scaled 3D scene geometry prediction.** Given multi-frame, multi-view images, DVGT predicts metric-scaled 3D point map in the ego coordinate system of the first frame and the ego pose from each frame to the first frame.

at a relative scale and require post-alignment with external data like LiDAR points to obtain the metric scale of the scene [35, 37, 42, 45, 61]. In contrast, our DVGT directly predicts metric-scaled scene geometry, which can be directly exploited in downstream tasks. Also, existing models treat the multi-frame and multi-view images equally and estimate the geometry and pose for each image. Instead, we employ a spatial-temporal architecture to efficiently process the driving inputs and directly generate the ego-aligned global point maps and ego poses in an end-to-end manner.

3. Proposed Approach

3.1. Ego-centric 3D Point Map Reconstruction

We first formulate the driving scene geometry perception as **ego-centric** 3D Point Map Reconstruction. This task involves predicting a continuous, metric-scaled 3D coordinate $\hat{\mathbf{P}}_{u,v} = (x, y, z)$ in a unified coordinate system for each pixel (u, v) of an input image $\mathbf{I}$. The resulting dense and continuous 3D point map $\hat{\mathbf{P}} = \{\hat{\mathbf{P}}_{u,v}\}_{u=1..H,v=1..W}$, where H, W are the image resolution, yields two key advantages: 1. **High-Fidelity**: Its continuous coordinates eliminate quantization errors, enabling the precise modeling of scene geometry. 2. **Completeness**: Its pixel-aligned density ensures full coverage of visible regions, unifying the foreground objects and the background environment.

Conventional visual geometry models often reconstruct 3D point maps in a reference camera coordinate system [19, 37]. They tightly couple the model output to the reference camera intrinsics and extrinsics, hindering its adaptability to diverse sensor setups. To resolve this, we decouple the scene geometry representation from camera parameters by predicting 3D point maps in the ego-vehicle coordinate system of a reference frame. This ego-centric formulation yields a unified scene geometry representation that is invariant to camera focal lengths, poses, and the number of views, which is critical to a universal driving perception model.

Formally, given an input image sequence $\mathcal{I} = \{\mathbf{I}_{t,n}\}_{t=1..T,n=1..N}$ from T frames and N views per frame, our model $\mathcal{M}$ jointly predicts the 3D point maps $\mathcal{P} = \{\hat{\mathbf{P}}_{t,n}\}_{t=1..T,n=1..N}$ and the ego-vehicle pose sequence $\mathcal{T}_{\text{ego}} = \{\hat{\mathbf{T}}_t\}_{t=1..T}$. Here, $\hat{\mathbf{P}}_{t,n} \in \mathbb{R}^{H\times W\times 3}$ is the point map for image $\mathbf{I}_{t,n}$, with all points expressed in the unified ego-vehicle coordinate system of the reference frame. $\hat{\mathbf{T}}_t \in \text{SE}(3)$ is the ego motion from the reference frame to the frame t. The overall mapping is expressed as:

$$(\mathcal{P}, \mathcal{T}_{\text{ego}}) = \mathcal{M}(\mathcal{I}). \tag{1}$$

3.2. Spatial-Temporal Geometry Transformer

To achieve this goal, we propose **DVGT**, a universal driving visual geometry transformer. As shown in Figure 3, our model consists of three components: an image encoder $\mathcal{E}$, a geometry transformer $\mathcal{F}$, and a set of prediction heads $\mathcal{H}$.

Overall Framework. Given multi-frame, multi-view image inputs, we adopt a pretrained vision foundation model [30] to convert each input image $\mathbf{I}_{t,n}$ into a set of image tokens:

$$\mathbf{F}_{t,n} = \mathcal{E}(\mathbf{I}_{t,n}). \tag{2}$$

To facilitate pose prediction, we augment each image's tokens with a dedicated, learnable ego token $\mathbf{E}_{t,n}$:

$$\mathbf{Z}_{t,n} = \text{concat}([\mathbf{F}_{t,n}, \mathbf{E}_{t,n}]), \tag{3}$$

where concat denotes the concatenation operation. To distinguish tokens from different frames, all tokens are then augmented with temporal positional embeddings:

$$\mathbf{Z}'_{t,n} = \mathbf{Z}_{t,n} + \mathbf{Pos}_t. \tag{4}$$

where $\mathbf{Pos}_t$ denotes the temporal positional embeddings of the frame t. The resulting sequence $\mathcal{Z} = \{\mathbf{Z}'_{t,n}\}_{t=1..T,n=1..N}$ is then fed into the geometry transformer $\mathcal{F}$ and subsequently prediction heads $\mathcal{H}$ for joint 3D point map and ego pose prediction:

$$(\mathcal{P}, \mathcal{T}_{ego}) = \mathcal{H}(\mathcal{F}(\mathcal{Z})). \tag{5}$$

Geometry Transformer. Most existing visual geometry models rely on global attention for interaction among all image tokens. They incurs substantial computational costs and

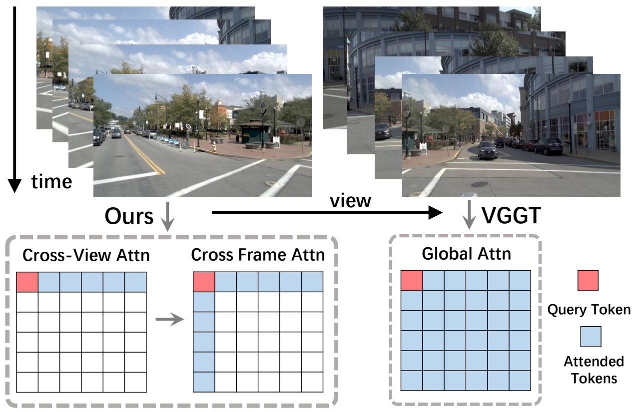

Figure 4. **Comparison of our factorized spatial-temporal attention (left) against the global attention used in VGGT (right).**

are infeasible for real-time autonomous driving. To leverage the strong spatial-temporal structure in visual inputs of driving scenarios, we introduce an efficient factorized attention mechanism. Specifically, the geometry transformer $\mathcal{F}$ is composed of L cascaded blocks, where each block sequentially performs three targeted attention operations:

- **Intra-View Local Attention** operates exclusively on tokens within each image to refine local features.
- **Cross-View Spatial Attention** aggregates spatial information by attending to tokens across different views within the same frame.
- **Cross-Frame Temporal Attention** captures consistent statics and temporal dynamics by attending to tokens from the same view across different frames.

As shown in Figure 4, this factorization efficiently factorizing the computationally expensive global attention while maintaining effective spatial-temporal information fusion.

Prediction Heads. The geometry transformer $\mathcal{F}$ outputs refined image tokens $\mathbf{F}'_{t,n}$ and ego tokens $\mathbf{E}'_{t,n}$. We then employ a 3D point map head $\mathcal{H}_{\text{point}}$ to decode the image tokens into the metric-scaled 3D point map:

$$\hat{\mathbf{P}}_{t,n} = \mathcal{H}_{\text{point}}(\mathbf{F}'_{t,n}). \tag{6}$$

For pose prediction, the ego tokens from all views within the same frame are first aggregated into a global representation:

$$\bar{\mathbf{E}}_t = \sum_{n=1}^{N} \mathbf{E}'_{t,n}. \tag{7}$$

This representation is then fed into a pose head $\mathcal{H}_{\text{pose}}$ to regress the ego pose of the frame t:

$$\hat{\mathbf{T}}_t = \mathcal{H}_{\text{pose}}(\bar{\mathbf{E}}_t). \tag{8}$$

Prior-Free Design. A core principle of DVGT is the 3D prior-free design. Unlike traditional driving perception models that rely on explicit camera parameters, our model is structurally independent of camera parameters and 2D-to-3D geometric projection designs. Instead, it learns to infer the 3D scene geometry directly from 2D image features in a purely data-driven, end-to-end manner. This fundamental design choice decouples the model from specific camera configurations, thereby granting it robust adaptability to diverse camera setups and driving scenarios.

3.3. Dense Geometry Ground Truth Construction

Accurate 3D reconstruction in autonomous driving is hindered by the scarcity of dense geometry ground truths. While aligning monocular depth predictions [51, 52] with projected sparse LiDAR depth [46] can generate pseudo-labels, it is often unreliable due to two primary factors: (1) the suboptimal generalization of general-purpose depth models in complex driving scenarios, and (2) the spatial unevenness of sparse LiDAR points, which often concentrates in small regions. These issues can lead to alignment failures or significant bias in estimated scale and shift parameters, resulting in grossly inaccurate pseudo-labels. To address this, we develop a robust pipeline to filter out samples where the depth model or alignment fails.

Failure Pattern Analysis. We conducted a rigorous analysis of failure cases in the pseudo-labeling process. We identified five dominant failure patterns: (a) **Semantic Misinterpretation:** Large, low-texture surfaces (e.g., truck trailers) are occasionally misclassified as sky. (b) **Photometric Instability:** Exposure issues can lead to stochastic depth predictions. (c) **Structural Ambiguity:** Complex planar patterns (e.g., billboards) are misinterpreted as having depth variance. (d) **Motion Artifacts:** Blur from high-speed driving or camera jitter degrades accuracy. (e) **Alignment Ill-Conditioning:** Extremely sparse or concentrated LiDAR points cause the alignment optimization to be ill-posed, resulting in significant errors for alignment.

Threshold Filtering. We evaluated multiple methods and selected MoGe-2 [40] for depth annotation and ROE [39] for alignment. We established thresholds based on metric statistics to filter out low-quality pseudo-labels. Our strategy includes: (1) Valid Point Overlap: We compute the ratio of valid LiDAR points that are also predicted as valid by the depth model. This effectively identifies and removes samples suffering from patterns (a) and (b). (2) Standard Depth Metrics: We use Absolute Relative Error (Abs Rel) and $\delta < 1.25$ metrics to discard poor predictions caused by planar structures (c) or motion blur (d). (3) Alignment Quality Metrics: To address (e), we filter out images with insufficient projected LiDAR points or low spatial variance among those points, and constrain the output scale and shift parameters from the alignment algorithm.

Large-Scale Dataset. We apply this low-quality pseudo-label filtering pipeline to generate dense and accurate 3D point maps for a large-scale, mixed-domain dataset by aggregating five public autonomous driving benchmarks: Waymo [31], nuScenes [1], OpenScene [3], DDAD [9], and KITTI [7]. As illustrated in Figure 5, our annotations capture complex scene geometry with high fidelity. By training and validating our model on this mixed dataset, we

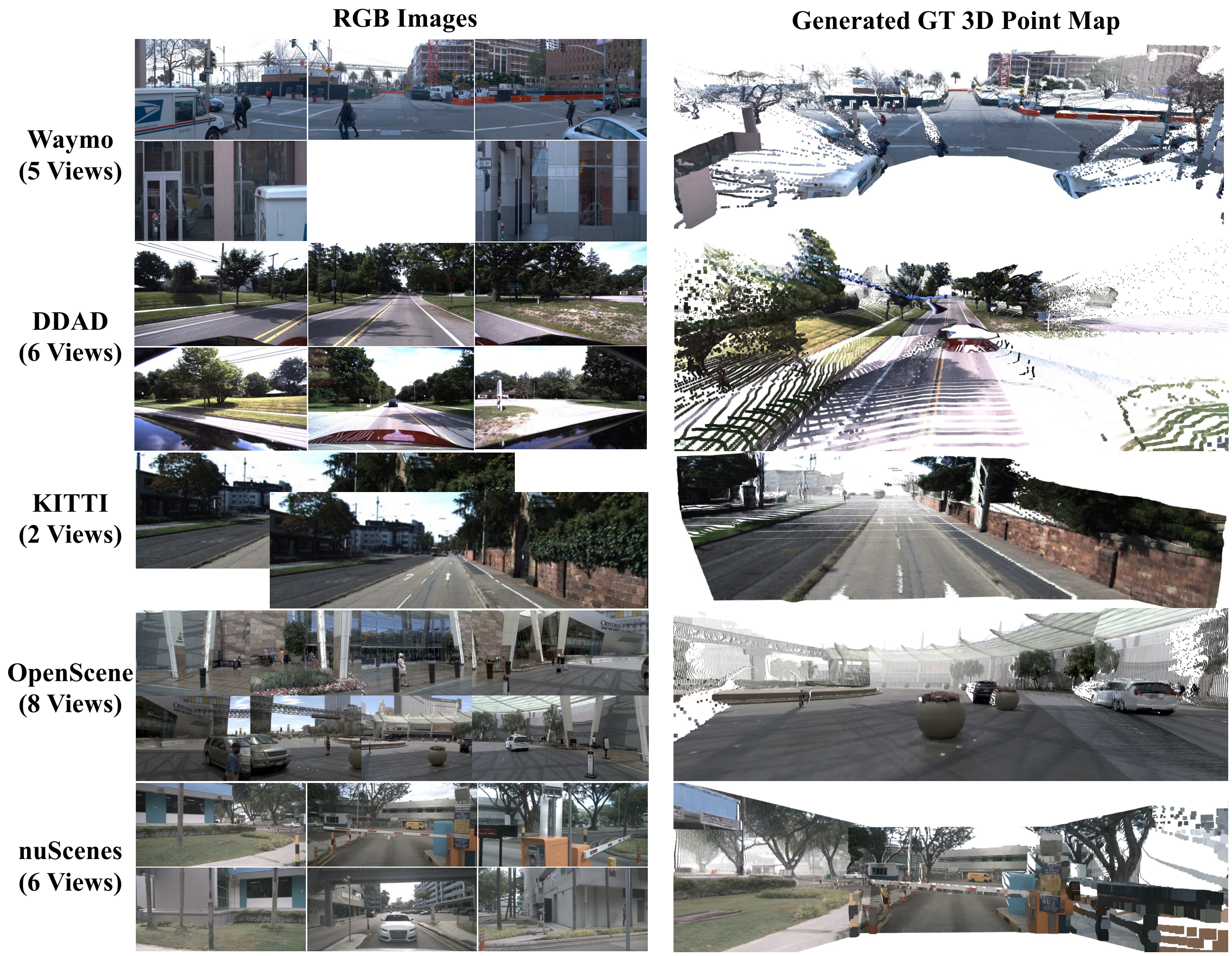

Figure 5. **We construct dense and accurate ground-truth 3D point maps for diverse driving scenarios.** This figure showcases examples from the Waymo [31], nuScenes [1], OpenScene [3], DDAD [9], and KITTI [7] datasets. For each scene, we display the multi-view RGB images (left) and the corresponding high-quality 3D point map (right), highlighting the diversity and precision of our data.

demonstrate its strong generalization ability and effectiveness across diverse, real-world driving scenarios.

3.4. Training of DVGT

Training Losses. We train the DVGT in an end-to-end manner using a multi-task loss:

$$\mathcal{L} = \lambda \mathcal{L}_{\text{epose}} + \mathcal{L}_{\text{pmap}}, \tag{9}$$

where both the ego pose loss $\mathcal{L}_{\text{epose}}$ and the point map loss $\mathcal{L}_{\text{pmap}}$ are metric-scaled. However, the numerical range of the point map values is significantly larger than that of the ego pose. We therefore set the ego pose loss weight $\lambda = 5.0$ to strike a balance. We describe each loss term in turn.

The ego pose loss $\mathcal{L}_{\text{epose}}$ supervises the predicted ego motion sequence $\mathcal{T}_{\text{ego}}$. We apply a standard L1 loss to the 7-dimensional pose representation, consisting of a 3D translation vector and a 4D rotation quaternion. We compute the difference between the predicted pose $\hat{\mathbf{T}}_t$ and the ground truth $\mathbf{T}_t$ for each frame t. It can be described as below:

$$\mathcal{L}_{\text{epose}} = \frac{1}{T} \sum_{t=1}^{T} \|\hat{\mathbf{T}}_t - \mathbf{T}_t\|_1. \tag{10}$$

We adopt the same loss functions as VGGT [37] for 3D point map supervision, which can be expressed as:

$$\mathcal{L}_{\text{pmap}} = \sum_{t,n} \Big(\|\mathbf{\Sigma}^P_{t,n} \odot (\hat{\mathbf{P}}_{t,n} - \mathbf{P}_{t,n})\|_2 + \|(\nabla \hat{\mathbf{P}}_{t,n} - \nabla \mathbf{P}_{t,n})\|_2 - \alpha \log \mathbf{\Sigma}^P_{t,n} \Big), \tag{11}$$

where the sum $\sum_{t,n}$ iterates over all input images across T frames and N views. $\hat{\mathbf{P}}_{t,n}$ is the predicted 3D point map and $\mathbf{P}_{t,n}$ is the corresponding ground-truth. $\mathbf{\Sigma}^P_{t,n}$ is the additionally predicted per-pixel uncertainty map, measuring the uncertainty of the model's prediction for each pixel. $\odot$ refers to the channel-broadcast element-wise product. ∇ denotes the 2D spatial gradient operator, and $\|\cdot\|_2$ is the L2 norm. The final term $-\alpha \log \mathbf{\Sigma}^P_{t,n}$ is a regularizer that encourages the model to be confident (with low uncertainty prediction), and we set the hyperparameter $\alpha = 2.0$.

Table 1. **Quantitative 3D reconstruction results across diverse datasets.** * indicates that the predicted 3D point map is aligned with the sparse LiDAR depth using the Umeyama [34] algorithm to recover the metric scale. The inference time test is conducted on 128 images (16 frames $\times$ 8 views). Note that for CUT3R [38] and StreamVGGT [61], we report the cumulative inference time for streaming all images.

Method	KITTI		NuScenes		Waymo		OpenScene		DDAD		Time
	Acc ↓	Comp ↓	Acc ↓	Comp ↓	Acc ↓	Comp ↓	Acc ↓	Comp ↓	Acc ↓	Comp ↓	
CUT3R* [38]	0.965	2.050	2.054	2.603	3.391	4.216	1.864	2.258	2.774	4.677	$\sim$5.6s
VGGT* [37]	1.154	1.294	1.300	1.498	1.641	2.053	1.422	1.496	1.741	2.473	$\sim$13.7s
MapAnything [19]	1.880	**1.014**	4.499	4.886	10.205	8.494	3.353	4.303	8.015	8.493	$\sim$5.8s
StreamVGGT* [61]	3.421	2.196	2.588	2.414	3.630	3.275	2.304	2.098	2.717	2.788	$\sim$31.0s
Driv3R* [4]	0.864	1.083	0.742	1.345	**0.800**	**1.311**	0.884	1.693	0.950	1.259	$\sim$9.0s
DVGT	**0.846**	1.468	**0.457**	**0.494**	1.714	2.216	**0.402**	**0.481**	**0.751**	**1.009**	**$\sim$4.0s**

Table 2. **Quantitative ray depth results across diverse datasets.**

Method	KITTI		NuScenes		Waymo		OpenScene		DDAD	
	Abs Rel ↓	δ<1.25 ↑	Abs Rel ↓	δ<1.25 ↑	Abs Rel ↓	δ<1.25 ↑	Abs Rel ↓	δ<1.25 ↑	Abs Rel ↓	δ<1.25 ↑
CUT3R [38]	0.217	0.659	0.332	0.547	0.291	0.562	0.278	0.593	0.870	0.315
VGGT [37]	0.158	0.801	0.243	0.729	0.176	0.811	0.241	0.719	0.613	0.476
MapAnything [19]	0.188	0.725	0.568	0.269	0.507	0.211	0.486	0.240	1.971	0.195
StreamVGGT [61]	0.362	0.469	0.412	0.540	0.339	0.584	0.319	0.607	0.838	0.415
Driv3R [4]	0.164	0.784	0.189	0.721	0.168	0.770	0.188	0.740	0.185	0.740
DVGT	**0.136**	**0.849**	**0.069**	**0.953**	**0.106**	**0.921**	**0.049**	**0.971**	**0.152**	**0.837**

Table 3. **Quantitative ego pose results across diverse datasets.**

Method	KITTI AUC@30 ↑	NuScenes AUC@30 ↑	Waymo AUC@30 ↑	OpenScene AUC@30 ↑	DDAD AUC@30 ↑
CUT3R [38]	51.8	43.5	50.1	34.7	48.6
VGGT [37]	**96.9**	**87.8**	**87.7**	66.3	92.8
MapAnything [19]	90.6	85.0	82.8	65.6	87.0
StreamVGGT [61]	95.8	86.2	85.6	74.1	91.9
DVGT	87.6	86.5	86.4	**74.7**	**95.1**

4. Experiments

We conduct extensive experiments to evaluate the performance of our DVGT. We provide dataset, implementation, and evaluation details in the appendix.

4.1. 3D Reconstruction and Depth Estimation

As shown in Table 1 and Table 2, we evaluate the 3D point reconstruction accuracy and ray depth accuracy (defined as the distance from a 3D point to the ego-vehicle center). Our analysis yields several key insights. First, DVGT demonstrates significantly superior 3D point and depth accuracy across driving datasets when compared to general visual geometry models [19, 37] and driving geometry models [4]. This highlights the effectiveness of our architecture in reformulating the autonomous driving perception problem.

A crucial distinction is the handling of scale. Models like CUT3R [38] and VGGT [37] necessitate post-hoc alignment with ground-truth point clouds to recover the metric scale. In contrast, DVGT directly predicts accurate, metric-scaled scene structure in a single, end-to-end forward pass. While MapAnything [19] also predicts metric scale without post-processing, its reconstruction error is substantially larger (e.g., 4.303 vs. 0.481 for DVGT on OpenScene [3]), underscoring the superior accuracy of our approach.

Furthermore, we note that DVGT's performance on Waymo [31] is less competitive. We attribute this to disproportionate data sampling: Waymo [31] was assigned the same weight as smaller datasets despite being 5× larger during training. Unlike nuScenes [1], which benefits from a distribution similar to the dominant OpenScene [3] data, Waymo lacks this alignment. We expect that optimizing the sampling weights of different datasets will bridge this gap.

4.2. Ego-Pose Estimation

In addition to scene geometry, DVGT jointly estimates ego pose. Table 3 presents the results for pose prediction. Our model achieves better results on OpenScene [3] and DDAD [9], while attaining comparable performance to VGGT [37] on nuScenes [1] and Waymo [31]. This demonstrates that DVGT effectively serves as a comprehensive visual geometry model, capable of accurately estimating joint scene structure and ego-vehicle pose.

We note that DVGT's pose prediction performance on KITTI [7] is slightly lower. We attribute this to KITTI's high-overlap dual-camera setup, which limits full 3D and ego-motion understanding compared to surround-view data.

4.3. Comparison with Driving Models

We perform a comparison with existing driving models for depth estimation on the widely used nuScenes [1] dataset. Our model directly reconstructs 3D point maps, whereas most driving models predict depth maps that are evaluated against sparse LiDAR ground truth. For a fair comparison, we convert our 3D point map prediction into depth maps

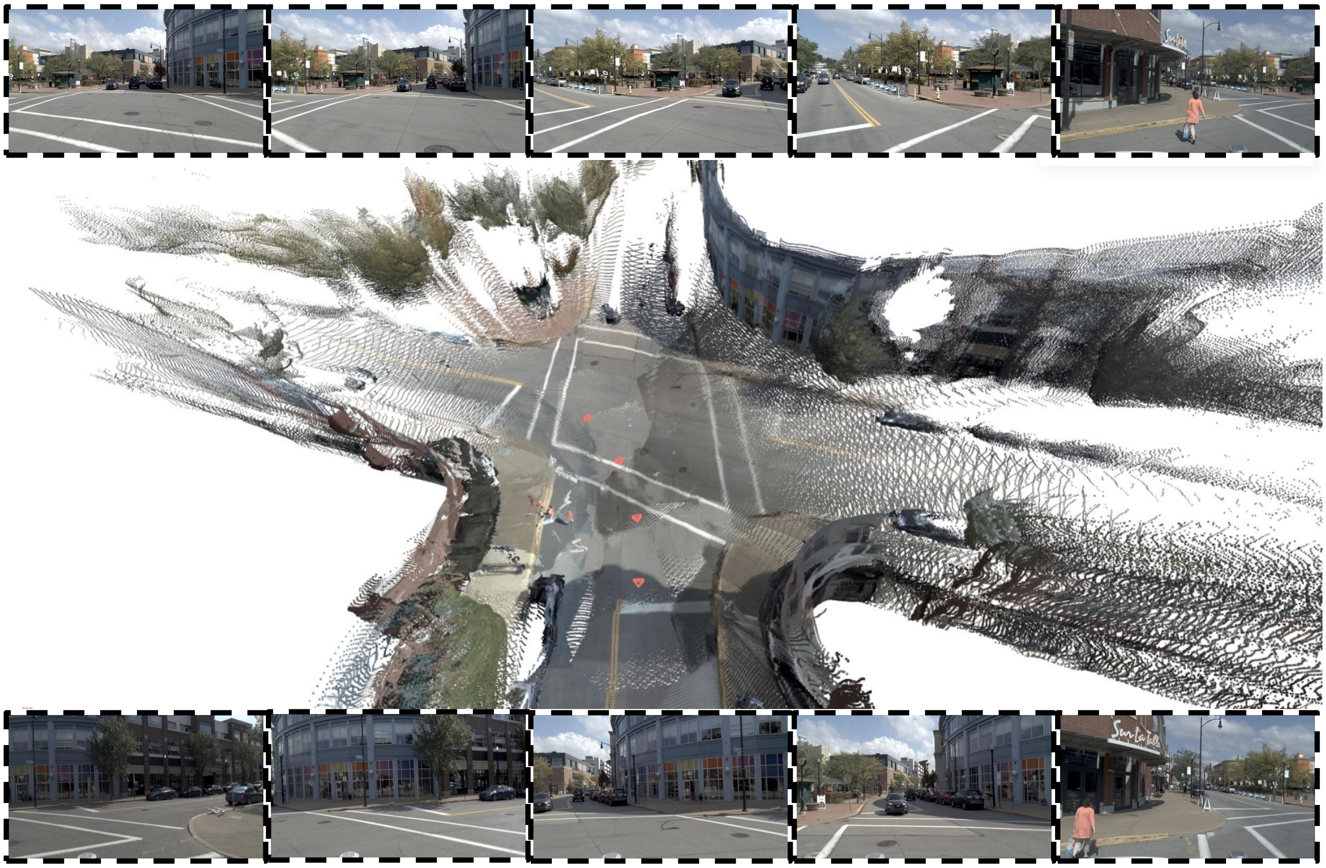

Figure 6. **Visualization of our DVGT on 3D point map reconstruction.**

Table 4. **Comparison with driving models for depth estimation compared with LiDAR GT on the nuscenes [1] dataset.**

Method	Scale Method	Abs Rel ↓	$\delta < 1.25$ ↑
MonoDepth2 [8]	Median Scaling	0.29	0.64
SurroundDepth [47]	SfM Pretrain	0.28	0.66
R3D3 [28]	Extrinsic	0.25	0.73
SelfOcc [15]	Pose GT	0.23	0.75
Dist4D [10]	Pose GT	0.39	0.58
OmniNWM [21]	Pose GT	0.23	0.81
DVGT	None	**0.13**	**0.86**

and compute the same metrics.

Table 4 shows that our DVGT achieves state-of-the-art performance, obtaining the best scores on both Abs Rel and $\delta < 1.25$ accuracy. Note that while other methods require post-processing like median scaling or depend on ground-truth camera poses for scale recovery, DVGT operates without any post-alignment or pose information. This demonstrates the effectiveness of our model in recovering metrically accurate 3D geometry directly from images.

4.4. Experimental Analysis

Ablation Study on Scene Scale. A significant challenge in autonomous driving perception is the vast dynamic range of scene geometry, with distances often exceeding 100 meters. Directly regressing large-valued 3D coordinates can introduce numerical instability during training, as model parameters may be pushed to large magnitudes, potentially leading to performance degradation.

Table 5. **Ablation study on ground truth coordinate normalization on the nuScenes [1] dataset.** 1, 10, and 100 denote linearly scale dividing by 1, 10, and 100, respectively. And asinh denotes the arcsinh function for non-linearly scaling.

Method	Acc ↓	Comp ↓	Abs Rel ↓	δ<1.25 ↑	AUC@30 ↑
1 (base)	1.584	1.424	0.261	0.676	68.4
10	**1.349**	**1.053**	**0.195**	**0.756**	79.8
100	1.646	1.431	0.257	0.694	80.7
asinh	1.411	1.390	0.222	0.719	**80.8**

Therefore, we investigate strategies for scaling the target 3D points to a more stable numerical range for regression. As shown in Table 5, we conduct an ablation study comparing three distinct scaling methods:

- Linear Scaling (10x): Linearly dividing all target coordinates by 10.
- Linear Scaling (100x): Linearly dividing all target coordinates by 100.
- Non-linear Scaling (arcsinh): Applying the arcsinh function to non-linearly compress the coordinate range.

The experimental results indicate that linear scaling by 10x achieves the best performance. We analyze that scaling

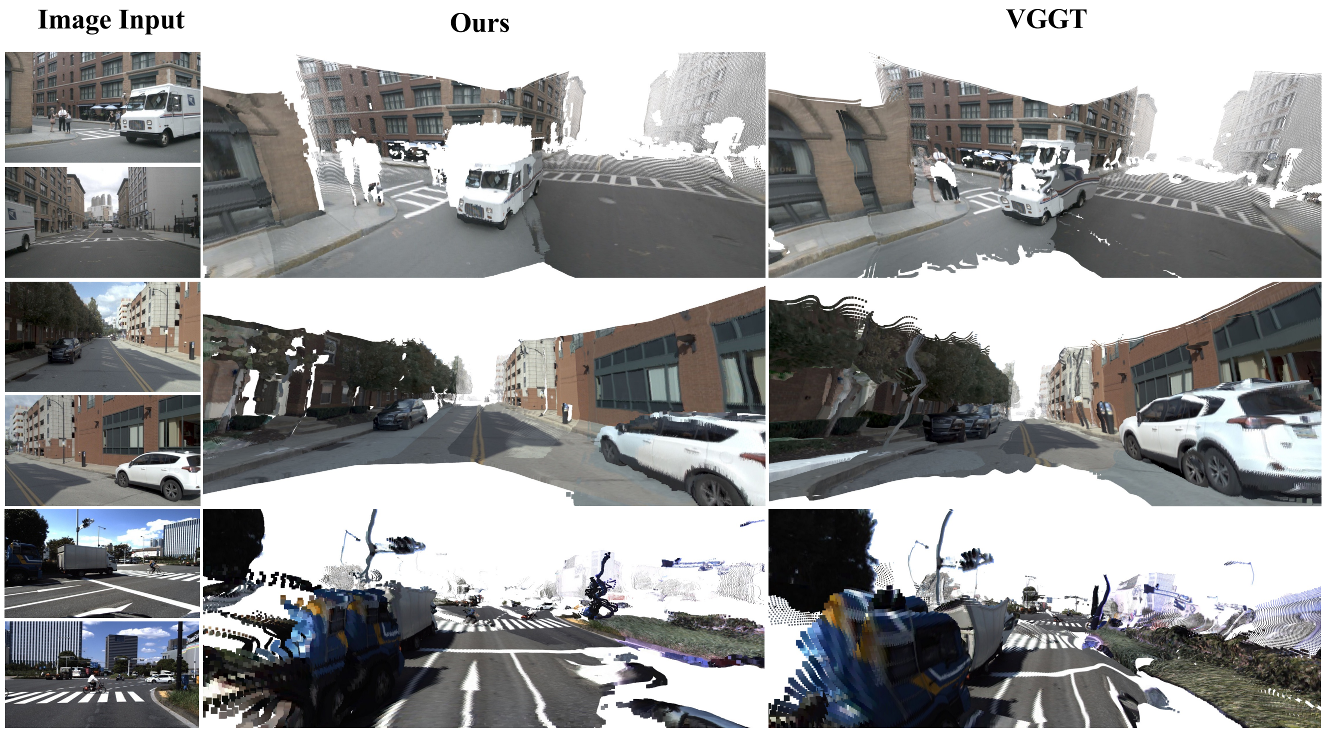

Figure 7. **Comparison of our DVGT with VGGT on 3D metric-scaled point map reconstruction.** Notably, VGGT performed post-alignment with LiDAR points GT while our DVGT eliminates the need for any post-processing.

Table 6. **Ablation study on attention mechanisms on the nuScenes [1] dataset.** We compare combinations of different attention components: **G** (Global Attention), **L** (Intra-View Local Attention), **S** (Cross-View Spatial Attention), **T** (Cross-Frame Temporal Attention), and **TE** (Temporal Positional Embedding).

Method	Acc ↓	Comp ↓	Abs Rel ↓	δ<1.25 ↑	AUC@30 ↑	Time
L+G	**1.131**	**1.129**	**0.178**	**0.789**	74.6	~8.2s
L+S+T	1.584	1.424	0.261	0.676	68.4	**~4.0s**
L+S+T+TE	1.458	1.139	0.227	0.725	**77.6**	**~4.0s**

by 100x makes the scene scale too small, which degrades precision for near-field geometry. The arcsinh transformation offers an adaptive scaling (less compression for near), but this non-linear operation may distort the inherent geometric structure, resulting in sub-optimal performance.

Ablation Study on Attention Design. We perform an ablation study on the attention mechanism within our geometry transformer in Table 6. We compare our decomposed attention mechanism against a standard global attention baseline, where all tokens interact with all others. Our method yields a significant improvement in efficiency and a faster inference speed. This decomposition, however, introduces a minor performance trade-off. To mitigate this, we incorporate temporal positional embeddings, providing the model with explicit sequential information. This addition successfully narrows the performance gap, allowing our model to achieve a compelling balance between high performance and computational efficiency.

Qualitative Analysis. We provide qualitative visualizations to further analyze the capabilities of our model. In Figure 6, the visualization confirms that the ego-pose is accurately estimated. Furthermore, the model exhibits high multi-frame and multi-view consistency for static elements, like road surfaces and buildings. Dynamic objects, such as pedestrians and vehicles at the intersection, are also faithfully reconstructed, while fine-grained details, including roadside trees, are well-preserved.

Figure 7 shows that VGGT struggles with both static elements (e.g., trees, lanes) and dynamic objects (e.g., vehicles, pedestrians), often failing to capture accurate geometry. In contrast, our DVGT, leveraging dense annotations and large-scale training, successfully recovers scene geometry while maintaining high efficiency.

5. Conclusion

In this paper, we introduced DVGT, a universal visual geometry model for autonomous driving. DVGT generates a dense, metric-scaled 3D global point map from unposed images, enabling an accurate and complete understanding of scene geometry. Its 3D prior-free transformer architecture enables adaptability to different camera configurations, overcoming a major bottleneck for scalability. To validate our approach and facilitate future research, we also constructed a large-scale driving dataset with dense 3D point annotations. Experiments demonstrate that DVGT significantly outperforms existing methods, paving the way for more robust and versatile vision-centric driving systems.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62576188, Grant 62336004, Grant 62321005, and Grant 62125603, in part by the Beijing Natural Science Foundation under Grant L247009, and in part by the Beijing National Research Center for Information Science and Technology.

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, 4, 5, 6, 7, 8

[2] Anh-Quan Cao and Raoul De Charette. Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields. In *ICCV*, pages 9387–9398, 2023. 1, 2

[3] OpenScene Contributors. Openscene: The largest up-to-date 3d occupancy prediction benchmark in autonomous driving. In *CVPR*, pages 18–22, 2023. 2, 4, 5, 6

[4] Xin Fei, Wenzhao Zheng, Yueqi Duan, Wei Zhan, Masayoshi Tomizuka, Kurt Keutzer, and Jiwen Lu. Driv3r: Learning dense 4d reconstruction for autonomous driving. *arXiv preprint arXiv:2412.06777*, 2024. 6

[5] Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. *arXiv preprint arXiv:2503.19755*, 2025. 1

[6] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *NeurIPS*, 37:91560–91596, 2024. 1

[7] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 32 (11):1231–1237, 2013. 2, 4, 5, 6

[8] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3828–3838, 2019. 1, 2, 7

[9] 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, 4, 5, 6

[10] Jiazhe Guo, Yikang Ding, Xiwu Chen, Shuo Chen, Bohan Li, Yingshuang Zou, Xiaoyang Lyu, Feiyang Tan, Xiaojuan Qi, Zhiheng Li, et al. Dist-4d: Disentangled spatiotemporal diffusion with metric depth for 4d driving scene generation. *arXiv preprint arXiv:2503.15208*, 2025. 2, 7

[11] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 1

[12] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *CVPR*, pages 17853–17862, 2023. 1

[13] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2

[14] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *CVPR*, pages 9223–9232, 2023. 1, 2

[15] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. Selfocc: Self-supervised vision-based 3d occupancy prediction. In *CVPR*, pages 19946–19956, 2024. 7

[16] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. In *ECCV*, pages 376–393. Springer, 2024. 1, 2

[17] Yuanhui Huang, Amonnut Thammatadatrakoon, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, and Jiwen Lu. Gaussianformer-2: Probabilistic gaussian superposition for efficient 3d occupancy prediction. In *CVPR*, pages 27477–27486, 2025. 2

[18] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *ICCV*, pages 8340–8350, 2023. 1

[19] 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. 2, 3, 6

[20] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, pages 71–91. Springer, 2024. 2

[21] Bohan Li, Zhuang Ma, Dalong Du, Baorui Peng, Zhujin Liang, Zhenqiang Liu, Chao Ma, Yueming Jin, Hao Zhao, Wenjun Zeng, et al. Omninwm: Omniscient driving navigation world models. *arXiv preprint arXiv:2510.18313*, 2025. 7

[22] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *ICCV*, pages 12578–12588, 2021. 1

[23] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *AAAI*, pages 1477–1485, 2023. 2

[24] Ye Li, Wenzhao Zheng, Xiaonan Huang, and Kurt Keutzer. Unidrive: Towards universal driving perception across camera configurations. *arXiv preprint arXiv:2410.13864*, 2024. 2

[25] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy

prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023. 2

[26] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *TPAMI*, 2024. 1, 2

[27] Kai-En Lin, Yen-Chen Lin, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, and Ravi Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *WACV*, pages 806–815, 2023. 1

[28] Aron Schmied, Tobias Fischer, Martin Danelljan, Marc Pollefeys, and Fisher Yu. R3d3: Dense 3d reconstruction of dynamic scenes from multiple cameras. In *ICCV*, pages 3216–3226, 2023. 7

[29] Yiang Shi, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Xinggang Wang. Occupancy as set of points. In *ECCV*, pages 72–87. Springer, 2024. 2

[30] 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. 3

[31] 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, 4, 5, 6

[32] Pin Tang, Zhongdao Wang, Guoqing Wang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma. Sparseocc: Rethinking sparse latent representation for vision-based semantic occupancy prediction. In *CVPR*, pages 15035–15044, 2024. 2

[33] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *NeurIPS*, 36:64318–64330, 2023. 1, 2

[34] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *TPAMI*, 13(4):376–380, 2002. 6

[35] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 2, 3

[36] Jiabao Wang, Zhaojiang Liu, Qiang Meng, Liujiang Yan, Ke Wang, Jie Yang, Wei Liu, Qibin Hou, and Ming-Ming Cheng. Opus: occupancy prediction using a sparse set. *NeurIPS*, 37:119861–119885, 2024. 2

[37] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, pages 5294–5306, 2025. 2, 3, 5, 6

[38] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, pages 10510–10522, 2025. 2, 6

[39] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *CVPR*, pages 5261–5271, 2025. 4

[40] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025. 2, 4

[41] Shuo Wang, Xinhai Zhao, Hai-Ming Xu, Zehui Chen, Dameng Yu, Jiahao Chang, Zhen Yang, and Feng Zhao. Towards domain generalization for multi-view 3d object detection in bird-eye-view. In *CVPR*, pages 13333–13342, 2023. 2

[42] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, pages 20697–20709, 2024. 2, 3

[43] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *ICCV*, pages 17850–17859, 2023. 2

[44] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *ECCV*, pages 55–72. Springer, 2024. 1

[45] 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. 2, 3

[46] Dongxu Wei, Zhiqi Li, and Peidong Liu. Omni-scene: Omni-gaussian representation for ego-centric sparse-view scene reconstruction. In *CVPR*, pages 22317–22327, 2025. 4

[47] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Yongming Rao, Guan Huang, Jiwen Lu, and Jie Zhou. Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation. In *CoRL*, pages 539–549. PMLR, 2023. 1, 2, 7

[48] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *ICCV*, pages 21729–21740, 2023. 1, 2

[49] Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. *arXiv preprint arXiv:2507.02863*, 2025. 2

[50] Shaoqing Xu, Fang Li, Shengyin Jiang, Ziying Song, Li Liu, and Zhi-xin Yang. Gaussianpretrain: A simple unified 3d gaussian representation for visual pre-training in autonomous driving. *arXiv preprint arXiv:2411.12452*, 2024. 2

[51] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, pages 10371–10381, 2024. 4

[52] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *ICCV*, pages 9043–9053, 2023. 4

[53] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, pages 4578–4587, 2021. 1

[54] Chubin Zhang, Juncheng Yan, Yi Wei, Jiaxin Li, Li Liu, Yansong Tang, Yueqi Duan, and Jiwen Lu. Occnerf: Self-supervised multi-camera occupancy prediction with neural radiance fields. *CoRR*, 2023. 2

[55] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *ICCV*, pages 9433–9443, 2023. 2

[56] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. In *ECCV*, pages 55–72. Springer, 2024. 1

[57] Wenzhao Zheng, Junjie Wu, Yao Zheng, Sicheng Zuo, Zixun Xie, Longchao Yang, Yong Pan, Zhihui Hao, Peng Jia, Xianpeng Lang, et al. Gaussianad: Gaussian-centric end-to-end autonomous driving. *arXiv preprint arXiv:2412.10371*, 2024.

[58] Wenzhao Zheng, Zetian Xia, Yuanhui Huang, Sicheng Zuo, Jie Zhou, and Jiwen Lu. Doe-1: Closed-loop autonomous driving with large world model. *arXiv preprint arXiv:2412.09627*, 2024.

[59] Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, and Alois C Knoll. Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. *arXiv preprint arXiv:2503.23463*, 2025.

[60] Zewei Zhou, Tianhui Cai, Seth Z Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. *arXiv preprint arXiv:2506.13757*, 2025. 1

[61] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025. 2, 3, 6

[62] Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Pointocc: Cylindrical tri-perspective view for point-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2308.16896*, 2023. 2

[63] Sicheng Zuo, Wenzhao Zheng, Xiaoyong Han, Longchao Yang, Yong Pan, and Jiwen Lu. Quadricformer: Scene as superquadrics for 3d semantic occupancy prediction. *arXiv preprint arXiv:2506.10977*, 2025. 2

[64] Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Gaussianworld: Gaussian world model for streaming 3d occupancy prediction. In *CVPR*, pages 6772–6781, 2025. 2