Gamma-from-Mono: Road-Relative, Metric, Self-Supervised Monocular Geometry for Vehicular Applications

Gasser Elazab[1,2] Maximilian Jansen[1] Michael Unterreiner[1] Olaf Hellwich[2]
[1]CARIAD SE [2]Technische Universität Berlin
gasser.elazab@cariad.technology

Abstract

Accurate perception of the vehicle's 3D surroundings, including fine-scale road geometry, such as bumps, slopes, and surface irregularities, is essential for safe and comfortable vehicle control. However, conventional monocular depth estimation often oversmooths these features, losing critical information for motion planning and stability. To address this, we introduce Gamma-from-Mono (GfM), a lightweight monocular geometry estimation method that resolves the projective ambiguity in single-camera reconstruction by decoupling global and local structure. GfM predicts a dominant road surface plane together with residual variations expressed by γ, a dimensionless measure of vertical deviation from the plane, defined as the ratio of a point's height above it to its depth from the camera, and grounded in established planar parallax geometry. With only the camera's height above ground, this representation deterministically recovers metric depth via a closed form, avoiding full extrinsic calibration and naturally prioritizing near-road detail. Its physically interpretable formulation makes it well suited for self-supervised learning, eliminating the need for large annotated datasets. Evaluated on KITTI and the Road Surface Reconstruction Dataset (RSRD), GfM achieves state-of-the-art near-field accuracy in both depth and γ estimation while maintaining competitive global depth performance. Our lightweight 8.88M-parameter model adapts robustly across diverse camera setups and, to our knowledge, is the first self-supervised monocular approach evaluated on RSRD.

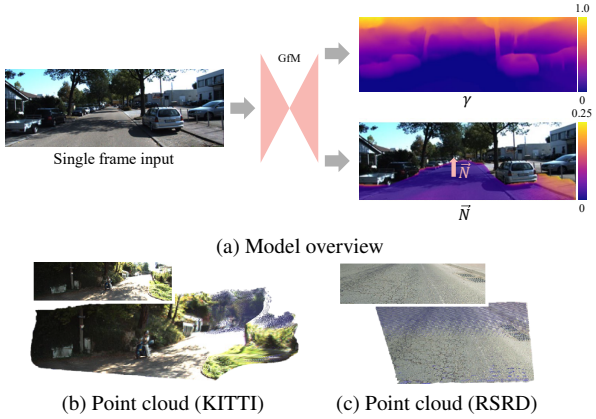

Figure 1. Overview of our self-supervised model. (a) Predicts γ (height/depth ratio) and road normal $\vec{N}_{\text{pred}}$. Height visualizations are clipped from 0 to 0.25 m. (b, c) 3D point clouds from KITTI [8] and RSRD [46], with input images in the top-left.

1. Introduction

Modern robotics and autonomous vehicles demand scalable 3D perception, but achieving it with a single camera remains challenging [23]. Monocular Geometry Estimation (MGE) reconstructs per-pixel 3D structure from a single image, enabling cost-effective 3D perception without specialized sensors [2]. Specifically, accurate 3D reconstruction of the near-road geometry is crucial for navigation in autonomous driving [16, 17], supporting obstacle avoidance and motion planning [32], and it is also vital for off-road and legged robotics, where elevation and local slope guide traversability and footstep planning [21]. However, monocular depth estimation methods, which are commonly used in these domains, struggle to accurately capture road topography [47]. Textureless pavements and low-contrast surfaces often lead to oversmoothing and underestimation of slopes, causing small obstacles and surface irregularities to be missed [1, 39, 47]. This limitation is critical, as small height variations, such as bumps or road-level changes, can differentiate drivable regions from hazards and negatively impact vehicle dynamics and safety [17, 22, 27].

Metric foundation depth models [4, 18, 19, 41, 42] generalize well across domains and provide strong scene-level predictions. However, they do not explicitly target road-relative quantities such as height above the ground or local slope. Moreover, they often suffer from residual scale drift when deployed in new envi-

ronments, typically requiring inference-time scale correction. In contrast, self-supervised monocular methods [5, 7, 11, 14, 28, 30, 37, 45] attempt to resolve scale ambiguity by incorporating metric anchors such as odometry sensors or ground-plane constraints. While these strategies stabilize global scale, they remain depth-centric, leaving road geometry underconstrained.

In practice, existing monocular pipelines recover height above the road indirectly via costly post-processing, such as elevation maps [21]. Recent top-down (BEV) approaches [47] model road surfaces explicitly but rely on discretized ground-plane grids and dense ground-truth supervision, limiting resolution and scalability in unlabeled settings. Since single-view depth is projectively ambiguous [24, 40], a road-relative height-to-depth ratio $\gamma = h/d$ offers a complementary representation, making γ dimensionless and tied to the ground. As shown in Fig. 2, doubling the scale of an object relative to the ground leaves the apparent vertical offset unchanged. From a single image the metric configuration is indistinguishable, so depth remains scale ambiguous, whereas in γ space, both configurations take the same dimensionless value, consistently tied to the road. With a known ground plane and camera height, this γ value converts to absolute height and depth in a simple closed-form way.

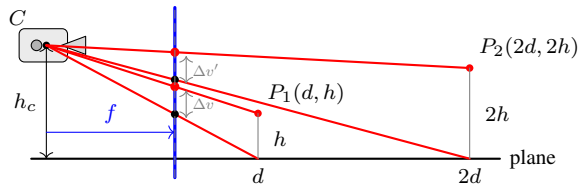

Figure 2. For points (d, h) and $(2d, 2h)$, $\Delta v = f\,h/d$ equals $\Delta v'$, the gap depends only on h/d. A known ground reference fixes metric scale, resolving monocular projective ambiguity.

Accordingly, we introduce Gamma from Mono (**GfM**), a single-frame approach that reframes monocular geometry around the dominant road plane, mitigating projective ambiguity and reducing scale to a single camera-height parameter. Our key contributions are:

- We propose a model that directly predicts a road-relative representation, comprising a global road-plane normal and a per-pixel height-to-depth ratio (γ), preserving near-road detail.
- To our knowledge, this is the first single-frame, self-supervised method that directly regresses γ for road-relative geometry, enabling explicit, interpretable estimates of road topography.
- We resolve metric scale from a dimensionless prediction, converting to metric depth using only known camera height and avoiding full extrinsic calibration or test-time fitting.

2. Related Work

Most prior work in monocular geometry estimation predicts per-pixel depth [20]. By contrast, the height-to-depth ratio γ has been a key parameter for multi-view reconstruction via planar parallax [13, 25, 26]. For example, MonoPP [7] uses γ only as a training-time scale cue distilled from multi-frame planar-parallax constraints. On the other hand, Yuan *et al*. [43] computes per-pixel γ from multi-frame homography alignment with LiDAR supervision. Beyond these, monocular methods remain depth-centric. Both approaches rely on multi-view cues or explicit ground-truth depth, and neither regresses γ directly from a single image in a self-supervised setting.

Scale ambiguity. Monocular depth estimation (MDE) has advanced greatly, yet models trained only on monocular images recover depth only up to an unknown scale factor [33, 40]. However, metric-scale depth is crucial for autonomous driving safety [31]. Recent foundational models such as Metric3D [42], UniDepth [18, 19], MoGe [33, 34], DepthAnything [40, 41], and Depth-Pro [4] include metric-depth variants, yet still exhibit small but persistent scale drift in novel environments. Moreover, these methods are trained on millions of images and use large transformer backbones, making them less suitable for resource-constrained real-time deployment on limited hardware. Traditional solutions recover scale through post-processing, such as fitting a known ground plane, or by integrating explicit cues such as vehicle speed, IMU data during training.

Motion-based supervision. Motion cues from vehicle sensors offer effective scale recovery for monocular depth. Velocity from odometry or GPS can directly constrain scale by aligning estimated motion with true displacement. Guizilini *et al*. [11] (PackNet) achieved metric-scaled depth by incorporating odometry-derived travel distances during training, supervising translation magnitude to align predicted scene motion with real-world displacement and reduce scale drift. Similarly, Zhang *et al*. [45] leveraged IMU-derived gravity and acceleration measurements to enforce vertical and horizontal consistency constraints, improving metric accuracy and robustness across scenes. Wagstaff and Kelly [30] extended this idea using full inter-frame pose supervision. These approaches demonstrate how motion-based cues, whether velocity, IMU, or pose, can yield scale-consistent monocular depth.

Ground plane constraints. Camera extrinsics with respect to a flat ground plane provide a valuable scale cue, widely used in autonomous driving. Sui *et al*. [28] enforced a ground-plane constraint to align predicted

depths with known geometry, but faced accuracy limitations due to calibration errors and road irregularities. VADepth [37] improved robustness by using attention-based losses on road pixels. In addition, GroCo [5] further introduced a network to detect the road plane and constrain its depth to calibrated metric values. This improves generalization to camera rotations but still requires precise extrinsics during both training and inference. Moreover, MonoPP [7] leveraged planar-parallax geometry in a multi-frame teacher-student framework, achieving state-of-the-art results but requiring computationally intensive training and accurate extrinsics. FUMET [14] recovered metric scale by first estimating camera height from the road plane. They then scale it with a per-frame factor from a learned vehicle-size prior (LSP) via silhouette–size matching. Training enforces sequence-wise camera-height invariance. However, reliance on the LSP (trained on DVM-CAR [12]) and segmentation quality can potentially bias scale estimates.

Summary and motivation. Monocular geometry estimation faces scale ambiguity, often addressed with external sensors or ground-plane assumptions, though these can be sensitive to calibration and scene geometry. In addition, most such approaches focus only on depth, limiting their ability to reconstruct fine road topography. Inspired by ground-plane priors [5, 7], we propose a height-sensitive monocular geometry estimation framework that directly predicts the ratio gamma from a single image. In contrast to MonoPP [7], which derives γ from two-view optical flow and is therefore view-dependent and prone to alignment errors, our method treats the ratio as a primary prediction target. This representation models extrinsic variations [13, 25], reduces calibration requirements, and encodes road-relative elevation, as illustrated in Fig. 3. Such height-aware perception benefits automotive tasks including slope estimation, elevation-change detection, road-damage identification, and obstacle recognition [1, 27].

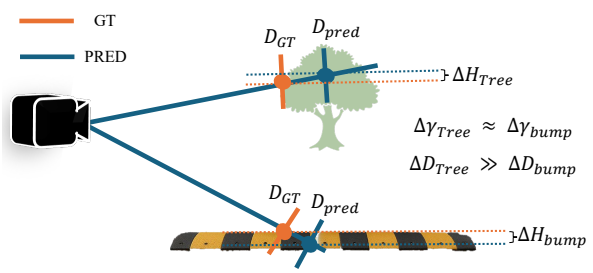

Figure 3. Smoothing a small bump causes negligible depth error compared to a tree, but in γ-space the errors are similar, revealing sensitivity to small height changes. A numerical example is provided in the supplementary material.

3. Method

Our method identifies the dominant road plane and estimates γ, the per-pixel height above the road plane divided by depth from the camera, in a fully self-supervised manner. In our experiments, direct height prediction in self-supervised training collapsed due to its unconstrained scale. Predicting γ instead couples height to depth, yielding a stable, scale-consistent value and sharper close-range geometry, as later shown in Tab. 1. We show in the inference model in Fig. 4, how our approach naturally decomposes the scene reconstruction into two interpretable components: a global road plane, described by the normal vector $\vec{N}_{\text{pred}}$, capturing the dominant road geometry and a local per-pixel structure γ modeling residual fine-grained variations [7, 13, 25, 43]. Consequently, our approach yields a scene-centric geometric representation that effectively captures slopes and recovers uneven terrain with high precision. Our method further uses camera height and intrinsics to recover metric geometry in a simple and fast post-processing step.

3.1. Problem Setup

Given a monocular target image I_t, the network θ_g predicts per-pixel $\gamma_t(u, v)$ and a single global road-plane normal $\vec{N}_{\text{pred}} \in \mathbb{R}^3$:

$$\gamma_t,\ \vec{N}_{\text{pred}} = \theta_g(I_t).$$

By definition,

$$\gamma_t(u, v) = \frac{h_t(u, v)}{d_t(u, v)},$$

where $d_t(u, v)$ is depth from the camera and $h_t(u, v)$ is height relative to the road plane. We adopt the convention that $\vec{N}_{\text{pred}}$ points upward, so heights are measured along $\vec{N}_{\text{pred}}$ and are positive above the road surface.

For training, we employ a self-supervised approach using neighboring source images $I_s = \{I_{t-1}, I_{t+1}\}$. A Pose Network θ_p estimates the 6-DoF relative pose $T_{t\to s}$ between I_t and I_s via $T_{t\to s} = \theta_p(I_t, I_s)$. Following [48], this enables self-supervised learning via novel-view synthesis. If accurate poses from other sensors are available, the pose network can be omitted.

3.2. Network Architecture

GfM is based on LiteMono-8M [44] and predicts γ_t and $\vec{N}_{\text{pred}}$. The final layer applies a sigmoid and maps its output to a predefined γ range using a signed log-space transform that increases precision near zero. Equation (1) defines the forward transform from γ to its log-space representation $\tilde{\gamma}$:

$$\tilde{\gamma} = \text{sign}(\gamma)\ \log\big(1 + \alpha|\gamma|\big), \tag{1}$$

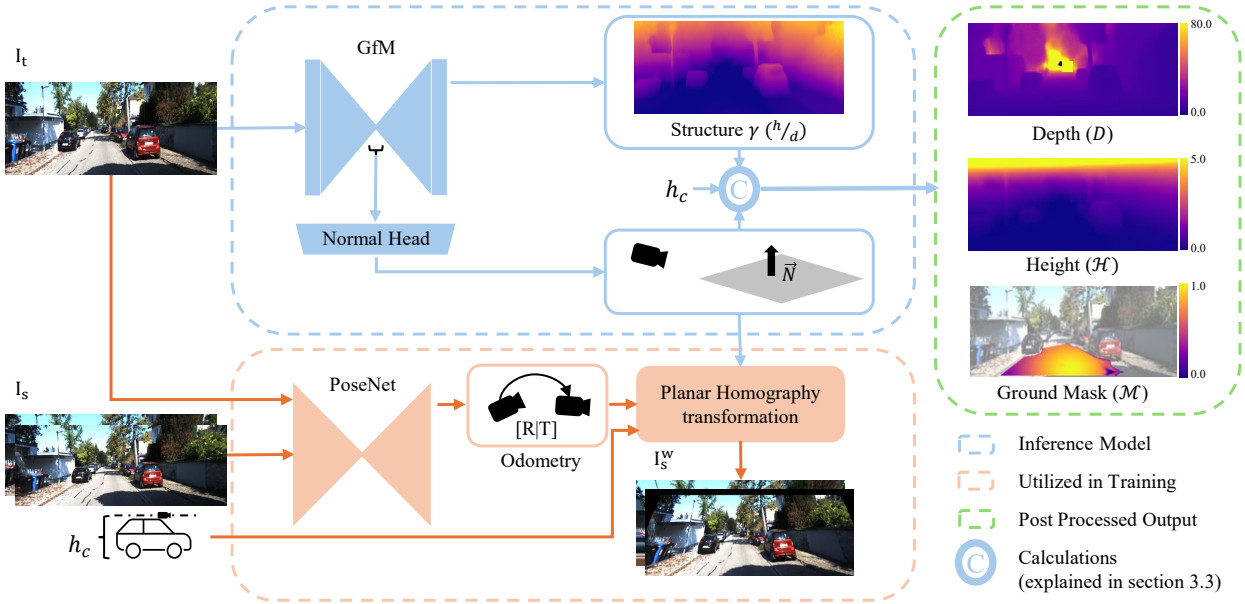

Figure 4. **Overview of our model architecture.** The main network predicts a per-pixel parameter γ, while PoseNet estimates the relative pose between I_s and I_t. From the relative pose, we compute a planar homography to align the planar road surface between the two images. Additionally, γ is used to infer depth, scene height, and a probabilistic road mask, as detailed in Sec. 3.3.

where $\alpha > 0$ controls the nonlinearity. Apply the same transform to the range endpoints to get $\tilde{\gamma}_{\min}$ and $\tilde{\gamma}_{\max}$. Given the sigmoid output $\sigma_t \in [0, 1]$, map it linearly in the transform domain:

$$\tilde{\gamma}_t = \tilde{\gamma}_{\min} + (\tilde{\gamma}_{\max} - \tilde{\gamma}_{\min})\, \sigma_t. \tag{2}$$

Finally, invert the transform to obtain γ_t:

$$\gamma_t = \operatorname{sign}(\tilde{\gamma}_t)\, \frac{\exp\big(|\tilde{\gamma}_t|\big) - 1}{\alpha}. \tag{3}$$

To model global scene geometry, we add a lightweight global normal head that predicts a single road-plane normal $\vec{N}_{\text{pred}}$. It applies attention pooling to the bottleneck feature map $x \in \mathbb{R}^{B\times C\times H\times W}$ to extract a road-focused descriptor, followed by a linear layer and ℓ_2-normalization to produce a unit vector. Further details are given in the supplementary material.

$$\vec{N}_{\text{pred}} = \frac{\text{Linear}\big(\text{AttnPool}(x)\big)}{\big\|\text{Linear}\big(\text{AttnPool}(x)\big)\big\|_2}. \tag{4}$$

Pose network. Following [10, 36, 44], we use a ResNet-18 backbone for pose estimation, taking two concatenated RGB frames (6 channels) as input.

3.3. Post-processing Network Outputs

Using a simple post-processing step, our network outputs are transformed to interpretable representations: metric depth, residual parallax flow, and a road mask for subsequent loss computation and analysis.

Computing depth. Given $\gamma_t(u, v)$ and intrinsics matrix K, we compute the metric depth $d_t(u, v)$ [43] as:

$$d_t(u, v) = \frac{h_c}{\gamma_t(u, v) + \vec{N}_{\text{pred}} \cdot (K^{-1}\mathbf{p}(u, v))}, \tag{5}$$

with $\mathbf{p}(u, v) = [u,\ v,\ 1]^\top$. The height relative to the road follows directly as $h_t(u, v) = \gamma_t(u, v) \cdot d_t(u, v)$.

Road-plane homography. Assuming the road is the dominant planar surface, we compute the homography between source and target views using the road planar homography computation in [7, 43]. Warping I_s with $H_{s\to t}$ yields the aligned image I_s^w, where the ground plane should ideally coincide with that of I_t.

Residual parallax flow. Points not lying on the dominant plane do not align under the homography. Their resulting misalignment, as explained in [7, 43], known as the planar parallax residual flow $\mathbf{u}^{\text{res}}_{s\to t}$ [26], is computed directly from γ_t:

$$u^{\text{res}}_{s\to t} = \frac{-\gamma_t \frac{T_z}{h_c}}{1 - \gamma_t \frac{T_z}{h_c}} \big(p_t - e_t\big), \tag{6}$$

where $T_z \neq 0$ is the translation along the z-axis, p_t the pixel coordinate, and e_t the epipole in I_t.

Probabilistic road mask. We compute a probabilistic road mask by comparing the predicted global normal $\vec{N}_{\text{pred}}$ with local surface normals $\mathbf{n}(u, v)$, obtained

by backprojecting depth [7, 38]. To prevent unintended interactions with the loss, we detach the depth before computing $\mathbf{n}(u,v)$, ensuring that the road mask remains unaffected during optimization. We then measure the angular deviation between $\mathbf{n}(u,v)$ and $\vec{N}_{\text{pred}}$:

$$\theta(u,v) \;=\; \cos^{-1}\left(\frac{\mathbf{n}(u,v)\cdot\vec{N}_{\text{pred}}}{\|\mathbf{n}(u,v)\|\;\|\vec{N}_{\text{pred}}\|}\right). \tag{7}$$

We convert this deviation into an angle-based probability map $p_{\text{angle}}(u,v)\in[0,1]$. We then multiply p_{angle} by a Gaussian spatial prior $g(u,v)$. The prior is centered horizontally, and its vertical mean is placed at 0.25 of the image height measured from the bottom edge to emphasize the near-field road region.

$$\mathcal{M}_{\text{road}}(u,v) \;=\; p_{\text{angle}}(u,v)\cdot g(u,v). \tag{8}$$

This weighting improves robustness to uphill/downhill road slopes and other geometric changes. An illustrative example is in Fig. 4, with further details in the supplementary material.

Target view synthesis. To reconstruct the target view I_t from the source I_s, we synthesize two images: $\hat{I}^d_{s\to t}$ by depth-based warping [5, 10, 44], and $\hat{I}^f_{s\to t}$ by planar homography plus residual flow [7]. First, we warp I_s with the homography $H_{s\to t}$:

$$I^w_s = \big\langle I_s,\ \text{warp}(H_{s\to t})\big\rangle, \tag{9}$$

then warp off-plane regions with a per-pixel residual flow $\mathbf{u}^{\text{res}}_{s\to t}$:

$$\hat{I}^f_{s\to t} = \big\langle I^w_s,\ \text{warp}(\mathbf{u}^{\text{res}}_{s\to t})\big\rangle, \tag{10}$$

where $\text{warp}(\cdot)$ denotes the sampling grid and $\langle I,\cdot\rangle$ bilinear resampling [3].

3.4. Losses

All pixel-level losses are averaged over valid pixels.

(1) Photometric loss. We measure photometric consistency between synthesized views ($\hat{I}^d_{s\to t}$, $\hat{I}^f_{s\to t}$) and the target view I_t using photometric error (pe) [10]:

$$\text{pe}\big(\hat{I}_t, I_t\big) = \alpha\,\frac{1-\text{SSIM}(\hat{I}_t, I_t)}{2} + (1-\alpha)\,\big\|\hat{I}_t - I_t\big\|_1, \tag{11}$$

where α balances SSIM [35] and L_1. Following Monodepth2 [10], we employ minimum reprojection across source images and use an implicit auto-masking strategy [10] to mitigate occlusions and dynamic objects:

$$\mathcal{L}_{\text{photo}} = \mathcal{L}^d + \mathcal{L}^f = \min_s \text{pe}\big(\hat{I}^d_{s\to t}, I_t\big) + \min_s \text{pe}\big(\hat{I}^f_{s\to t}, I_t\big). \tag{12}$$

While the first photometric loss $\mathcal{L}^d$, which compares warped images synthesized using depth, ensures consistency on large scales, the second term $\mathcal{L}^f$ supervises fine-grained structures near the road surface.

(2) Homography alignment loss. We compute a masked photometric loss using I^w_s and I_t, restricting the loss to the dominant road plane:

$$\mathcal{L}_{\text{homo}} \;=\; \mathcal{M}_{\text{road}}\cdot\text{pe}\big(I^w_s, I_t\big), \tag{13}$$

where $\mathcal{M}_{\text{road}}$ is the probabilistic road mask defined in Sec. 3.3. This loss enforces accurate alignment of planar road regions between views [7].

(3) Normal consistency loss. To avoid degenerate or inverted normals, we penalize deviations between the predicted normal $\vec{N}_{\text{pred}}$ and a reference normal $\vec{N}_{\text{ref}}$ using a hinge-based angular loss:

$$\cos(\Delta\theta) = \frac{\vec{N}_{\text{pred}}\cdot\vec{N}_{\text{ref}}}{\|\vec{N}_{\text{pred}}\|\;\|\vec{N}_{\text{ref}}\|}, \tag{14}$$

$$\mathcal{L}_{\text{norm}} = \big[\,1-\cos(\Delta\theta)\,\big] \;+\; \text{ReLU}\big(\cos(\theta_{\text{thres}}) - \cos(\Delta\theta)\big)^2. \tag{15}$$

This formulation encourages physically plausible normals, imposing larger penalties when angular deviations exceed the threshold θ_{thres}.

Total loss. Including an additional edge-aware smoothness term $\mathcal{L}_{\text{smooth}}$ [9, 10], our final objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{photo}} + \mathcal{L}_{\text{homo}} + \lambda_{\text{norm}}\mathcal{L}_{\text{norm}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}}, \tag{16}$$

where λ_{norm}, λ_{smooth} balance loss components.

4. Experiments

Evaluation overview. We evaluate on KITTI [8], the standard monocular depth benchmark for vehicular application, and RSRD [46], which contains varied height perturbations. KITTI allows benchmarking against prior work, while RSRD tests road-topography sensitivity and enables evaluation on diverse terrain. All experiments use an image resolution of 640×192.

For KITTI [8], we use the Eigen split [6]: 39,180 monocular triplets for training, 4,424 images for validation, and 697 for testing. Evaluation follows the improved benchmark [29] with a GT camera height of ≈ 1.65 m [15]. Depth is clipped at 80 m for both depth and γ evaluations. On the other hand, given the textureless surface of RSRD, we rely on the provided ground-truth poses. We train on 7,040 sparse triplets and evaluate on the 600 images with dense ground truth.

(a) KITTI [8, 29]

	Published in	Method	Train	#Params	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Depth	WACV 2025	MonoPP [7]	M+camH	34.57M	0.089	0.545	3.864	**0.134**	0.913	0.983	0.995
	RA-L 2022	VADepth [37]	M+camH	18.8M	0.091	0.555	3.871	**0.134**	0.913	0.983	0.995
	ECCV 2024	GroCo [5]	M+camH	34.65M	0.089	0.517	3.815	**0.134**	0.910	**0.984**	0.995
	arXiv 2025	UniDepthV2-L [19]	F	353.22M	0.123	0.576	**3.612**	0.139	**0.933**	0.991	**0.997**
	arXiv 2025	MoGe2-L [34]	F	330.90M	0.180	0.732	4.122	0.217	0.633	0.969	0.994
	ICLR 2025	DepthPro [4]	F	951.99M	0.117	**0.502**	3.721	0.149	0.874	0.979	0.995
	ICLR 2025	DepthPro [4] †	F	951.99M	0.072	0.349	3.489	0.112	0.941	0.990	0.998
	-	GfM (ours)	M+camH	**8.88M**	**0.085**	0.554	4.082	0.139	0.916	0.983	0.995
Depth ≤ 20	WACV 2025	MonoPP [7]	M+camH	34.57M	0.067	0.121	1.109	0.093	0.957	0.994	**0.999**
	ECCV 2024	GroCo [5]	M+camH	34.65M	0.067	0.125	1.141	0.094	0.955	0.995	**0.999**
	arXiv 2025	UniDepthV2-L [19]	F	353.22M	0.105	0.188	1.383	0.114	0.966	**0.996**	**0.999**
	ICLR 2025	DepthPro [4]	F	951.99M	0.105	0.202	1.362	0.122	0.913	0.990	0.997
	ICLR 2025	DepthPro [4] †	F	951.99M	0.051	0.071	0.888	0.073	0.982	0.998	0.999
	-	GfM (ours)	M+camH	**8.88M**	**0.058**	**0.103**	**1.036**	**0.085**	**0.967**	**0.996**	**0.999**

	Published in	Method	Train	#Params	PP	Abs Diff ↓	RMSE ↓	RMSE log ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Gamma	ECCV 2024	GroCo [5]	M+camH	34.65M	✓	**0.012**	**0.018**	0.840	0.66	0.768	**0.822**
	ICLR 2025	DepthPro [4]	F	951.99M	✓	0.014	0.020	0.885	0.655	0.748	0.797
	ICLR 2025	DepthPro [4] †	F	951.99M	✓	0.011	0.016	0.718	0.726	0.811	0.854
	-	GfM (ours)	M+camH	**8.88M**	-	**0.012**	**0.018**	**0.801**	**0.706**	**0.780**	0.820

(b) RSRD [46]

	Published in	Method	Train	#Params	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Depth	-	Monodepth2 [10] ‡	M+Pose	14.3M	0.059	0.032	0.393	0.076	0.987	**0.996**	**0.999**
	ICLR 2025	DepthPro [4]	F	951.99M	0.344	0.763	1.807	0.484	0.294	0.588	0.781
	ICLR 2025	DepthPro [4] †	F	951.99M	0.065	0.064	0.600	0.086	0.971	0.996	0.999
	arXiv 2025	UniDepthV2-L [19]	F	353.22M	0.627	2.339	3.417	0.478	0.112	0.380	0.880
	arXiv 2025	UniDepthV2-L [19] †	F	353.22M	0.043	0.027	0.417	0.058	0.994	0.999	1.000
	arXiv 2025	MoGe2-L [34]	F	330.90M	0.576	1.666	2.664	0.445	0.078	0.550	0.910
	arXiv 2025	MoGe2-L [34] †	F	330.90M	0.037	0.021	0.377	0.053	0.993	0.999	1.000
	-	GfM (ours)	M+Pose	**8.88M**	**0.034**	**0.028**	**0.372**	**0.055**	**0.990**	**0.996**	**0.999**

	Published in	Method	Train	#Params	PP	Abs Diff ↓	RMSE ↓	RMSE log ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Gamma	-	Monodepth2 [10] ‡	M+Pose	14.3M	✓	0.015	0.023	**1.124**	0.459	0.463	0.471
	ICLR 2025	DepthPro [4]	F	951.99M	✓	0.211	0.226	2.807	0.057	0.064	0.077
	ICLR 2025	DepthPro [4] †	F	951.99M	✓	0.016	0.024	1.234	0.500	0.521	0.558
	arXiv 2025	UniDepthV2-L [19]	F	353.22M	✓	0.120	0.132	2.191	0.045	0.101	0.168
	arXiv 2025	UniDepth [19] †	F	353.22M	✓	0.037	0.044	0.260	0.516	0.921	0.991
	arXiv 2025	MoGe2-L [34]	F	330.90M	✓	0.142	0.155	2.216	0.018	0.034	0.056
	arXiv 2025	MoGe2-L [34] †	F	330.90M	✓	0.027	0.032	0.243	0.629	0.902	0.977
	-	GfM (ours)	M+Pose	**8.88M**	-	**0.009**	**0.015**	1.142	**0.587**	**0.591**	**0.605**

Table 1. **Quantitative evaluation on KITTI and RSRD datasets.** The upper rows present depth metrics evaluated against their respective ground-truth, while the lower rows present the height-to-depth ratio (γ) performance, computed using the improved KITTI ground-truth [29] and dense point clouds for RSRD [46]. Abbreviations: *M* (monocular-only), *PP* (RANSAC plane-fitting post-processing), *camH* (camera-height supervision), † (median-scaled by ground-truth. Thus, excluded from direct comparison due to scale drift correction), *V* velocity regression, *Pose* full GT pose supervision, *SI* object size priors from internet data, *F* foundational models trained on massive data, ‡ Monodepth2 baseline trained on RSRD with GT poses.

Implementation details. Our method is implemented in PyTorch and trained on a single NVIDIA V100 GPU with a batch size of 12. For KITTI, training consists of two phases: (1) Pretraining as monocular depth estimation for 10 epochs, jointly learning depth and metric-scaled pose from PoseNet. (2) Switching to γ-prediction, training for 25 additional epochs on KITTI (35 epochs total).

For RSRD, we use ground-truth poses rather than jointly optimizing pose and structure, as learning both simultaneously on textureless data is unreliable. Consequently, we omit the homography loss ($\mathcal{L}_{\text{homo}}$) for RSRD, since its primary function is enforcing metric scale in jointly learned depth and pose.

The network outputs a sigmoid map that is converted to γ via the signed log-space transform in Eq. (1) with $\alpha = 0.5$. We set the γ range to $[-0.1, 5.0]$ for KITTI and $[-0.5, 2.0]$ for RSRD. For the normal-

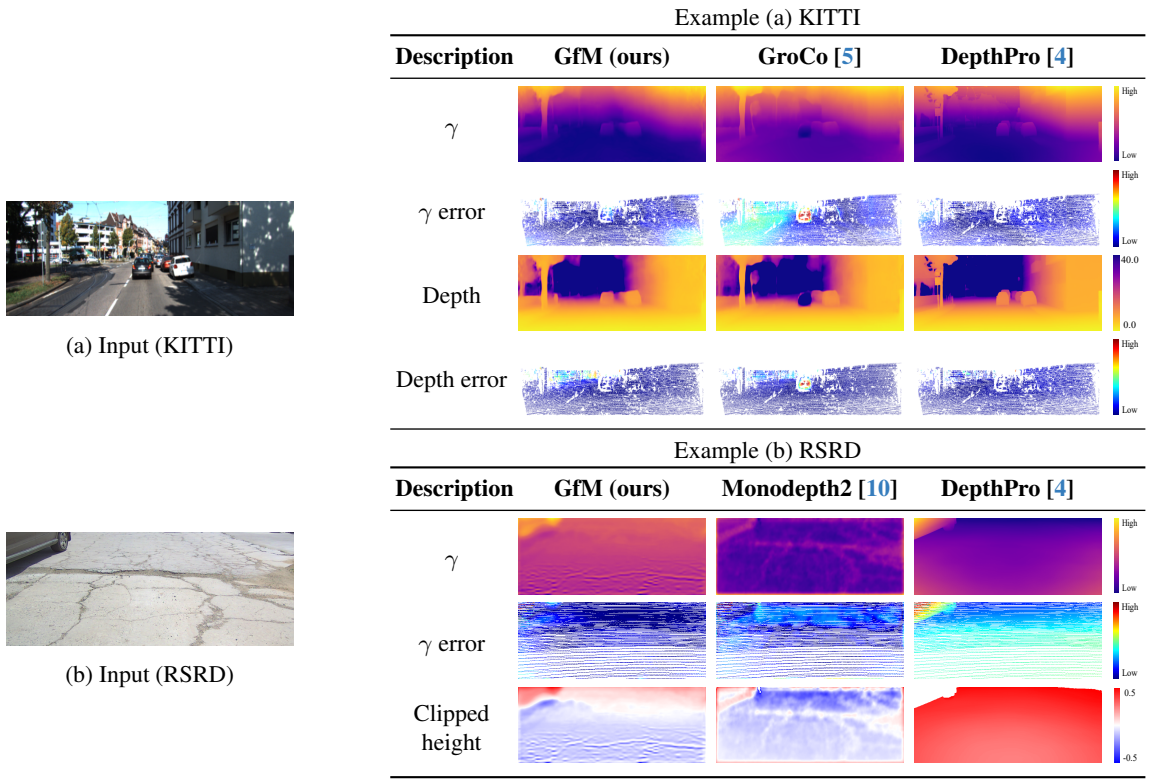

Figure 5. **Qualitative comparison on KITTI and RSRD.** Left: (a) KITTI [8] input and (b) RSRD [46] input. Right: (a) KITTI example comparing GfM, GroCo [5], and DepthPro [4] on γ, γ error (Abs Diff), depth, and depth error (Abs Rel). (b) RSRD example comparing GfM, Monodepth2 [10], and DepthPro on γ prediction, γ error map (Abs Diff), and height predictions clipped to $[-0.5, 0.5]$. Colormaps for each metric are shown in the last column.

consistency loss we use an angular threshold $\theta_{\text{thres}} = 5°$ and reference normals $\vec{N}_{\text{ref}} = [0, -1, 0]$ for KITTI and $[0, -0.95, -0.29]$ for RSRD. We optimize with AdamW (weight decay 10^{-2}) and a cosine learning-rate schedule initialized at 5×10^{-4}; loss weights are $\lambda_{\text{norm}} = 0.1$ and $\lambda_{\text{smooth}} = 10^{-2}$. During training, we follow the data-augmentation and auto-masking strategies of [7, 10].

Baselines. We focus on metric-scaled self-supervised methods as well as several foundation models, with DepthPro [4] highlighted in the main tables as a representative foundation model due to its high sensitivity to fine geometric variations. For KITTI, comparisons include methods such as GroCo [5], using publicly available pretrained weights. Since no prior self-supervised method was trained on RSRD and training code for recent models is unavailable, we construct a baseline by training Monodepth2 [10] with known poses and identical inputs to GfM. For each depth map (prediction and GT), γ is computed from depth by fitting the ground plane with RANSAC (10,000 iterations, $0.01\,\text{m}$ inlier threshold). More details and further evaluations are in the supplementary material.

Evaluation metrics. For depth estimation, we use standard metrics: *Abs Rel*, *Sq Rel*, *RMSE*, *RMSE log*, and δ-accuracies $(\delta_1, \delta_2, \delta_3)$, following prior work [7, 10, 36, 44]. For γ (signed height-to-depth), standard relative metrics can become unstable near or below zero. As a result, we define:

- *Abs Diff*: mean absolute difference $\left|\gamma_{\text{pred}} - \gamma_{\text{gt}}\right|$,
- *RMSE* & *RMSE log*: computed with a small offset to handle sign changes,
- Modified δ-accuracies: pixels are correct if $\left|\gamma_{\text{pred}} - \gamma_{\text{gt}}\right| < 0.01$ or satisfy the ratio-based condition [10].

These adjustments ensure a robust evaluation for the signed γ predictions.

Results. As summarized in Tab. 1a, GfM delivers competitive performance in general depth estimation on KITTI while using significantly fewer parameters and less supervision, and achieves state-of-the-art results in near-range depth and γ metrics. Qualitative examples in Fig. 5 further illustrate these strengths: in example (a), our method outperforms prior work GroCo [5] in challenging dynamic-object scenarios, though far-away dynamic objects remain difficult, as presented in

Table 2. Ablation study on KITTI benchmark (capped at 80 m). Lower is better except $\delta \uparrow$. **Bold** indicates best in column.

Variant	Gamma					Depth				
	Abs Diff ↓	RMSE ↓	RMSE log ↓	δ_1 ↑	δ_3 ↑	Abs Rel ↓	RMSE ↓	RMSE log ↓	δ_1 ↑	δ_3 ↑
GfM (full)	**0.012**	**0.018**	**0.801**	**0.706**	**0.801**	**0.085**	**4.082**	0.139	**0.916**	**0.995**
w/o predicted normal (fixed $\vec{N}$)	0.013	0.020	0.909	0.658	0.796	0.095	4.258	0.150	0.903	0.995
w/o log-space transform	0.016	0.024	1.119	0.558	0.695	0.095	4.302	0.144	0.907	0.994
w/o ImageNet pretraining	0.015	0.024	1.033	0.645	0.777	0.111	4.648	0.168	0.872	0.989
w/o probabilistic road mask	0.013	0.021	0.944	0.678	0.796	0.087	4.210	**0.138**	0.908	0.995

the supplementary material. On the new RSRD dataset (Tab. 1b), GfM surpasses all evaluated methods in both γ and depth metrics, and produces more reliable 3D reconstructions, as seen in Fig. 6, where foundation models often misinterpret road slopes without domain adaptation. Example (b) in Fig. 5 shows how our accurate γ prediction captures the fine details of road topography with higher fidelity than both conventional depth-based self-supervised models and large-scale foundation models. Importantly, γ evaluation is invariant to camera height assumptions: perturbing the height only rescales recovered metric depth, leaving γ unchanged. This makes γ a physically interpretable indicator of road-relative geometry.

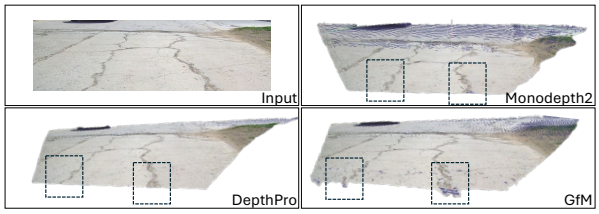

Figure 6. **Point Cloud Comparison.** Computed 3D point clouds from different models on a sample image from RSRD [46]. DepthPro exhibits scale drift and slightly flattens the road surface. Meanwhile, Monodepth2 [10], despite being trained on the dataset, still struggles to capture road topography accurately. Additional qualitative examples are in the supplementary material.

Ablation study. Table 2 reports ablations on KITTI. Although omitted from the table due to catastrophic metric failures, some components have particularly strong influence. Removing the normal constraint $\mathcal{L}_{\text{norm}}$ causes $\vec{N}_{\text{pred}}$ to collapse into degenerate or flipped orientations, degrading road-topography estimation and photo-consistency; this results in a γ Abs Diff of 2.282, $\delta_1^{\gamma} = 0.00$, depth Abs Rel of 0.937, and $\delta_1^{\text{depth}} = 0.00$. Removing the homography alignment loss $\mathcal{L}_{\text{homo}}$ eliminates the dominant-plane scale cue, leaving γ largely unaffected but severely degrading metric depth. Other components contribute as follows (Tab. 2, rows 1–4).

1. **Learned normal** Predicting $\vec{N}_{\text{pred}}$ outperforms using KITTI extrinsics directly, consistent with prior findings that vehicle tilt varies in real driving due to suspension and load [15].
2. **Signed log-space mapping.** Replacing the signed log-space transformation (Sec. 3.2) with a linear mapping, as in [7, 10, 36, 44], caused training to diverge, unless the learning rate was reduced by 10×, even then, performance was worse.
3. **ImageNet pretraining.** Pretraining yields consistent gains for both γ and depth, matching trends in self-supervised depth [10, 44].
4. **Probabilistic road mask $\mathcal{M}_{\text{road}}$.** Replacing our probabilistic mask with a binary mask, as in [7, 38], reduces performance; the probabilistic near-field weighting is more robust.

5. Conclusion

In this work, we tackled a core limitation of monocular depth estimation: its inability to deliver explicit scene geometry for real-world vehicular tasks due to scale ambiguity, extrinsic calibration sensitivity, and limited structural understanding. We introduced a lightweight model that predicts the γ representation and a global road-plane normal, which unifies depth and height and reduces reliance on extrinsic calibration. In addition, it is trained self-supervised on monocular driving video. Given known camera height, GfM turns a single image into metric, detailed near-field geometry that exposes subtle road irregularities for safer planning.

Limitations. Metric scale depends on camera-height accuracy, as miscalibration introduces a global scale factor while preserving relative geometry. As with other self-supervised methods, performance can degrade on dynamic objects, especially distant objects.

Future work. We will extend the framework toward a foundation model for vehicular perception by self-supervising γ on large-scale monocular driving videos, recovering metric scale at test time via the known camera height. In addition, we will explore using γ directly for on-road control and planning where full 3D reconstruction is unnecessary.

References

[1] Akmalbek Abdusalomov, Sabina Umirzakova, Makhkamov Bakhtiyor Shukhratovich, Azamat Kakhorov, and Young-Im Cho. Breaking new ground in monocular depth estimation with dynamic iterative refinement and scale consistency. *Applied Sciences (2076-3417)*, 15(2), 2025. 1, 3

[2] Mehrnaz Farokhnejad Afshar, Zahra Shirmohammadi, Seyyed Amir Ali Ghafourian Ghahramani, Azadeh Noorparvar, and Ali Mohammad Afshin Hemmatyar. An efficient approach to monocular depth estimation for autonomous vehicle perception systems. *Sustainability*, 15 (11):8897, 2023. 1

[3] Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2800–2810, 2018. 5

[4] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, 2025. 1, 2, 6, 7

[5] Aurélien Cecille, Stefan Duffner, Franck Davoine, Thibault Neveu, and Rémi Agier. GroCo: Ground constraint for metric self-supervised monocular depth. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2, 3, 5, 6, 7

[6] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 5

[7] Gasser Elazab, Torben Gräber, Michael Unterreiner, and Olaf Hellwich. MonoPP: Metric-scaled self-supervised monocular depth estimation by planar-parallax geometry in automotive applications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. 2, 3, 4, 5, 6, 7, 8

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. 1, 5, 6, 7

[9] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 5

[10] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 4, 5, 6, 7, 8

[11] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2485–2494, 2020. 2

[12] Jingmin Huang, Bowei Chen, Lan Luo, Shigang Yue, and Iadh Ounis. Dvm-car: A large-scale automotive dataset for visual marketing research and applications. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 4140–4147. IEEE, 2022. 3

[13] Michal Irani and Prabu Anandan. Parallax geometry of pairs of points for 3d scene analysis. In *Computer Vision—ECCV'96: 4th European Conference on Computer Vision Cambridge, UK, April 15–18, 1996 Proceedings, Volume I 4*, pages 17–30. Springer, 1996. 2, 3

[14] Genki Kinoshita and Ko Nishino. Camera height doesn't change: Unsupervised training for metric monocular road-scene depth estimation. In *European Conference on Computer Vision*, pages 57–73. Springer, 2025. 2, 3

[15] Karlo Koledic, Luka Petrovic, Ivan Markovic, and Ivan Petrovic. Gvdepth: Zero-shot monocular depth estimation for ground vehicles based on probabilistic cue fusion. *arXiv preprint arXiv:2412.06080*, 2024. 5, 8

[16] Shuaijun Li, Guilin Zhang, Xiangyu Lei, Xiao Yu, Huihuan Qian, and Yangsheng Xu. Trajectory tracking control of a unicycle-type mobile robot with a new planning algorithm. In *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 780–786. IEEE, 2017. 1

[17] Guanqun Liang, Tong Zhao, Zhengwei Shangguan, Ningfei Li, Mingyu Wu, Jingcheng Lyu, Yongchang Du, and Yintao Wei. Experimental study of road identification by lstm with application to adaptive suspension damping control. *Mechanical Systems and Signal Processing*, 177:109197, 2022. 1

[18] LuigiPiccinelli, Yung-HsuYang, ChristosSakaridis, MattiaSegu, SiyuanLi, LucVanGool, and FisherYu. UniDepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

[19] LuigiPiccinelli, ChristosSakaridis, Yung-HsuYang, MattiaSegu, SiyuanLi, WimAbbeloos, and LucVanGool. UniDepthv2: Universal monocular metric depth estimation made simpler, 2025. 1, 2, 6

[20] Armin Masoumian, Hatem A Rashwan, Julián Cristiano, M Salman Asif, and Domenec Puig. Monocular depth estimation using deep learning: A review. *Sensors*, 22 (14):5353, 2022. 2

[21] Takahiro Miki, Lorenz Wellhausen, Ruben Grandia, Fabian Jenelten, Timon Homberger, and Marco Hutter. Elevation mapping for locomotion and navigation using gpu. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2273–2280. IEEE, 2022. 1, 2

[22] Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. Lost and found: detecting small road hazards for self-driving vehicles. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1099–1106. IEEE, 2016. 1

[23] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocu-

lar depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 1

[24] Tom Roussel, Luc Van Eycken, and Tinne Tuytelaars. Monocular depth estimation in new environments with absolute scale. In *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 1735–1741. IEEE, 2019. 2

[25] Sawhney. 3d geometry from planar parallax. In *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 929–934. IEEE, 1994. 2, 3

[26] Shashua and Navab. Relative affine structure: Theory and application to 3d reconstruction from perspective views. In *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 483–489. IEEE, 1994. 2, 4

[27] Gautam Shetty, Sabir Hossain, Chuan Hu, and Xianke Lin. Road slope prediction and vehicle dynamics control for autonomous vehicles. *arXiv preprint arXiv:2210.05741*, 2022. 1, 3

[28] Wei Sui, Teng Chen, Jiaxin Zhang, Jiao Lu, and Qian Zhang. Road-aware monocular structure from motion and homography estimation. *arXiv preprint arXiv:2112.08635*, 2021. 2

[29] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017. 5, 6

[30] Brandon Wagstaff and Jonathan Kelly. Self-supervised scale recovery for monocular depth and egomotion estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2620–2627. IEEE, 2021. 2

[31] Peide Wang. Research on comparison of lidar and camera in autonomous driving. In *Journal of Physics: Conference Series*, page 012032. IOP Publishing, 2021. 2

[32] Ruihao Wang, Jian Qin, Kaiying Li, Yaochen Li, Dong Cao, and Jintao Xu. Bev-lanedet: An efficient 3d lane detection based on virtual camera via key-points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1002–1011, 2023. 1

[33] 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, 2024. 2

[34] 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, 2025. 2, 6

[35] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[36] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1164–1174, 2021. 4, 7, 8

[37] Jie Xiang, Yun Wang, Lifeng An, Haiyang Liu, Zijun Wang, and Jian Liu. Visual attention-based self-supervised absolute depth estimation using geometric priors in autonomous driving. *IEEE Robotics and Automation Letters*, 7(4):11998–12005, 2022. 2, 3, 6

[38] Feng Xue, Guirong Zhuo, Ziyuan Huang, Wufei Fu, Zhuoyue Wu, and Marcelo H Ang. Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2330–2337. IEEE, 2020. 5, 8

[39] Lei Yang, Kaicheng Yu, Tao Tang, Jun Li, Kun Yuan, Li Wang, Xinyu Zhang, and Peng Chen. Bevheight: A robust framework for vision-based roadside 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21611–21620. IEEE, 2023. 1

[40] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 2

[41] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2025. 1, 2

[42] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 1, 2

[43] Haobo Yuan, Teng Chen, Wei Sui, Jiafeng Xie, Lefei Zhang, Yuan Li, and Qian Zhang. Monocular road planar parallax estimation. *IEEE Transactions on Image Processing*, 2023. 2, 3, 4

[44] Ning Zhang, Francesco Nex, George Vosselman, and Norman Kerle. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18537–18546, 2023. 3, 4, 5, 7, 8

[45] Sen Zhang, Jing Zhang, and Dacheng Tao. Towards scale-aware, robust, and generalizable unsupervised monocular depth estimation by integrating imu motion dynamics. In *European Conference on Computer Vision*, pages 143–160. Springer, 2022. 2

[46] Tong Zhao, Chenfeng Xu, Mingyu Ding, Masayoshi Tomizuka, Wei Zhan, and Yintao Wei. Rsrd: A road surface reconstruction dataset and benchmark for safe and comfortable autonomous driving. *arXiv preprint arXiv:2310.02262*, 2023. 1, 5, 6, 7, 8

[47] Tong Zhao, Lei Yang, Yichen Xie, Mingyu Ding, Masayoshi Tomizuka, and Yintao Wei. Roadbev: Road

surface reconstruction in bird's eye view. *arXiv preprint arXiv:2404.06605*, 2024. 1, 2

[48] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3