Visual Implicit Geometry Transformer for Autonomous Driving

Arsenii Shirokov , **Mikhail Kuznetsov** , **Danila Stepochkin** , **Egor Evdokimov** , **Daniil Glazkov** , **Nikolay Patakin** , **Anton Konushin** and **Dmitry Senushkin**
Lomonosov Moscow State University

Abstract

We introduce the Visual Implicit Geometry Transformer (ViGT), an autonomous driving geometric model that estimates continuous 3D occupancy fields from surround-view camera rigs. ViGT represents a step towards foundational geometric models for autonomous driving, prioritizing scalability, architectural simplicity, and generalization across diverse sensor configurations. Our approach achieves this through a calibration-free architecture, enabling a single model to adapt to different sensor setups. Unlike general-purpose geometric foundational models that focus on pixel-aligned predictions, ViGT estimates a continuous 3D occupancy field in a bird's-eye-view (BEV) addressing domain-specific requirements. ViGT naturally infers geometry from multiple camera views into a single metric coordinate frame, providing a common representation for multiple geometric tasks. Unlike most existing occupancy models, we adopt a self-supervised training procedure that leverages synchronized image-LiDAR pairs, eliminating the need for costly manual annotations. We validate the scalability and generalizability of our approach by training our model on a mixture of five large-scale autonomous driving datasets (NuScenes, Waymo, NuPlan, ONCE, and Argoverse) and achieving state-of-the-art performance on the pointmap estimation task, with the best average rank across all evaluated baselines. We further evaluate ViGT on the Occ3D-nuScenes benchmark, where ViGT achieves comparable performance with supervised methods.

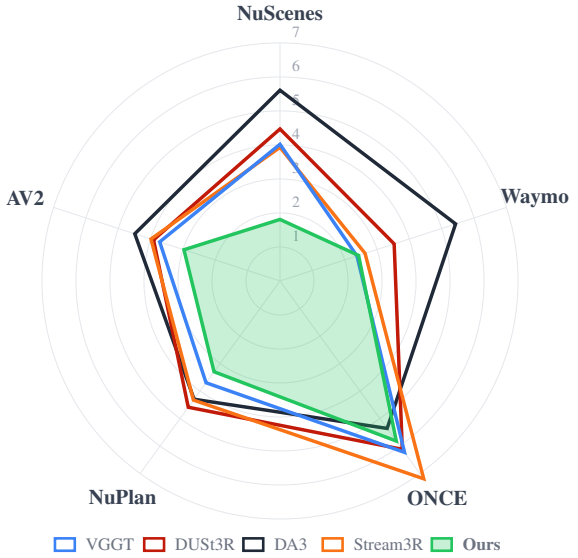

Figure 1. Our Visual Implicit Geometry Transformer outperforms the most recent fundamental geometric models on publicly available autonomous driving datasets in Chamfer Distance↓. Positions closer to the center indicate better performance.

1 Introduction

Accurate, metric geometric perception models are crucial for reliable and safe autonomous driving. Learning 3D-aware geometric representations from images and other sensor modalities has become central to progress in both autonomous driving and general scene understanding.

Recent advances in multi-view reconstruction [Wang *et al.*, 2024a; Lin *et al.*, 2025] have enabled a new class of end-to-end geometric models that directly estimate dense 3D point maps from uncalibrated image collections, implicitly recovering camera parameters. While these methods achieve strong performance on monocular and multi-view reconstruction in general domain, their application to autonomous driving scenarios remains challenging. Firstly, surround-view camera rigs commonly used in autonomous vehicles prioritize full 360° coverage with limited redundancy, resulting in reduced overlap between camera views and weakening the multi-view geometric assumptions underlying these approaches. Secondly, many recent multi-view reconstruction models do not explicitly enforce accurate metric scale, which is critical for autonomous driving applications. Thirdly, per-pixel depth and point predictions are typically aligned with individual images rather than the global scene, making it difficult to reconcile inconsistencies across views. In contrast, discretized 3D voxel representations that explicitly encode occupied and free space have become widely adopted in modern autonomous driving systems due to their direct compatibility with downstream tasks such as collision checking and motion planning.

However, voxel-based approaches often face scalability challenges, as they typically require dense and costly ground-truth annotations for training.

Occupancy fields provide an attractive alternative by modeling 3D scenes as continuous functions over space, rather than relying on explicit voxel grids or per-view depth maps. Such representations are compact, differentiable, resolution- and sensor-agnostic, and naturally align geometry from multiple views into a unified scene-centric coordinate frame, making them well suited for autonomous driving applications.

In this work, we introduce the Visual Implicit Geometry Transformer, which estimates a continuous 3D occupancy field directly from surround-view camera images. Our model learns a compact continuous bird's-eye-view (BEV) field that can be decoded into binary occupancy using a point-wise decoder. This field provides a flexible geometric representation of the scene and can be further transformed into downstream representations by querying point-wise occupancy probabilities and accumulating them through rendering. In contrast to previous approaches, we simplify the image-to-BEV transformation by making it calibration-free and fully data-driven. This design reduces inductive bias and improves the generality and scalability of the model. Finally, we adopt a scalable, sensor-agnostic, self-supervised training procedure that leverages synchronized but uncalibrated multi-view images and LiDAR data without requiring additional annotations.

2 Related Work

General geometric models. Foundational geometric models have been an active area of research for several decades. Learning-based monocular depth estimation was shown to generalize better when scaled with large and diverse datasets [Ranftl *et al.*, 2022; Patakin *et al.*, 2022]. In particular, pixel-aligned depth prediction models trained with uncalibrated [Ranftl *et al.*, 2022; Patakin *et al.*, 2022], or metric supervision [Bhat *et al.*, 2023] were shown to produce geometry estimates that generalize effectively to in-the-wild scenarios. However, monocular depth estimation remains fundamentally ill-posed due to depth ambiguity, which limits its generalization when relying on single-view inputs [Yin *et al.*, 2023]. To address this limitation, recent works proposed transformer-based architectures [Wang *et al.*, 2024a; Leroy *et al.*, 2024; Zhang *et al.*, 2025] that predict dense point maps instead of per-pixel depth. By leveraging multi-view supervision during training, these models reconstruct image-aligned visible geometry [Wang *et al.*, 2024a; Leroy *et al.*, 2024; Lin *et al.*, 2025] and, in some cases, estimate camera parameters from uncalibrated image collections [Wang *et al.*, 2025a]. Subsequent approaches [Wang *et al.*, 2025b; Wu *et al.*, 2025] introduced latent memory embedding that can be rendered from novel viewpoints, albeit with a limited number of views. While these methods significantly improve scalability with respect to both training data and the number of input images, they still rely heavily on multi-view correspondences and therefore exhibit substantial performance degradation under low-overlap image collections [Li *et al.*, 2025]. Moreover, most existing approaches regress scale-invariant depth or point representations, which are insufficient for recovering 3D metric volumetric geometry required by downstream applications such as autonomous driving. In this work, we introduce a scalable transformer-based model that estimates a continuous 3D occupancy field at metric scale directly from surround-view images. Differently, our method is designed for surround-view driving scenarios with limited camera overlap, alleviating multi-view constraints.

Geometric models for autonomous driving. Prior geometric models for autonomous driving have predominantly focused on semantic occupancy estimation, as it provides a structured representation suitable for downstream tasks. Vision-based perception models in this domain build upon earlier works [Li *et al.*, 2022; Philion and Fidler, 2020] that introduced forward [Li *et al.*, 2022] and backward [Philion and Fidler, 2020] mappings from images to bird's-eye-view (BEV) space, where subsequent estimations are performed. More recent methods [Marinello *et al.*, 2025; Pan *et al.*, 2024; Huang *et al.*, 2024; Tang *et al.*, 2024] typically lift multi-view image features into voxel grids using explicit geometric projections [Li *et al.*, 2022; Philion and Fidler, 2020], followed by refinement through 3D convolutional or transformer-based fusion [Zhou and Krähenbühl, 2022]. Alternative approaches [Liu *et al.*, 2023b; Zhang *et al.*, 2024] incorporate explicit LiDAR encoders to improve prediction accuracy. Despite their effectiveness, these architectures rely on precise camera calibration, which limits generalization to novel camera configurations or scenarios with calibration errors. Additionally, they require expensive annotations, such as voxel-level semantic labels, which constrains scalability. In contrast, we introduce the simple transformer-based model architecture that does not use camera calibration for BEV construction. Our model learns a continuous 3D occupancy field in BEV representation that incorporates depth, point clouds, and occupancy from images, enabling scalable and transferable 3D perception across diverse self-driving datasets.

Recent advances [Agro *et al.*, 2024; Boulch *et al.*, 2023; Yang *et al.*, 2024] have introduced continuous occupancy forecasting in both space and time, leveraging pretext geometric tasks to support downstream applications. Early methods [Boulch *et al.*, 2023; Agro *et al.*, 2024; Khurana *et al.*, 2023] operate on LiDAR point clouds and employ convolutional architectures to predict binary occupancy at the point level. Subsequent works [Agro *et al.*, 2024; Liu *et al.*, 2024] extend this paradigm by incorporating temporal modeling. Similarly, visual point cloud forecasting methods, such as ViDAR [Yang *et al.*, 2024], predict future LiDAR-like point clouds from monocular image sequences rather than estimating occupancy directly. However, the majority of these models remain biased by calibration-dependent architectures, limiting their ability to generalize across diverse sensor configurations and environments. We leverage continuous occupancy estimation task for training our models since it provides annotation-free supervision from raw data.

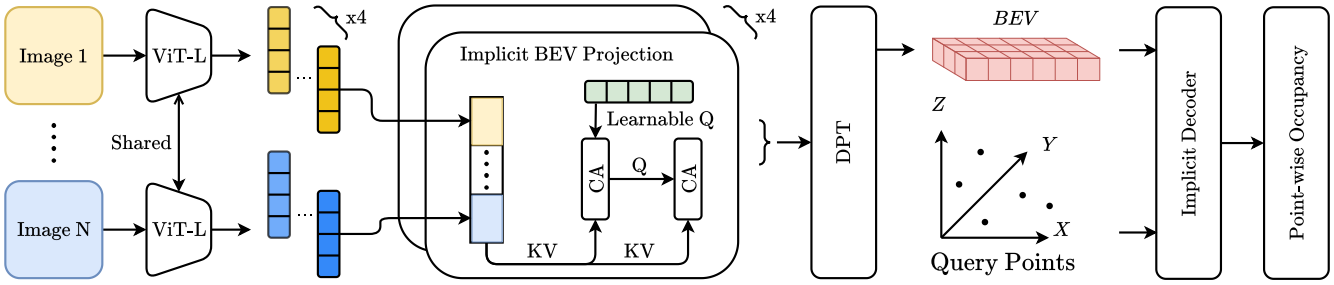

Figure 2. Our architecture consists of three main components: (1) an image encoder (ViT-L) that independently processes each image and extracts feature tokens from the last four layers, producing four sequences of tokens per image; (2) a calibration-free Implicit BEV Projection module that projects tokens from each encoder layer across all images to their corresponding BEV space, generating four layer-specific BEV representations, which are then aggregated and upsampled into a single unified BEV representation using DPT; and (3) a query-based Implicit Decoder that predicts occupancy probabilities for 3D points from the final BEV features. This design enables pure data-driven scene modeling without geometric inductive biases.

3 Method

In this work, we introduce the Visual Implicit Geometry Transformer (ViGT) that estimates continuous 3D occupancy field directly from surround-view camera images. The remainder of this section is organized as follows: section 3.1 presents ViGT architecture that incorporates scalable blocks without inductive biases; section 3.2 describes unsupervised occupancy estimation problem as the scalable task for training; finally, section 3.3 details the label-free training procedure enabling learning continuous 3D occupancy fields end-to-end.

3.1 Visual Implicit Geometry Transformer

We step towards general foundational geometric models for autonomous driving by designing our model architecture, prioritizing scalability, simplicity, and generalization across diverse sensor configurations. Our architecture consists of three main components (Fig. 2): an image encoder that independently processes each image, a calibration-free implicit camera-to-BEV projection module that unifies independent image features in a scene-centric BEV representation and a query-based implicit decoder that predicts occupancy values for 3D points.

Image Encoder. Following principles of simplicity and scalability, we employ a ViT-Large backbone [Dosovitskiy *et al.*, 2021] to extract visual features from camera images. In contrast to prior autonomous driving models that rely on convolutional neural networks with limited receptive fields [Li *et al.*, 2022; Philion and Fidler, 2020], transformer architectures capture long-range contextual dependencies through self-attention while imposing minimal inductive bias. Each image is encoded independently by the ViT-Large model, producing a compact set of token embeddings that serve as input to subsequent processing stages.

Implicit BEV projection. Autonomous driving applications require scene-centric geometric representations defined at metric scale within a unified coordinate frame. Unlike image-aligned representations, such as depth or point maps, scene-centric representations enable direct reasoning about spatial occupancy, which is critical for autonomous driving tasks. To this end, we adopt a bird's-eye-view (BEV) representation, as it provides comprehensive scene coverage while being more computationally and memory efficient than dense 3D volumetric grids. Existing self-driving approaches [Li *et al.*, 2022; Philion and Fidler, 2020] rely on known camera intrinsics and extrinsics to explicitly project image features into BEV space. Guided by principles of simplicity and scalability, we seek a calibration-free transformation from multi-view visual features to BEV space that can be applied across diverse camera setups.

We design our implicit camera-to-BEV projection module to learn the geometric transformation from data (Figure 3, Figure 5). Specifically, the implicit camera-to-BEV projection employs two sequential cross-attention blocks between image patches and BEV latent queries. The number of BEV latent queries is smaller than the number of image patch tokens, addressing computationally efficiency. To capture multiscale geometric information, we project outputs of the four last layers of ViT-L encoder independently to BEV space, using our implicit BEV projection with non-shared weights and individual BEV queries. The resulting multi-layer BEV representations are then aggregated and upscaled using DPT [Ranftl *et al.*, 2021] to obtain the final fine-grained scene-level BEV representation. We validate our design choices through ablation studies comparing different projection structures and encoder layer selections (Table 3).

Implicit decoder. To transform the discrete BEV grid into a continuous occupancy field, we employ an implicit decoder inspired by ImplicitIO [Agro *et al.*, 2023] that maps 3D query points to occupancy probabilities. We linearly interpolate query features and concatenate them with normalized query points. The concatenated features are decoded using Convolutional Occupancy Network [Peng *et al.*, 2020] to predict the occupancy probabilities for the query point.

3.2 Unsupervised Occupancy Estimation

To build a scalable training framework, we leverage a training objective that relies exclusively on data that can be collected efficiently on real-world autonomous platforms. Specifically, we assume the availability of synchronized multi-view images and corresponding LiDAR point clouds. Unlike con-

Figure 3. Cross-attention visualization demonstrating consistent correspondences between BEV queries and camera tokens. (Left) Image attention heatmaps for selected BEV queries, (Right) BEV attention heatmaps for selected image queries. These visualizations confirm that the implicit BEV projection learns geometrically correct camera-to-BEV transformations.

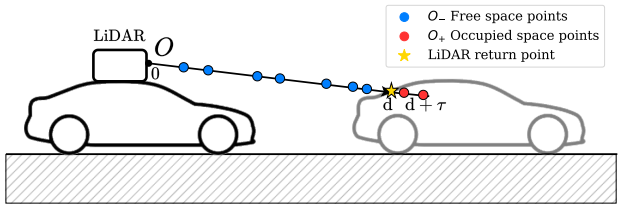

Figure 4. We construct two sets of training points sampled along each LiDAR ray. Points before the reflection point are labeled as free space (negative label), while points near the reflection point correspond to occupied space (positive label).

ventional occupancy estimation approaches that operate on discretized voxel grids, we avoid explicit spatial discretization and instead formulate the task in continuous space, operating directly on points. LiDAR is a high-precision active sensor that emits laser pulses and measures the return time from reflecting surfaces to estimate distances. By construction, LiDAR measurements guarantee that the space between the sensor origin and the first surface intersection along a ray is free. Accordingly, we construct two sets of training points sampled along each LiDAR ray. Points located before the reflection point are labeled as free space and thus assigned negative occupancy labels, while points near reflection point correspond to solid surfaces and are assigned positive occupancy labels. More formally, given a normalized LiDAR ray parameterized by origin $\boldsymbol{o}$ and direction $\boldsymbol{d}$, we define two occupancy classes as follows:

$$\begin{cases} O_- : \boldsymbol{p}_i = \boldsymbol{o} + t_i \boldsymbol{d}, & t_i \in [0, d], \\ O_+ : \boldsymbol{p}_j = \boldsymbol{o} + t_j \boldsymbol{d}, & t_j \in [d, d + \tau], \end{cases} \tag{1}$$

where d is the distance from the ray origin $\boldsymbol{o}$ to the reflected surface and τ is the thickness hyperparameter that regulates the minimal surface depth along rays. This value is usually small for models that should perceive thin surfaces.

Overall, the training objective is formulated as a binary occupancy classification problem, where the model predicts the occupancy label of each queried 3D point given the available input context. Prior works [Boulch *et al.*, 2023; Agro *et al.*, 2024] primarily rely on single or multiple LiDAR point clouds as input, resulting in LiDAR-only models. In contrast, we adopt a vision-only paradigm and leverage this training objective to supervise the Visual Implicit Geometry Transformer.

3.3 Training

Following the unsupervised occupancy estimation formulation in Section 3.2, we sample query points along LiDAR rays to construct positives (occupied space) and negatives (free space) occupancy labels. We employ a balanced sampling strategy, that ensures equal representation of negative and positive targets. We divide the ray segment $[0, d)$ into K bins and sample uniformly within each bin, ensuring coverage across the entire free space. However, simple uniform sampling for negative queries can lead to inefficient supervision near object surfaces, where accurate boundary predictions are crucial. To this end, we augment this with symmetric negative sampling from the region $[d - \tau, d)$ where τ is the same hyperparameter used for positive sampling, better utilizing object borders and improving surface localization. This symmetric approach samples negative points near the surface boundary, providing explicit supervision for the model to learn object borders. We validate our query sampling strategy through ablation studies comparing random, stratified, and stratified with symmetric interval sampling (Table 3).

We train the model using binary cross-entropy loss. For a batch of query points $\{\boldsymbol{p}_i\}_{i=1}^N$ with ground truth occupancy labels $\{y_i \in \{0, 1\}\}_{i=1}^N$ and predicted occupancy probabilities $\{o_i \in [0, 1]\}_{i=1}^N$, the loss is defined as:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(o_i) + (1 - y_i) \log(1 - o_i) \right]. \tag{2}$$

3.4 Rendering

Our model estimates a continuous occupancy field, which provides a flexible geometric representation that can be transformed into various downstream representations by querying point-wise occupancy probabilities and accumulating them through volume rendering. This unified representation enables conversion to point clouds and occupancy voxel grids, facilitating application to diverse autonomous driving tasks while maintaining a single, consistent geometric model.

Ray-based rendering. To extract point clouds, we leverage volumetric integration to render LiDAR rays. Given the ray with origin $\boldsymbol{o}$ and direction $\boldsymbol{d}$, we sample points

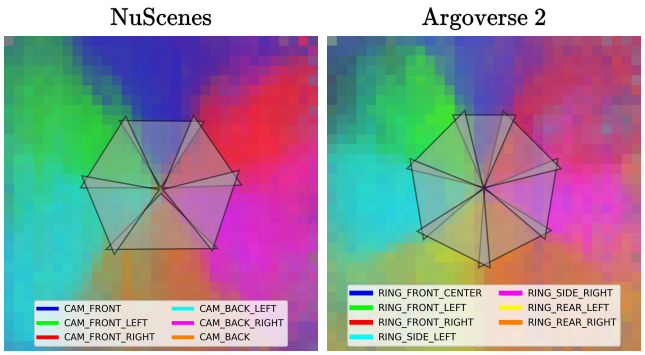

Figure 5. The camera-to-bev attention responses (colored regions) accurately partition BEV space according to each camera's field of view, demonstrating that the implicit BEV projection correctly learns to map image features to their corresponding spatial sectors in BEV space without explicit camera parameters. Shown for different camera rigs configuration.

$\{\boldsymbol{p}_i = \boldsymbol{o} + t_i\boldsymbol{d}\}_{i=1}^{N}$, where t_i are evenly spaced sampling distances. We decode occupancy probability $o(\boldsymbol{p}_i)$ for every point and integrate the depth value accordingly:

$$d = \sum_{i=1}^{N} t_i \cdot o(\boldsymbol{p}_i) \cdot T_i, \quad T_i = \prod_{j=1}^{i-1}(1 - o(\boldsymbol{p}_j)), \tag{3}$$

where T_i is the transmittance accumulated from predicted occupancy.

Voxel grid rendering. To generate binary voxel grids from continuous occupancy field, we query M points randomly sampled within each voxel and aggregate their occupancy predictions. For a voxel V the occupancy is computed as the maximum response across all predictions:

$$O(V) = \max_{j=1}^{M} o(\boldsymbol{p}_j), \tag{4}$$

4 Experiments

4.1 Training details

We train the model with the AdamW optimizer for 200K iterations, using a cosine learning rate scheduler with a peak learning rate of 5×10^{-5} and a warmup of 10K iterations. To ensure diverse training data, we use balanced sampling across five autonomous driving datasets: NuScenes [Caesar *et al.*, 2020], ONCE [Mao *et al.*, 2021], Waymo [Sun *et al.*, 2020], Argoverse 2 [Wilson *et al.*, 2021], and NuPlan [Caesar *et al.*, 2021]. For each batch, we sample equally from all datasets to prevent bias towards larger datasets. We use a batch size of 6 per GPU. The input images are resized with the shortest side set to 192 pixels. We use 1024 BEV queries with 1024 channels in each Implicit BEV projector, aggregating features to 256×256 resolution with 256 channels after DPT. The training runs on 128 A100 GPUs over five days. We employ gradient norm clipping with a threshold of 1.0 to ensure training stability.

Method	Conf.	Supervision	Calib.	F1-score↑	IoU↑
FB-Occ	ICCV 2023	3D GT	+	0.8181	0.7022
Sparse-Occ	ECCV 2024	3D GT	+	0.6271	0.4680
PanoOcc	CVPR 2024	3D GT	+	0.8347	0.7271
Offset-Occ	CVPR 2025	3D GT	+	0.6240	0.4637
RenderOcc	ICRA 2024	2D GT	+	0.6442	0.4824
Self-Occ	CVPR 2024	Self-sup	+	0.6552	0.4960
Ours		Self-sup	-	0.7115	0.5658

Table 1. Occ3D-NuScenes. Our method achieves third-best overall performance and first-best performance over previous self-supervised methods and methods with 2d labels supervision. Our method is the only approach which don't use camera calibration.

4.2 Implicit BEV projection

We demonstrate that our calibration-free implicit BEV projection correctly learns geometric transformations without explicit camera parameters. Figure 3 illustrates attention responses between BEV queries and camera tokens, showing consistent correspondences in both forward (camera-to-BEV) and inverse (BEV-to-camera) attention directions. Figure 5 demonstrates that the projection correctly aligns image features to their corresponding spatial sectors in BEV space for different camera rigs (6 cameras on NuScenes and 7 cameras on Argoverse 2) from different datasets, with each camera's field of view accurately mapped to its correct frustum region. These figures empirically confirm that the implicit projection successfully learns geometrically correct camera-to-BEV transformations in a calibration-free manner across diverse datasets with different camera rig configurations.

4.3 Occupancy estimation

The first experiment compares our model with recent geometric models designed for autonomous driving applications.

Evaluation protocol. We evaluate all models on the Occ3D-NuScenes benchmark [Tian *et al.*, 2023] which comprises 150 validation scenes with voxelized occupancy annotations defined within a region of interest of $[-40\text{m}, -40\text{m}, -1\text{m}] \times [+40\text{m}, +40\text{m}, +5.4\text{m}]$ in ego-vehicle coordinate frame with voxel resolution $[0.4\text{m}, 0.4\text{m}, 0.4\text{m}]$. Unlike the original benchmark, which provides semantic occupancy annotations across 17 categories, we focus exclusively on geometric occupancy estimation. To this end, all semantic classes are mapped to a single occupied label, reducing the task to binary occupancy prediction. We evaluate occupancy estimation using two metrics: F1-score and Intersection over Union (IoU). Both metrics are computed over the binary occupancy predictions (occupied vs. free) within the camera's visible region.

Baselines. We compare our method against recent vision-based occupancy prediction approaches as shown in Table 1. All models are grouped by their supervision type. Most approaches employ dense voxel grids with explicit 3D occupancy annotations: FB-Occ [Li *et al.*, 2023], PanoOcc [Wang *et al.*, 2024b], Sparse-Occ [Liu *et al.*, 2023a], and Offset-Occ [Marinello *et al.*, 2025]. Methods with reduced supervision include RenderOcc [Pan *et al.*, 2024], which trains with 2D supervision, and Self-Occ [Huang *et al.*, 2024], which

Method	Conf.	Rendering	NuScenes		AV2		Waymo		ONCE		NuPlan		Avg. Rank↓
			AbsRel↓	CD↓	AbsRel↓	CD↓	AbsRel↓	CD↓	AbsRel↓	CD↓	AbsRel↓	CD↓	
VGGT	CVPR 2025	D	0.195	4.019	0.164	3.724	0.084	2.381	0.450	6.230	0.167	3.696	3.9
		P	0.303	4.794	0.220	4.305	0.113	3.576	0.568	7.566	0.215	5.047	
DUSt3R	CVPR 2024	D	0.250	4.466	0.230	3.906	0.152	3.534	0.515	6.124	0.236	4.590	6.9
		P	0.375	5.321	0.445	6.309	0.347	7.726	0.887	10.341	0.508	6.749	
Mast3R	ECCV 2024	D	0.191	4.583	0.190	4.752	0.120	2.949	0.498	8.402	0.181	4.826	6.4
		P	0.285	5.487	0.355	6.876	0.260	4.673	0.843	10.906	0.635	8.475	
Monst3R	ICLR 2025	D	0.217	3.945	0.201	4.013	0.127	3.037	0.390	5.039	0.182	3.443	4.8
		P	0.415	7.024	0.279	5.726	0.210	5.442	1.308	15.659	0.263	7.221	
Stream3R	—	D	0.173	3.931	0.162	3.992	0.083	2.629	0.465	7.190	0.163	4.331	4
		P	0.397	5.985	0.237	5.799	0.114	3.522	0.705	10.859	0.207	5.668	
Cut3R	CVPR 2025	D	0.189	4.062	0.194	3.681	0.101	2.478	0.447	5.970	0.156	3.000	3.5
		P	0.240	5.233	0.342	6.760	0.176	3.866	0.811	11.442	0.172	3.688	
DA3	—	D	0.289	5.606	0.174	4.488	0.383	5.426	0.403	5.360	0.265	4.300	6.4
RenderOcc	ICRA 2024	PR	0.245	3.637	0.316	10.864	0.526	8.442	0.298	7.490	0.332	5.295	7.3
Ours		PR	0.068	1.807	0.131	2.965	0.121	2.431	0.169	5.821	0.118	3.298	1.8

Table 2. Pointmap Estimation. Our method achieves the best Average Rank (1.8) across all metrics and datasets, demonstrating the best overall performance. Rendering types: D denotes depth map unprojection from camera views, P denotes direct point cloud from pointmaps, and PR denotes points rendering from predicted occupancy fields.

learns occupancy in a self-supervised manner without voxel labels, similar to our approach.

Experimental results. We evaluate our model without voxel finetuning to demonstrate that our learned representation can generalize to new geometric tasks. Our model achieves third-best overall performance with an F1-score of 0.7115 and IoU of 0.5658, while being the only approach that is both fully self-supervised and calibration-free. Unlike top-performing methods such as PanoOcc and FB-Occ that require expensive 3D annotations, our method uses only raw LiDAR point clouds for supervision, eliminating the need for costly manual labeling. Compared to the previous best self-supervised method (Self-Occ), we achieve improvements of 0.056 in F1-score and 0.07 in IoU, while additionally removing the calibration requirement.

4.4 Pointmap Estimation

Second, we evaluate our model against general-domain vision-based methods that are similarly calibration-free but produce pixel-aligned predictions.

Evaluation protocol. We evaluate all methods on the public test splits of widely used autonomous driving datasets, including NuScenes, Argoverse 2, Waymo, ONCE, and NuPlan, using LiDAR point clouds as ground-truth targets. Due to the large scale of nuPlan, we subsample the official test split by selecting every 100th LiDAR frame for evaluation. We report two evaluation metrics: Absolute Relative Error (AbsRel↓), computed along LiDAR rays, and Chamfer Distance (CD↓). For pixel-aligned methods, predictions are sampled at image pixels corresponding to the projections of LiDAR points onto the image plane. To provide an aggregate measure of overall performance across all datasets and metrics, we compute the Average Rank (Avg Rank↓) for each method as the mean of performance places (1st, 2nd, 3rd, 4th, etc.) assigned to each metric-dataset combination, where lower rank values indicate better overall performance.

Baselines Table 2 presents a comprehensive comparison of our method against state-of-the-art approaches. The evaluated methods can be categorized into two groups: (1) general geometry methods that produce image-aligned representations, and (2) self-driving domain methods that estimate 3D scene representation. For general geometry methods (VGGT [Wang *et al.*, 2025a], DUSt3R [Wang *et al.*, 2024a], Mast3R [Leroy *et al.*, 2024], Monst3R [Zhang *et al.*, 2025], Stream3R [Lan *et al.*, 2025], Cut3R [Wang *et al.*, 2025b], and DA3 [Lin *et al.*, 2025]), we evaluate them using two variants: depth maps and pointmaps, both filtered by LiDAR ray masks to extract valid 3D points for comparison. Note that we use metric-scale DA3 that only produces depth maps, so we evaluate it using the depth map variant only. For scene-centric methods including ours and RenderOcc [Pan *et al.*, 2024], we render depth by LiDAR rays using Eq. (3).

Experimental results. As shown in Table 2, our method achieves the best average rank (1.8) across all metrics and datasets. Figure 6 provides a visual comparison of the top-3 methods by Average Rank. We outperform all competing methods in both AbsRel and CD metrics on NuScenes and Argoverse 2. On NuScenes, we achieve AbsRel of 0.068 and CD of 1.807, representing improvements of 0.105 in AbsRel and 1.83 in CD over the second-best method (Stream3R with depth-maps). On Argoverse 2, we achieve AbsRel of 0.131, outperforming the best competitor by 0.031 (Stream3R with depth-maps); and we achieve CD of 2.965, outperforming the best competitor by 0.716 (Cut3R with depth-maps). On Waymo, we achieve the second-best CD performance (2.431), trailing the best method by only 0.05. On ONCE and NuPlan, we achieve the best AbsRel performance, with third-best and second-best CD performance respectively.

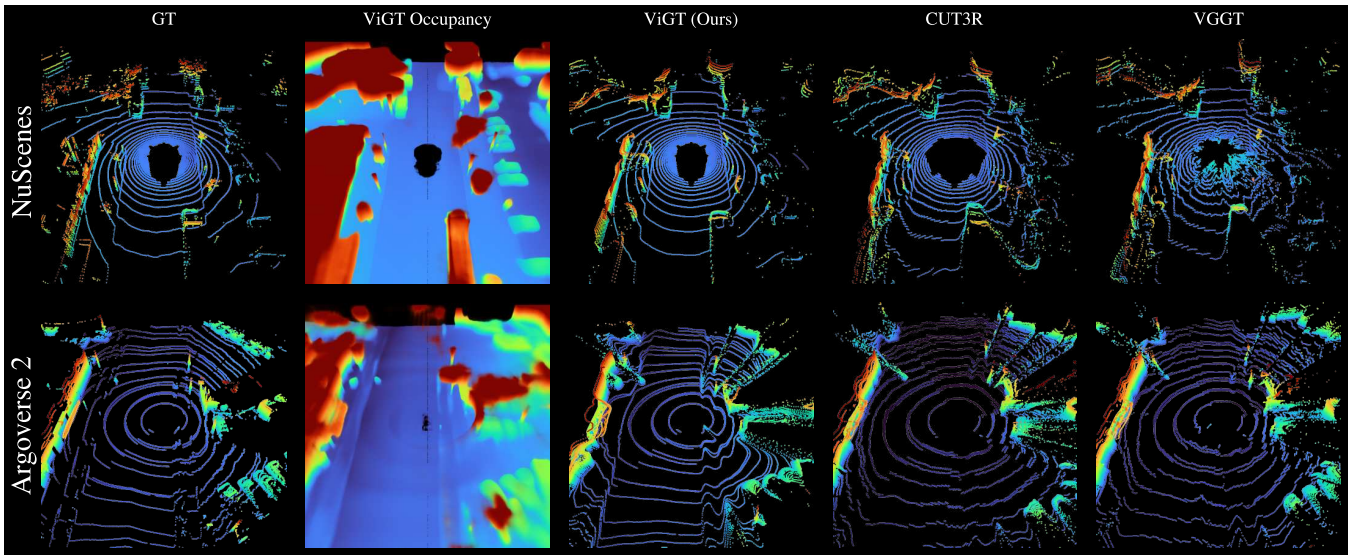

Figure 6. Visual comparison of GT and predicted point clouds across the top-3 methods ranked by Average Rank (Ours, Cut3R, VGGT). Our method produces more geometrically accurate point clouds. We also visualize a render of the predicted occupancy, demonstrating accurate capture of continuous geometric structures.. The color encodes height along the z-axis.

Component	Point Est.	Voxel Occ.	
Implicit BEV Projector Architecture	CD↓	F1-score↑	IoU↑
CA	3.051	0.697	0.547
CA + CA	**2.699**	**0.713**	**0.566**
CA + SA + CA	2.814	0.697	0.548
Encoder Output Layers	CD↓	F1-score↑	IoU↑
Last 4 layers	**2.699**	**0.713**	**0.566**
Layers 5, 11, 16, 23	3.093	0.70	0.55
Query Sampling Strategy	CD↓	F1-score↑	IoU↑
Random	2.895	0.701	0.552
Stratified	3.039	0.697	0.547
Stratified + Sym. Interval	**2.699**	**0.713**	**0.566**

Table 3. Ablation study of key design choices: BEV projector architecture, encoder layer selection, and query sampling strategy.

Compared to RenderOcc, the most similar autonomous driving model, we consistently outperform it across all evaluated metrics. These results demonstrate the effectiveness of our calibration-free, self-supervised approach in accurately estimating 3D geometry from camera rig images across diverse autonomous driving datasets.

4.5 Ablation Studies

We evaluate three main components: the implicit BEV projector architecture, encoder output layer selection, and query sampling strategy (see Table 3.). All ablation experiments are trained on NuScenes only, using 4 GPUs for 100k iterations.

BEV Projector Architecture We compare three variants of implicit BEV projection module: a single cross-attention block (CA), two sequential cross-attention blocks (CA + CA), and a variant with self-attention (CA + SA + CA). The CA + CA configuration achieves the best performance, demonstrating that two sequential cross-attention operations provide sufficient capacity to learn the geometric transformation without the added complexity of self-attention.

Encoder Layer Selection We evaluate which ViT-L encoder layers to project to BEV space for DPT aggregation and upsampling. We compare using the last four layers versus default ViT-L output layers (5, 11, 16, 23). The last four layers selection shows better performance.

Query Sampling Strategy We ablate the strategy for sampling query points along LiDAR rays for occupancy supervision. We compare random sampling, stratified bin-based sampling, and stratified sampling augmented with symmetric interval sampling near object boundaries. The stratified + symmetric interval strategy performs best, confirming that explicit supervision near object boundaries is crucial for accurate occupancy estimation.

5 Conclusion

In this paper, we propose ViGT, a scalable vision-based geometric model for autonomous driving that learns a continuous 3D occupancy field directly from uncalibrated multi-camera images. ViGT is trained jointly on five large-scale datasets using a self-supervised learning procedure based on synchronized image–LiDAR pairs. We show that the resulting unified representation can be directly rendered for multiple downstream tasks, including pointmap estimation and occupancy prediction, without task-specific retraining. Extensive evaluations demonstrate that ViGT generalizes effectively, achieving state-of-the-art performance on pointmap estimation across five datasets with diverse camera rig configurations while remaining competitive with supervised methods on voxel-based occupancy benchmark. We consider such models representing a promising research direction toward building foundational geometric perception systems for autonomous driving applications.

Acknowledgements

This work was supported by the The Ministry of Economic Development of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000C313925P4H0002; grant No 139-15-2025-012).

References

[Agro *et al.*, 2023] Ben Agro, Quinlan Sykora, Sergio Casas, and Raquel Urtasun. Implicit occupancy flow fields for perception and prediction in self-driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1379–1388, 2023.

[Agro *et al.*, 2024] Ben Agro, Quinlan Sykora, Sergio Casas, Thomas Gilles, and Raquel Urtasun. Uno: Unsupervised occupancy fields for perception and forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14487–14496, 2024.

[Bhat *et al.*, 2023] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *CoRR*, abs/2302.12288, 2023.

[Boulch *et al.*, 2023] Alexandre Boulch, Corentin Sautier, Björn Michele, Gilles Puy, and Renaud Marlet. ALSO: Automotive lidar self-supervision by occupancy estimation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[Caesar *et al.*, 2020] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[Caesar *et al.*, 2021] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[Huang *et al.*, 2024] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. Selfocc: Self-supervised vision-based 3d occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19946–19956, 2024.

[Khurana *et al.*, 2023] Tarasha Khurana, Peiyun Hu, David Held, and Deva Ramanan. Point cloud forecasting as a proxy for 4d occupancy forecasting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[Lan *et al.*, 2025] Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. Stream3r: Scalable sequential 3d reconstruction with causal transformer. *arXiv preprint arXiv:2508.10893*, 2025.

[Leroy *et al.*, 2024] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024.

[Li *et al.*, 2022] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, page 1–18, Berlin, Heidelberg, 2022. Springer-Verlag.

[Li *et al.*, 2023] 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.

[Li *et al.*, 2025] Samuel Li, Pujith Kachana, Prajwal Chidananda, Saurabh Nair, Yasutaka Furukawa, and Matthew Brown. Rig3r: Rig-aware conditioning for learned 3d reconstruction. *arXiv preprint arXiv:2506.02265*, 2025.

[Lin *et al.*, 2025] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[Liu *et al.*, 2023a] Haisong Liu, Haiguang Wang, Yang Chen, Zetong Yang, Jia Zeng, Li Chen, and Limin Wang. Fully sparse 3d panoptic occupancy prediction. *CoRR*, 2023.

[Liu *et al.*, 2023b] Zhijian Liu, Haotian Tang, Alexander Amini, Xingyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[Liu *et al.*, 2024] Xinhao Liu, Moonjun Gong, Qi Fang, Haoyu Xie, Yiming Li, Hang Zhao, and Chen Feng. Lidar-based 4d occupancy completion and forecasting. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11102–11109. IEEE, 2024.

[Mao *et al.*, 2021] Jiageng Mao, Niu Minzhe, ChenHan Jiang, hanxue liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, Chunjing XU, and Hang Xu. One million scenes for autonomous driving: Once dataset. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

[Marinello *et al.*, 2025] Nicola Marinello, Simen Cassiman, Jonas Heylen, Marc Proesmans, and Luc Van Gool. Camera-only 3d panoptic scene completion for autonomous driving through differentiable object shapes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2520–2529, 2025.

[Pan *et al.*, 2024] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Hongwei Xie, Bing Wang, Li Liu, and Shanghang Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12404–12411. IEEE, 2024.

[Patakin *et al.*, 2022] Nikolay Patakin, Anna Vorontsova, Mikhail Artemyev, and Anton Konushin. Single-stage 3d geometry-preserving depth estimation model training on dataset mixtures with uncalibrated stereo data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1705–1714, June 2022.

[Peng *et al.*, 2020] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision*, pages 523–540. Springer, 2020.

[Philion and Fidler, 2020] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, pages 194–210. Springer, 2020.

[Ranftl *et al.*, 2021] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.

[Ranftl *et al.*, 2022] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022.

[Sun *et al.*, 2020] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

[Tang *et al.*, 2024] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15035–15044, 2024.

[Tian *et al.*, 2023] 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. *Advances in Neural Information Processing Systems*, 36:64318–64330, 2023.

[Wang *et al.*, 2024a] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024.

[Wang *et al.*, 2024b] Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17158–17168, 2024.

[Wang *et al.*, 2025a] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[Wang *et al.*, 2025b] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025.

[Wilson *et al.*, 2021] 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 *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021.

[Wu *et al.*, 2025] Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[Yang *et al.*, 2024] Zetong Yang, Li Chen, Yanan Sun, and Hongyang Li. Visual point cloud forecasting enables scalable autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14673–14684, 2024.

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

[Zhang *et al.*, 2024] Shuo Zhang, Yupeng Zhai, Jilin Mei, and Yu Hu. Fusionocc: Multi-modal fusion for 3d occupancy prediction. In *ACM Multimedia 2024*, 2024.

[Zhang *et al.*, 2025] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3r: A simple approach for estimating geometry in the presence of motion. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Zhou and Krähenbühl, 2022] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *CVPR*, 2022.