GVDepth: Zero-Shot Monocular Depth Estimation for Ground Vehicles based on Probabilistic Cue Fusion

Karlo Koledić Luka Petrović Ivan Marković Ivan Petrović

University of Zagreb Faculty of Electrical Engineering and Computing

{karlo.koledic, luka.petrovic, ivan.markovic, ivan.petrovic}@fer.unizg.hr

Abstract

Generalizing metric monocular depth estimation presents a significant challenge due to its ill-posed nature, while the entanglement between camera parameters and depth amplifies issues further, hindering multi-dataset training and zero-shot accuracy. This challenge is particularly evident in autonomous vehicles and mobile robotics, where data is collected with fixed camera setups, limiting the geometric diversity. Yet, this context also presents an opportunity: the fixed relationship between the camera and the ground plane imposes additional perspective geometry constraints, enabling depth regression via vertical image positions of objects. However, this cue is highly susceptible to overfitting, thus we propose a novel canonical representation that maintains consistency across varied camera setups, effectively disentangling depth from specific parameters and enhancing generalization across datasets. We also propose a novel architecture that adaptively and probabilistically fuses depths estimated via object size and vertical image position cues. A comprehensive evaluation demonstrates the effectiveness of the proposed approach on five autonomous driving datasets, achieving accurate metric depth estimation for varying resolutions, aspect ratios and camera setups. Notably, we achieve comparable accuracy to existing zero-shot methods, despite training on a single dataset with a single-camera setup. Project website: ***https://unizgfer-lamor/gvdepth.github.io/***

1. Introduction

Accurate 3D reconstruction is one of the fundamental computer vision challenges, with far-reaching applications across various fields. Unlike multi-view approaches that rely on complex dense matching and pose estimation, learning-based Monocular Depth Estimation (MDE) exploits semantic and geometric cues to infer pixel-wise depth maps from a single image. However, *these cues are often specific to a narrow distribution of the training data*, resulting in poor generalization to out-of-distribution scenarios.

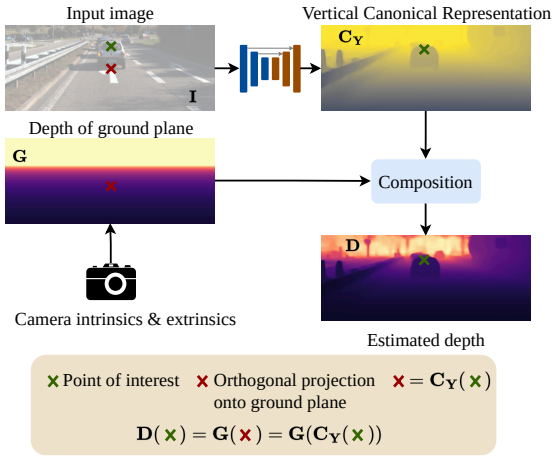


Figure 1. **Proposed Vertical Canonical Representation.** We introduce a novel intermediate representation that enables exploitation of underlying planar geometry priors, facilitating learning and generalization across varying camera setups. Under this representation, camera parameters and depth are completely disentangled, enabling depth regression via known depth of the ground plane.

In addition to the environmental domain gap, MDE is particularly sensitive to the domain gap induced by the perspective geometry parameters of the camera system. This issue is present in most supervised [1, 8, 18, 26, 41] and self-supervised methods [10, 11, 33, 42–45] that are trained on a dataset collected with a single camera system. Models tend to overfit to this perspective geometry bias, leading to significant performance degradation when inferring depth with different cameras. Unfortunately, training a MDE model on data with a wide distribution of perspective geometries is far from trivial. Due to the inherent ambiguity of the MDE problem and entanglement of depth and camera parameters [27], many methods resort to relative depth estimation with unknown shift and scale [6, 14, 28, 34, 36, 38, 39]. Recently, zero-shot metric depth estimation has been achieved with various approaches, including embedding [12], estimating [3, 27] or disentangling the camera

parameters [40]. While these methods present impressive results, *they require large-scale training on data acquired from multiple camera systems*, which is not always practical or cost-effective.

One specific area with numerous applications across varying fields is MDE for ground-based vehicles and robots. In this context, an additional perspective geometry constraint arises due to the fixed position of the camera relative to the ground plane. This constraint induces an additional depth cue: the vertical image position of the ground-contact point, which has been shown to be a dominant cue in autonomous driving scenarios [7].

These findings motivate us to utilize the vertical image position cue and the inherent planarity of the scene in order to enable accurate depth estimation and improve generalization for arbitrary intrinsic and extrinsic camera parameters. Contrary to the object size cue, widely used by the aforementioned zero-shot methods [3, 12, 27, 40], *we argue that the vertical image position presents a cue that is more fundamentally rooted in the perspective geometry configuration of the scene*. However, without proper modeling, this cue is highly susceptible to overfitting to the specific camera system present in the training data [16, 17, 25].

To leverage this cue, while enabling training with diverse perspective geometries and zero-shot generalization, in this paper we propose a method to ensure its consistency for arbitrary camera setups. Specifically, we introduce a novel intermediate depth representation termed *Vertical Canonical Representation*. Rather than directly regressing depth values, our model predicts vertical image positions of the orthogonal projections of points onto the ground plane, as visualized in Fig. 1. By utilizing the known depth of the ground plane, this approach allows the mapping between the canonical representation and depth space to adapt with varying camera parameters, effectively disentangling the vertical position cue from specific perspective geometry. To further address potential errors in certain regions, we introduce an architecture that adaptively fuses depth predictions from both vertical position and object size cues, guided by dynamically estimated uncertainties.

We comprehensively evaluate our method's zero-shot capability on five autonomous driving datasets, analyzing its robustness and accuracy across various cues. We demonstrate that our approach matches the accuracy of existing zero-shot methods, despite using a single datasets collected with a single camera system.

To summarize, we present the following contributions:

- a novel canonical transformation that enables the exploitation of vertical image position cue across diverse perspective geometries, enhancing both training scalabilty and zero-shot generalization,
- a model architecture for the probabilistic fusion of depth maps estimated using object size and vertical image position cues,
- a thorough ablation of different depth cues, with insights previously unseen in the literature.

2. Related Work

Domain-dependent Monocular Depth Estimation. In a seminal work [8], learning-based MDE has been established as a supervised dense regression task via ground-truth data obtained by reprojection of LiDAR scans. Subsequent works improve upon the original approach by introducing various architectural novelties, such as convolutional models with skip connections between encoder and decoder [18], integration of conditional random fields [19, 41], adversarial training [13, 21] or attention-based architectures [1, 26]. Unfortunately, these methods are specifically fine-tuned for accuracy on well established indoor [31] and outdoor [9] benchmarks, *experiencing significant accuracy degradation under distribution shifts* [17].

Domain-agnostic Monocular Depth Estimation. The inherent ambiguity and interdependence between camera parameters and depth present challenges for training a robust MDE model that generalizes across a wide range of environments. As datasets are usually collected with a single camera system, models tend to act discriminatively and associate perspective geometry parameters with the pertaining environment. Relative depth estimation methods [6, 14, 28, 34, 36, 38, 39] bypass camera parameter dependency by estimating depth with an unknown scale. While these approaches achieve visually impressive results, *their inability to estimate metric depth limits their applicability*, often necessitating integration with additional sensors.

ZoeDepth [2] is the first work to show satisfactory metric accuracy across different domains by fine-tuning a pretrained relative depth estimation model on a single dataset with metric scale. ZeroDepth [12] chooses a different approach by providing the model with embedding of camera parameters, effectively disambiguating the features from the specifics of the camera system. Metric3D [40] addresses the focal length-depth ambiguity by estimating depth in a canonical space defined through focal length denormalization. UniDepth [27] introduces a universal approach, simultaneously predicting both the depth and camera parameters, enabling metric depth estimation for images acquired with unknown sensors. Although these methods demonstrate impressive results, *they rely on abundance of training data and computational resources*. Furthermore, they are not truly zero-shot, as they tend to overfit to the typically high training resolution, hindering real-time performance on resource-constrained devices that require low-resolution estimation.

Planar Constraints for Depth Estimation. Real-world scenes are often highly regular, featuring numerous planar structures that can serve as additional cues to guide depth estimation [24, 30]. A highly relevant application of MDE

is in autonomous driving and mobile robotics, where a fixed relation between the camera system and the ground plane induces additional perspective geometry constraints. It has been shown that models heavily utilize the corresponding ground-contact cues, completely disregarding the object size cue [7]. Moreover, incorporating ground plane constraint within the model enhances overall model accuracy, both in supervised [37], and self-supervised training [4, 33]. However, these methods leverage the known perspective geometry constraints to enhance depth estimation only for road pixels, leading to limited generalization capabilities.

3. Proposed Method

Preliminaries. In this work, we examine an extremely common application domain of MDE: ground vehicles or a robot with a fixed camera configuration. Assuming an ego coordinate system $\boldsymbol{\mathcal{F}}_{\mathcal{E}}$ $(x \rightarrow, y \nearrow, z \odot)$ defined at the ground level, and camera coordinate system $\boldsymbol{\mathcal{F}}_{\mathcal{C}}$ $(x \rightarrow, y \otimes, z \nearrow)$, we consider a camera system parameterized by intrinsic matrix $\mathbf{K}$ and extrinsic matrix ${}^{\mathcal{C}}_{\mathcal{E}}[\mathbf{R}|\mathbf{t}]$, where:

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \quad {}^{\mathcal{C}}_{\mathcal{E}}\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sin\theta & -\cos\theta \\ 0 & \cos\theta & \sin\theta \end{bmatrix}, \quad (1)$$

and $\mathbf{t} = [0, 0, -h]^T$, with (θ, h) representing camera pitch and height, respectively. Transformation of points from ego to camera coordinate frame is then: $\mathbf{p}_{\mathcal{C}} = {}^{\mathcal{C}}_{\mathcal{E}}\mathbf{R}\mathbf{p}_{\mathcal{E}} + \mathbf{t}$. We disregard the camera yaw and roll in our parameterization; yaw has no effect on our canonical transformations, while roll is typically negligible in real-world setups.

Problem Statement. We aim to learn a dense mapping $f : (\mathbf{I}, \boldsymbol{\mathcal{C}}) \mapsto \mathbf{D}$, with $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}, \mathbf{D} \in \mathbb{R}^{H \times W}$ being RGB image and a correspondending depth map, and $\boldsymbol{\mathcal{C}}$ representing camera parameterization in Eq. (1). Most importantly, the learned mapping should be able to generalize to images acquired with arbitrary camera system. In contrast to existing zero-shot methods [3, 12, 27, 40], which rely on multiple datasets with varying camera parameters, we aim to learn this mapping using data acquired with a single camera system $\boldsymbol{\mathcal{C}}_{\text{Train}}$.

3.1. Depth Cues

Monocular depth estimation relies on semantic and geometric cues in the training data to solve the depth ambiguity of monocular perspective geometry. As illustrated in Fig. 2, one of the possible cues to regress depth d is the relation between object's real-world size S and imaging size s:

$$d = f_y \frac{S}{s}. \quad (2)$$

However, this cue leads to entanglement between focal length and depth, leading to generalization issues when

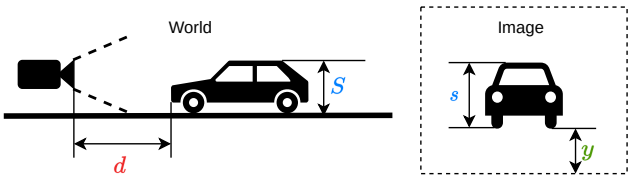


Figure 2. **Depth cues in MDE for ground vehicles**. Depth d can be calculated both as a function of imaging size s (with corresponding known object size S) and as a function of vertical image position y.

camera parameters change. Recent zero-shot methods address this issue by embedding [12], estimating [3, 27] or disentangling the focal length [40]. A major drawback of this cue is its reliance on the object's real world size S, which is highly dependent on the particular environmental domain.

For ground vehicles, a fixed position of the camera system relative to the ground introduces an additional perspective geometry constraint – known depth of the ground plane, which is constant for all images acquired by the same camera setup. For each pixel (u, v), depth d is calculated from the intersection of the camera rays $d\,\mathbf{K}^{-1}[u, v, 1]^T$ and the ground plane $({}^{\mathcal{C}}_{\mathcal{E}}\mathbf{R}\mathbf{n}_{\mathcal{E}})^T\mathbf{p}_{\mathcal{C}} + h = 0$, with $\mathbf{n}_{\mathcal{E}} = [0, 0, 1]^T$ being the upward ground normal. This constraint enables depth regression via object's ground contact point:

$$\begin{aligned} d &= \frac{-h}{({}^{\mathcal{C}}_{\mathcal{E}}\mathbf{R}\mathbf{n}_{\mathcal{E}})^T\mathbf{K}^{-1}[u, v, 1]^T} \\ &= \frac{f_y h}{(H - c_y - y)\cos(\theta) - f_y \sin(\theta)}, \end{aligned} \quad (3)$$

where $y = H - v$ represents vertical image position of the ground projection, as visualized in Fig. 2. Compared to the object size cue, which is entirely semantic, this cue implicitly encodes the known perspective geometry constraints of the scene, thereby inducing a valuable prior to the ill-posed MDE problem. *However, it is highly reliant on accurate and consistent camera calibration.*

3.2. Canonical Transforms

Metric3D [40] proposed a canonical transform disentangling the predicted depth and camera parameters, enabling the use of object size cue for arbitrary camera systems. To achieve this, depth is predicted in canonical space $\mathbf{C}_{\mathbf{F}}$ with a canonical focal length f_c, and post-processed to metric depth $\mathbf{D}_{\mathbf{F}}$ by applying Focal Canonical Transform $\text{FCT}_{\boldsymbol{\mathcal{C}}}(\cdot)$, where $\mathbf{D}_{\mathbf{F}} = \text{FCT}_{\boldsymbol{\mathcal{C}}}(\mathbf{C}_{\mathbf{F}}) = \frac{f_y}{f_c}\mathbf{C}_{\mathbf{F}}$, and $\boldsymbol{\mathcal{C}} = \{f_y\}$.

Besides the object size cue, in this paper we propose to leverage the vertical image position cue, effectively encoding the configuration of the scene's perspective geometry. Unfortunately, this cue also induces entanglement between camera parameters and depth. To enhance cross-dataset generalization, we introduce an additional canoni-

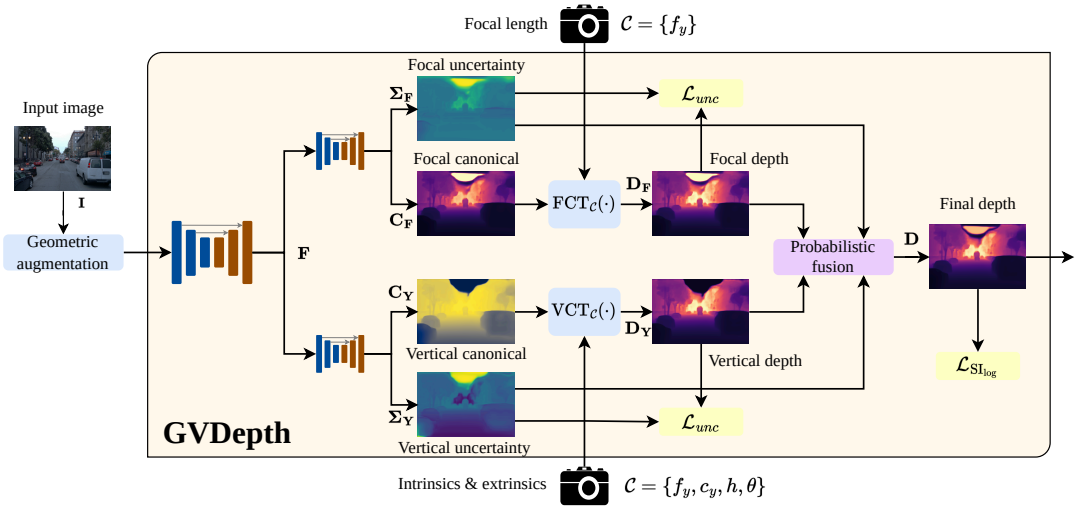


Figure 3. **Model architecture**. Given the input image and corresponding camera parameters, our model predicts two canonical depth representations, $(\mathbf{C_F}, \mathbf{C_Y})$ and corresponding uncertanties $(\mathbf{\Sigma_F}, \mathbf{\Sigma_Y})$. Canonical representations are disentangled from the corresponding camera parameters, enabling separate exploitation of object size cue and vertical image position cue. Canonical representations are then transformed into depth maps $(\mathbf{D_F}, \mathbf{D_Y})$, which are subsequently adaptively fused into a final depth map $\mathbf{D}$ based on estimated uncertanties.

cal transform, which enables consistent exploitation of the ground plane prior. Under this formulation, the model estimates a dense map $\mathbf{C_Y}$ representing vertical image position of the ground projection point. Metric depth $\mathbf{D_Y}$ is then obtained via Vertical Canonical Transform VCT$_\mathcal{C}(\cdot)$, where VCT$_\mathcal{C}(\mathbf{C_Y})$ corresponds to mapping $y \mapsto d$ obtainable from Eq. (3). Therefore, this transformation requires knowledge of camera parameters $\mathcal{C} = \{f_y, c_y, h, \theta\}$. Crucially, *the proposed canonical space remains consistent under variations of camera parameters, facilitating generalization to diverse scenarios.*

The proposed VCT$_\mathcal{C}(\cdot)$ is related to the depth of the ground plane, frequently used to enhance MDE accuracy in autonomous driving scenarios [4, 33, 37]. However, these methods exploit the ground-plane constraint exclusively for road pixels, resulting in limited generalization capability.

3.3. Model Architecture

Both cues have their advantages and drawbacks. For example, vertical position cue is highly consistent across environments due to the reduced reliance on semantic information. However, it is highly sensitive to calibration errors, and performs worse as depth increases. An ideal model should leverage both cues, taking into account the confidence for respective image regions. To that end, as visualized in Fig. 3, we design our model to predict depth maps with both cues separately, and then fuse them into a final depth map based on the estimated aleatoric uncertainty.

We use a standard encoder-decoder model with skip connections to process and upsample image features, gaining $\mathbf{F} \in \mathbb{R}^{h \times w \times C}$, where $(h, w) = (\frac{H}{2}, \frac{W}{2})$. We forward these features to two very small UNet networks [29], producing our canonical representations $(\mathbf{C_F}, \mathbf{C_Y})$ and accompanying uncertainty estimates $(\mathbf{\Sigma_F}, \mathbf{\Sigma_Y})$. Canonical representations are forwarded to FCT$(\cdot)$ and VCT$(\cdot)$, obtaining $(\mathbf{D_F}, \mathbf{D_Y})$, which are then fused into a final depth map $\mathbf{D} \in \mathbb{R}^{h \times w}$:

$$\mathbf{D} = \frac{\mathbf{\Sigma_Y} \odot \mathbf{D_F} + \mathbf{\Sigma_F} \odot \mathbf{D_Y}}{\mathbf{\Sigma_Y} + \mathbf{\Sigma_F}}. \tag{4}$$

3.4. Geometric Augmentation

Contrary to zero-shot methods trained on multiple datasets [3, 12, 27, 40], we aim to achieve similar generalization capabilities by training on a single camera system. However, without geometric diversity induced by varying camera parameters, model can simply disregard our proposed canonical transformations. Therefore, we apply geometric augmentations, namely random cropping and resizing, to simulate different camera perspectives. Crucially, compared to [3, 27, 40], we train our model with multiple resolutions, enabling generalization for arbitrary image sizes.

3.5. Optimization

Our main training objective is to guide the fused depth map $\mathbf{D}$ towards the ground-truth depth $\mathbf{D}^*$. Given the distance in the log domain $\mathbf{E} = \log(\mathbf{D}) - \log(\mathbf{D}^*)$, the SI$_{\log}$ loss [8] is defined as:

$$\mathcal{L}_{\text{SI}_{\log}} = \alpha \sqrt{\mathbb{V}[\mathbf{E}] + \lambda \mathbb{E}^2[\mathbf{E}]}, \tag{5}$$

with $\alpha = 10$ and $\lambda = 0.15$ as in [1], and $\mathbb{V}$, $\mathbb{E}$ being variance and expectation, respectively.

Additionally, we enhance depth estimation accuracy from both canonical transformations by employing an uncertainty-weighted loss:

$$\mathcal{L}_{\text{unc}} = \boldsymbol{\Sigma}_{\mathbf{F}}{}^{-1} \odot ||\mathbf{D}_{\mathbf{F}} - \mathbf{D}^*||_1 + \boldsymbol{\Sigma}_{\mathbf{Y}}{}^{-1} \odot ||\mathbf{D}_{\mathbf{Y}} - \mathbf{D}^*||_1 + \log(\boldsymbol{\Sigma}_{\mathbf{F}}) + \log(\boldsymbol{\Sigma}_{\mathbf{Y}}), \tag{6}$$

which is equivalent to maximizing the log-likelihood, but with L1 norm instead of L2. This approach down-weights regions where each cue exhibits higher uncertainty, encouraging each cue to focus on specific parts of the image where it performs better. To facilitate training stability [15], we predict $\log(\cdot)$ uncertainties, and perform $\exp(\cdot)$ operation accordingly.

Finally, same as in [27], we employ geometric consistency loss. Given two different sets of geometric augmentations of the same image ($\mathcal{T}_1$, $\mathcal{T}_2$), geometric consistency loss is defined as:

$$\mathcal{L}_{\text{con}} = \left\| \mathcal{T}_2 \circ \mathcal{T}_1^{-1} \circ \mathbf{D}_1 - \text{sg}\left(\mathbf{D}_2\right) \right\|_1, \tag{7}$$

where $\mathbf{D}_i$ are estimated depth maps for i'th augmentation, and $\text{sg}(\cdot)$ corresponds to the stop-gradient operation.

Our final loss is a linear combination:

$$\mathcal{L} = \mathcal{L}_{\text{SI}_{\log}} + \lambda_{\text{unc}} \mathcal{L}_{\text{unc}} + \lambda_{\text{con}} \mathcal{L}_{\text{con}}, \tag{8}$$

with $\lambda_{\text{unc}} = 0.5$ and $\lambda_{\text{con}} = 0.1$.

4. Experiments

Datasets. We use a diverse set of autonomous driving datasets, each collected with a different camera setup. Specifically, we utilize an aggregation of KITTI [9], DDAD [11], DrivingStereo [35], Waymo [32] and Argoverse Stereo [5] datasets. Due to the missing and sometimes inaccurate information, we calibrate extrinsic parameters, i.e., camera height and pitch, which are required for our canonical representation. Further details are provided in the supplementary materials.

Implementation Details. We use a standard encoder-decoder with skip connections [10]. In all experiments we use a ConvNext-Large [20] encoder initialized with ImageNet weights. To provide the model with additional contextual information, we fuse ground plane embedding in the decoder layers [17, 37]. In our depth cue fusion model, secondary U-Nets are designed with a lightweight structure, containing only three convolutional layers for both downsampling and upsampling.

Our models are trained in PyTorch [23] with an AdamW optimizer [22]. Learning rate is progressively reduced with Cosine Annealing from 1e-4 to 1e-5. For all datasets, we use random cropping and resizing to $H \in [200, 500]$, with W chosen to preserve the aspect ratio after cropping. To increase resolution diversity in a single batch update, we perform gradient accumulation for 2 training steps. All models are trained for 150000 optimizer steps with an effective batch size of 16, on a single NVIDIA RTX A6000 GPU. Compared to concurrent zero-shot methods [12, 27, 40], this constitutes a fairly simple training setup, both in computational and data resources, demonstrating the efficiency of our approach.

Evaluation Details. We use established depth metrics: absolute relative error (A.Rel), root mean squared error (RMS), root mean squared log error ($\text{RMS}_{\log}$) and accuracy δ_i under a threshold 1.25^i. Depth is evaluated in the metric range of [0 m, 80 m].

We evaluate GVDepth at an image resolution of $H \times 640$, where H varies according to the dataset to preserve the original aspect ratio. For a fair comparison, domain-specific models are trained and tested using the same image dimensions, which depend on specific dataset. All models are evaluated using a ConvNext-L backbone if available; otherwise, we select the backbone with the most similar complexity. For zero-shot methods, we use the provided checkpoints. For metric zero-shot models which are optimized for a specific training resolution, we report results at both their original training resolution and our testing resolution for a comprehensive evaluation.

4.1. Evaluation of Zero-shot Accuracy

Comparison with State-of-the-Art. We perform extensive evaluation, comparing GVDepth with SotA domain-dependent and domain-agnostic methods. Results for KITTI and DDAD datasets are shown in Tab. 1. Both supervised [26, 41] and self-supervised [10, 44] domain-dependent methods perform well on datasets matching their training domain but overfit to environmental and geometric biases, leading to reduced accuracy in KITTI $\rightarrow$ DDAD transfer. Relative depth estimation models exhibit greater adaptability across domains; however, they require ground-truth scale and shift alignment, limiting their practical applicability. Recent zero-shot metric methods, such as ZeroDepth, UniDepth, and Metric3D, demonstrate strong generalization capabilities but rely on extensive data acquisition and large-scale training, which is not always feasible. Furthermore, UniDepth and Metric3D suffer accuracy degradation when inferring on resolutions lower than their high-resolution training data. While resizing and padding can address this, these adjustments increase computational demands, limiting their suitability for resource-constrained systems. GVDepth delivers comparable results while being adaptable to arbitrary camera setups, aspect ratios, and image resolutions. Depending on the train-test dataset combination, our method ranks first or second across most metrics, despite training on only a single dataset and using a fraction of the data required by other zero-shot MDE methods. This is particularly impressive, as each dataset features unique environments and camera setups, demonstrating that our model relies on domain-agnostic perspective geometry cues rather than domain-specific semantic cues.

Table 1. **Generalization capability comparison.** Domain-dependent methods [10, 26, 41, 44] are trained and evaluated on our evaluation setup at an $H \times 640$ resolution. Zero-shot methods [12, 27, 28, 39, 40] are evaluated with provided checkpoints. (†): Shift and scale alignment with ground-truth. (‡): Resizing and padding to native training resolution. Best zero-shot results without resolution alignment are **bolded**, second best are underlined. In-domain evaluation results are shaded.

Method	Train dataset	# Train samples	KITTI						DDAD					
			A.Rel ↓	RMS ↓	RMS_{log} ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	A.Rel ↓	RMS ↓	RMS_{log} ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Monodepth2 [10]	KITTI	23K	9.02	3.94	0.137	91.4	98.3	99.5	37.8	14.6	0.599	20.1	45.7	74.0
DIFFNet [44]			8.15	3.86	0.129	92.0	98.4	99.6	37.3	14.4	0.563	16.1	42.6	79.1
NeWCRFs [41]			5.51	2.20	0.084	97.3	99.7	99.9	33.7	15.6	0.571	30.2	62.9	79.2
iDisc [26]			5.33	2.11	0.081	97.1	99.7	99.9	27.4	13.3	0.461	38.4	74.3	88.4
MiDaS† [28]	Mix	2M	13.2	5.33	0.199	82.5	95.5	98.5	13.7	8.01	0.319	83.2	95.0	98.1
LeReS† [39]		354K	14.1	4.89	0.190	81.5	95.4	98.3	17.0	7.04	0.331	76.1	93.9	97.6
ZeroDepth [12]	Mix	17M	11.1	3.65	0.154	87.9	96.8	98.9	8.62	5.83	0.156	91.9	97.4	98.8
UniDepth [27]		3M	13.6	**3.13**	0.149	88.0	**99.3**	**99.9**	13.0	**6.62**	**0.191**	84.4	**97.1**	**98.9**
Metric3D [40]		8M	7.91	3.34	0.118	**93.0**	98.9	99.7	12.0	6.52	0.178	87.7	94.5	98.1
GVDepth	KITTI	23K	5.67	2.61	0.090	95.7	99.4	99.9	**11.8**	8.35	0.251	82.5	93.4	96.5
	DDAD	13K	12.3	4.38	0.187	86.0	96.3	98.9	9.67	5.62	0.162	90.3	97.2	98.9
	Argoverse	5K	11.2	4.43	0.181	86.1	97.1	99.4	13.5	8.38	0.240	82.6	93.7	96.9
	Waymo	156K	10.9	3.23	0.141	89.8	98.4	99.6	13.9	7.68	0.259	**86.2**	95.2	97.2
	DrivingStereo	168K	**6.96**	3.17	**0.110**	92.7	98.6	99.7	12.0	7.72	0.215	82.4	93.6	97.6
UniDepth‡ [27]	Mix	3M	5.61	2.34	0.084	97.0	99.6	99.9	15.1	5.82	0.180	89.3	97.7	99.0
Metric3D‡ [40]		8M	5.41	2.25	0.085	97.2	99.5	99.9	10.6	5.72	0.164	92.4	97.6	98.8

Table 2. **Comparison with GEDepth [37].** Vertical – depth regression via Vertical Canonical Transform - $VCT_{\mathcal{C}}(\cdot)$. Fusion – depth regression with model as designed in Fig. 3. All models are trained with equivalent setup and model complexity on KITTI.

	Testing dataset	GEDepth[37]		Vertical		Fusion	
		A.Rel ↓	δ_1 ↑	A.Rel ↓	δ_1 ↑	A.Rel ↓	δ_1 ↑
KITTI	KITTI	5.68	95.4	5.70	95.5	**5.67**	**95.7**
	DStereo	18.25	66.7	10.43	87.3	**10.24**	**87.4**
	Waymo	22.11	69.1	**13.42**	**78.6**	14.34	78.4
	Argo	19.27	52.8	14.72	64.7	**10.45**	**84.8**
	DDAD	19.46	61.2	14.26	73.6	**11.81**	**82.5**

Comparison with Ground Plane Methods. Our proposed $VCT_{\mathcal{C}}(\cdot)$ implicitly encodes depth of the ground plane, which has been utilized in [4, 33, 37] to improve MDE accuracy in autonomous driving scenes. However, as demonstrated in Tab. 2, $VCT_{\mathcal{C}}(\cdot)$ significantly outperforms GEDepth [37], with the performance gap further widened after fusion with $FCT_{\mathcal{C}}(\cdot)$. As we will show, the primary cause of such discrepency is inability of GEDepth to leverage vertical image positions and disentangle perspective geometry for pixels located outside the road surface. In the Supp. materials we provide a comprehensive analysis and critique of existing models that rely on ground plane depth.

Qualitative Results. In Fig. 4 we illustrate qualitative results of our DrivingStereo → KITTI evaluation. Notably, it can be seen that vertical uncertainty $\boldsymbol{\Sigma_Y}$ is lower in closer regions, and decreasing with higher distances. An interesting detail appears in the left image, where the vertical uncertainty adapts to the scene's leftward slope, indicating that the model is aware of ground perturbations. Additionally, in the right image, the vertical uncertainty is significantly higher for a car lacking a visible ground-contact point, creating ambiguity in its vertical image position cue.

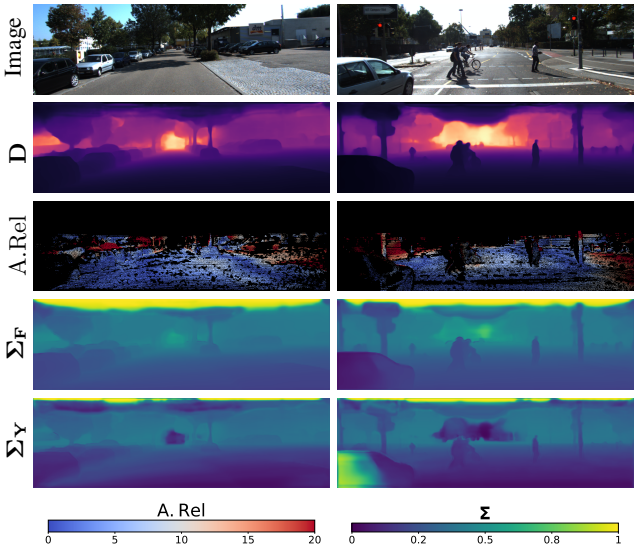


Figure 4. **Qualitative results.** Predicted depths, errors and uncertanties for DrivingStereo → KITTI evaluation.

4.2. Ablation studies

Evaluation of Canonical Representations. In Tab. 3 we report extensive evaluation results, spanning different representations and train-test dataset combinations. Moreover, in Supplementary we report qualitative results accompanying the table. By carefully examining the results, we can draw the following conclusions: i) **The baseline model struggles to transfer effectively to out-of-distribution data.** Without the invariance introduced by canonical representa-

Table 3. **Ablation of canonical representations.** Baseline – standard depth regression with geometric augmentations. Focal – depth regression via Focal Canonical Transform - $\text{FCT}_{\mathcal{C}}(\cdot)$. This is equivalent to Metric3D[40], but trained on our setup. Vertical – depth regression via Vertical Canonical Transform - $\text{VCT}_{\mathcal{C}}(\cdot)$. Fusion – depth regression with model as designed in Fig. 3. Training datasets on rows, testing datasets on columns. Best results are **bolded**, second best are underlined. In-domain evaluation results are shaded. **Note:** Focal and vertical results presented here are from separate models leveraging $\text{FCT}_{\mathcal{C}}(\cdot)$ and $\text{VCT}_{\mathcal{C}}(\cdot)$, not from submodels in Fig. 3.

	Representation	KITTI			DrivingStereo			Waymo			Argoverse			DDAD		
		A.Rel ↓	RMS ↓	δ_1 ↑	A.Rel ↓	RMS ↓	δ_1 ↑	A.Rel ↓	RMS ↓	δ_1 ↑	A.Rel ↓	RMS ↓	δ_1 ↑	A.Rel ↓	RMS ↓	δ_1 ↑
KITTI	Baseline	6.01	2.65	95.2	25.60	8.31	23.0	31.61	10.86	11.0	46.28	15.33	6.0	16.12	8.72	76.3
	Focal	5.78	2.78	94.8	12.01	5.77	86.1	19.12	8.30	72.5	10.51	**4.86**	**86.6**	14.15	8.41	80.0
	Vertical	5.70	**2.58**	95.5	10.43	**5.42**	87.3	**13.42**	**7.46**	**78.6**	14.72	7.11	64.7	14.26	9.39	73.6
	Fusion	**5.67**	2.61	**95.7**	**10.24**	5.66	**87.4**	14.34	7.56	78.4	**10.45**	5.06	84.8	**11.81**	**8.35**	**82.5**
DStereo	Baseline	35.62	5.56	20.1	4.82	2.76	97.9	20.21	9.02	65.1	20.62	8.51	50.7	25.32	8.10	47.7
	Focal	8.47	3.74	89.8	3.05	1.76	99.5	16.61	8.04	74.2	9.94	4.95	86.8	14.92	7.86	82.0
	Vertical	7.42	3.33	92.6	3.07	**1.75**	99.5	**11.82**	**6.36**	**85.5**	11.41	6.56	84.7	14.71	8.42	80.0
	Fusion	**6.96**	**3.17**	**92.7**	**3.01**	1.76	**99.5**	12.15	6.50	83.1	**9.91**	**4.92**	**86.9**	**12.02**	**7.72**	**82.4**
Waymo	Baseline	18.32	5.65	84.0	22.50	5.98	64.9	4.43	2.86	96.8	28.74	9.51	10.9	24.33	8.92	55.1
	Focal	18.21	5.87	80.2	13.81	5.27	85.5	3.62	**2.46**	98.6	18.41	5.87	85.3	20.41	8.75	66.0
	Vertical	**8.30**	**3.17**	**92.3**	**11.40**	**5.25**	86.8	3.61	2.48	**98.9**	12.21	5.12	88.6	**11.51**	**7.34**	84.4
	Fusion	10.92	3.23	89.8	12.78	5.68	**87.1**	**3.51**	2.57	98.8	**9.62**	**4.74**	**94.1**	13.99	7.68	**86.2**
Argo	Baseline	72.31	9.91	7.6	46.32	9.99	20.2	34.41	8.71	34.8	4.72	2.45	97.6	41.92	9.35	25.4
	Focal	13.16	6.41	80.2	16.61	10.21	74.6	21.13	9.40	60.6	3.16	1.81	98.6	18.21	10.51	62.6
	Vertical	14.60	5.81	84.2	18.22	8.10	75.6	14.98	**7.38**	72.6	**2.95**	**1.77**	98.6	13.86	8.88	81.5
	Fusion	**11.21**	**4.43**	**86.1**	**12.81**	**7.81**	**82.1**	**12.92**	7.38	**78.4**	3.14	1.82	**98.7**	**13.48**	**8.38**	**82.6**
DDAD	Baseline	59.42	10.17	5.6	18.61	6.72	77.0	26.61	8.40	30.9	31.81	9.98	15.1	10.98	5.86	88.2
	Focal	12.45	**4.07**	86.0	14.31	**5.64**	83.1	21.91	7.60	53.9	15.01	6.16	83.1	10.91	5.63	88.7
	Vertical	22.21	6.17	72.5	17.00	6.45	79.1	**15.51**	**6.27**	**80.6**	14.10	5.96	83.1	10.20	5.73	89.1
	Fusion	**12.31**	4.38	**86.0**	**14.21**	6.31	**83.8**	16.61	7.36	77.6	**12.66**	**5.76**	**84.9**	**9.67**	**5.62**	**90.3**

tions, baseline models fail to leverage diverse perspective geometries in the training data due to the ambiguities induced by depth-camera parameters entanglement; (ii) **Vertical image position cues generalize better than object size cues.** Models with our novel $\text{VCT}_{\mathcal{C}}(\cdot)$ representation generally outperform $\text{FCT}_{\mathcal{C}}(\cdot)$ models, except in rare cases, potentially stemming from calibration errors. This is particularly evident with models trained on the Waymo dataset, which encounter difficulties when relying solely on object size cues. The probable cause is that Waymo's cameras are positioned significantly higher, causing a variation of apparent object sizes due to the change of perspective. In contrast, vertical image position cues are more robust, implicitly encoding camera height within the representation; (iii) **Probabilistic fusion improves generalization by integrating both cues.** Our fusion strategy achieves the most accurate predictions by adaptively weighting the more reliable cue. For example, when object size cues are more accurate (*e.g.* DDAD → KITTI), the model adjusts accordingly, and the same applies for vertical position cues (*e.g.* Waymo → Argoverse).

Resolution Adaptability. SoTA zero-shot methods such as UniDepth [27] and Metric3D [40] require resizing and padding to training resolution to achieve optimal accuracy. However, when applied to datasets like KITTI, this approach results in the model processing nearly twice as many pixels as necessary due to padding. This significantly increases computational complexity without adding any meaningful detail to the image.

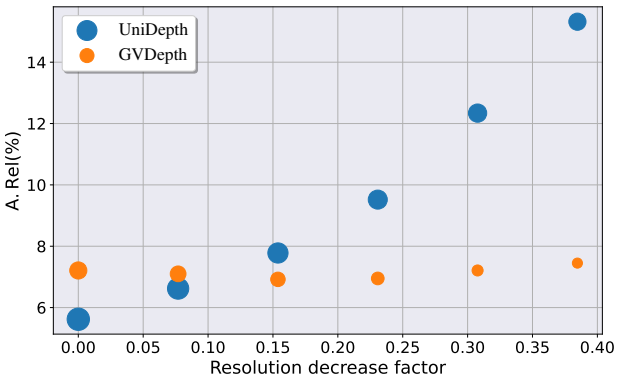


Figure 5. **Resolution adaptability.** A.Rel(%) of UniDepth [27] and GVDepth as height resolution H decreases from base resolution 462. Computational complexity is depicted with different circle radius. GVDepth results are from DrivingStereo → KITTI evaluation.

In contrast, GVDepth is designed to be fully resolution-agnostic, adapting seamlessly to the native resolution of the input image. As shown in Fig. 5, the error rate of GVDepth remains largely unaffected even as the input resolution decreases. This feature is especially advantageous for real-time systems, where dynamically adjusting image resolution is one of the simplest and most effective ways to control computational complexity.

Surround-view Generalization. In Tab. 4 we provide results for Waymo → DDAD zero-shot transfer for surround-

Table 4. **Surround-view generalization.** Vertical – model with $\mathrm{VCT}_{\mathcal{C}}(\cdot)$. Focal – model with $\mathrm{FCT}_{\mathcal{C}}(\cdot)$. All models evaluated on DDAD. $\mathrm{VCT}_{\mathcal{C}}(\cdot)$ and $\mathrm{FCT}_{\mathcal{C}}(\cdot)$ models trained solely on Waymo. (†): Waymo surround images added to training data. UniDepth and Metric3D evaluated with original checkpoints and resolutions.

Method	Front-left		Front-left		Back	
	A.Rel ↓	δ_1 ↑	A.Rel ↓	δ_1 ↑	A.Rel ↓	δ_1 ↑
Baseline	67.42	18.2	78.48	12.3	33.58	27.1
Focal	23.37	62.2	22.70	67.1	<u>20.93</u>	<u>67.7</u>
Vertical	<u>19.21</u>	<u>71.8</u>	<u>17.22</u>	**73.5**	22.12	66.2
Focal†	20.69	68.2	20.01	66.7	**19.60**	**68.6**
Vertical†	**16.44**	**76.5**	**16.82**	<u>71.2</u>	21.35	65.8
UniDepth† [27]	72.28	8.9	71.40	8.2	55.51	11.1
Metric3D† [40]	14.57	80.7	13.86	80.1	10.55	88.7

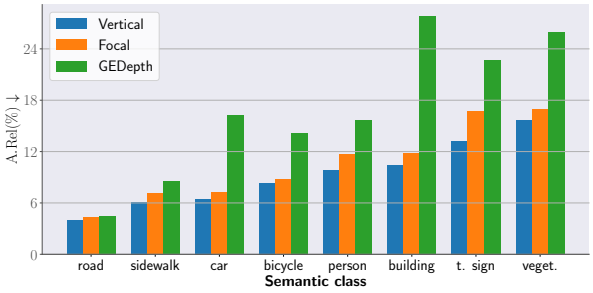


Figure 6. **Error per object class.** Vertical – model with $\mathrm{VCT}_{\mathcal{C}}(\cdot)$. Focal – model with $\mathrm{FCT}_{\mathcal{C}}(\cdot)$. All models are evaluated on DrivingStereo → KITTI zero-shot transfer.

view cameras. Despite the increased ambiguity and perturbations in ground regions captured by side views, our $\mathrm{VCT}(\cdot)$ model outperforms $\mathrm{FCT}(\cdot)$ on side cameras, with performance increasing after Waymo side cameras are included in training data. $\mathrm{FCT}(\cdot)$ performs better on back camera, possibly due to the occlusion of bottom part of the image by ego-vehicle. For unknown reasons, UniDepth [27] failed to infer correct camera parameters for DDAD surround cameras, resulting in subpar performance.

Generalization in Specific Regions. The strong quantitative generalization results of vertical position cues do not fully capture their performance. For instance, a model with $\mathrm{VCT}_{\mathcal{C}}(\cdot)$ might only perform better on road pixels due to its implicitly encoded planarity. However, as shown in Fig. 6, this is not the case: the $\mathrm{VCT}_{\mathcal{C}}(\cdot)$ model exhibits accurate generalization across a wide range of object classes. The results clearly reveal major flaws in GEDepth's [37] performance, as it struggles to generalize beyond road-related pixels, particularly for safety-critical objects.

Moreover, we examine the effects of ground plane inconsistencies and perturbations on the depth estimation accuracy. A reasonable assumption would be that the $\mathrm{VCT}_{\mathcal{C}}(\cdot)$ model is more sensitive to these perturbations. However, Fig. 7a illustrates that error for both models increases in similar pattern as ground perturbation increases. This indicates that our $\mathrm{VCT}_{\mathcal{C}}(\cdot)$ model can adapt and recognize perturbations which are present in the training data, atleast to the similar degree as competing methods.

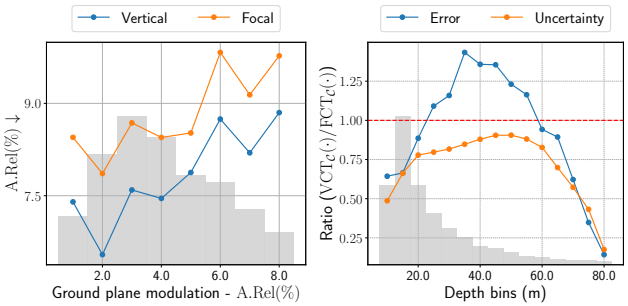


Figure 7. **Sensitivity analysis.** All models are evaluated on DrivingStereo → KITTI. Histograms represent data distribution. **a)** A.Rel of $\mathrm{VCT}_{\mathcal{C}}(\cdot)$ and $\mathrm{FCT}_{\mathcal{C}}(\cdot)$ models depending on ground plane modulation severity. Ground plane modulation is expressed as A.Rel of ground depth calculated from camera parameters via Eq. (3) and ground-truth depth of the road. **b)** Ratio of depth errors $(\mathrm{A.Rel}(\mathbf{Z_Y})/\mathrm{A.Rel}(\mathbf{Z_F}))$ and uncertainties $(\mathbf{\Sigma}_Y/\mathbf{\Sigma}_Z)$ depending on depth bin.

Finally, in Fig. 7b we analyze the errors and uncertainties of both cues across discrete depth bins. As shown, the error ratio skews towards $\mathrm{VCT}_{\mathcal{C}}(\cdot)$ in regions closer to the camera, which is expected. As depth increases, the object size cue becomes more accurate due to the reduced reliability of the vertical position cue at mid-range distances. However, for farther distances, $\mathrm{VCT}_{\mathcal{C}}(\cdot)$ once again achieves lower error. We hypothesize that this is because $\mathrm{VCT}_{\mathcal{C}}(\cdot)$ implicitly encodes the horizon level, aiding depth estimation for distant objects. Additionally, the estimated uncertainties follow a similar pattern, demonstrating the effectiveness of our adaptive fusion module.

5. Conclusion and Future Work

We present GVDepth, a novel MDE model designed for accurate zero-shot transfer across diverse environments and camera setups, with a focus on autonomous vehicles and mobile robotics. Leveraging the fixed camera-to-ground geometry in such platforms, we introduce a *Vertical Canonical Representation* that uses known ground depth to enforce perspective invariance, resulting in enhanced learning and generalization. Our architecture further adaptively fuses depth estimates from object size and vertical image position cues, guided by uncertainty.

Extensive ablations across five autonomous driving datasets show that our method generalizes significantly better than object-size-based baselines, matching or surpassing state-of-the-art zero-shot models, while using only a single training dataset and far fewer computational and data resources. Future work will explore scaling to multi-dataset training toward a foundation model for MDE in autonomous driving and mobile robotics.

Acknowledgment. This research has been funded by the H2020 project AIFORS under Grant Agreement No 952275.

References

[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 1, 2, 4

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 2

[3] 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. *arXiv preprint arXiv:2410.02073*, 2024. 1, 2, 3, 4

[4] Aurélien Cecille, Stefan Duffner, Franck Davoine, Thibault Neveu, and Rémi Agier. Groco: Ground constraint for metric self-supervised monocular depth. In *European Conference on Computer Vision (ECCV)*, 2024. 3, 4, 6

[5] Ming-Fang Chang, John W Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[6] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29, 2016. 1, 2

[7] Tom van Dijk and Guido de Croon. How do neural networks see depth in single images? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2183–2191, 2019. 2, 3

[8] 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. 1, 2, 4

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 2, 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*, pages 3828–3838, 2019. 1, 5, 6

[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. 1, 5

[12] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9233–9243, 2023. 1, 2, 3, 4, 5, 6

[13] Hyungjoo Jung, Youngjung Kim, Dongbo Min, Changjae Oh, and Kwanghoon Sohn. Depth prediction from a single image with conditional adversarial networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1717–1721. IEEE, 2017. 2

[14] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 1, 2

[15] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 5

[16] Karlo Koledić, Igor Cvišić, Ivan Marković, and Ivan Petrović. Moft: Monocular odometry based on deep depth and careful feature selection and tracking. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6175–6181. IEEE, 2023. 2

[17] Karlo Koledić, Luka Petrović, Ivan Petrović, and Ivan Marković. Gendepth: Generalizing monocular depth estimation for arbitrary camera parameters via ground plane embedding. *arXiv preprint arXiv:2312.06021*, 2023. 2, 5

[18] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 1, 2

[19] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1119–1127, 2015. 2

[20] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 5

[21] Kin Gwn Lore, Kishore Reddy, Michael Giering, and Edgar A Bernal. Generative adversarial networks for depth map estimation from rgb video. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1258–12588. IEEE, 2018. 2

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 5

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[24] Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. P3depth: Monocular depth estimation with a piecewise planarity prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1610–1621, 2022. 2

[25] Rui Peng, Ronggang Wang, Yawen Lai, Luyang Tang, and Yangang Cai. Excavating the potential capacity of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15560–15569, 2021. 2

[26] Luigi Piccinelli, Christos Sakaridis, and Fisher Yu. idisc: Internal discretization for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21477–21487, 2023. 1, 2, 5, 6

[27] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[28] 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):1623–1637, 2020. 1, 2, 6

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[30] Shuwei Shao, Zhongcai Pei, Weihai Chen, Xingming Wu, and Zhengguo Li. Nddepth: Normal-distance assisted monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7931–7940, 2023. 2

[31] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 2

[32] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[33] Ruoyu Wang, Zehao Yu, and Shenghua Gao. Planedepth: Self-supervised depth estimation via orthogonal planes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21425–21434, 2023. 1, 3, 4, 6

[34] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 311–320, 2018. 1, 2

[35] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[36] 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. 1, 2

[37] Xiaodong Yang, Zhuang Ma, Zhiyu Ji, and Zhe Ren. Gedepth: Ground embedding for monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12719–12727, 2023. 3, 4, 5, 6, 8

[38] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*, 2020. 1, 2

[39] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 204–213, 2021. 1, 2, 6

[40] 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. 2, 3, 4, 5, 6, 7, 8

[41] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Newcrfs: Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 5, 6

[42] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 340–349, 2018. 1

[43] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *2022 International Conference on 3D Vision (3DV)*, pages 668–678. IEEE, 2022.

[44] Hang Zhou, David Greenwood, and Sarah Taylor. Self-supervised monocular depth estimation with internal feature fusion. In *British Machine Vision Conference (BMVC)*, 2021. 5, 6

[45] 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*, pages 1851–1858, 2017. 1