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

Supplementary Material

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

A. Additional Ablations with Ground Plane Methods

Additional Results. Table 6 presents a more detailed ablation study of GEDepth [25], incorporating results from additional training datasets. GVDepth consistently outperforms GEDepth across multiple training-test dataset combinations. We do not report results on Argoverse and DDAD datasets, as GEDepth failed to achieve satisfactory convergence. We will provide further comments on this issue below.

Furthermore, in Tab. 7, we provide results for PlaneDepth [22] and GroCo [1], which utilize the ground plane constraint in a similar manner as GVDepth. Since these methods are self-supervised, we retrained PlaneDepth with supervision, using the same backbone as our method. The results suggest that the performance gap is not due to difference in supervision or model complexity, but due to the superior out-of-distribution generalization of our approach. For GroCo [1], we report the results from the original paper, as it lacks the open-source code implementation. However, GroCo is fundamentally a self-supervised version of GEDepth [25], preserving all of its core principles.

Our main argument is that none of these methods are explicitly designed to enhance generalization or facilitate training with diverse perspective geometries. Moreover, they improve performance only for road pixels, unlike the proposed $\mathrm{VCT}_{\mathcal{C}}(\cdot)$ that exploits the ground plane constraint for all objects. In the following sections, we provide an in-depth analysis, offering insights into the generalization and convergence challenges faced by this methods.

Detailed Analysis. We start with the following premises, which differentiate the proposed $\mathrm{VCT}_{\mathcal{C}}(\cdot)$ from aforementioned approaches:

(i) Unlike GEDepth and GroCo, which utilize the ground plane constraint only for road pixels, the proposed canonical representation enables utilization of vertical image position cue for all objects in the scene;

(ii) Similarly, in constrast to GEDepth and GroCo, our approach is inherently modeled to enable perspective

Table 6. **Comparison with GEDepth [25].** All models are trained with equivalent setup and model complexity on KITTI, DrivingStereo and Waymo datasets. Best results are **bolded**, second best are underlined. In-domain evaluation results are shaded.

	Testing dataset	GEDepth[25] A.Rel ↓	GEDepth[25] δ_1 ↑	Vertical A.Rel ↓	Vertical δ_1 ↑	Fusion A.Rel ↓	Fusion 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**
Dstereo	KITTI	11.77	74.2	7.42	92.6	**6.96**	**92.7**
	DStereo	**2.99**	**99.5**	3.07	99.5	3.01	99.5
	Waymo	18.88	73.3	**11.72**	**85.5**	12.15	83.1
	Argo	13.19	83.0	11.41	84.7	**9.91**	**86.9**
	DDAD	20.20	55.18	15.71	80.0	**12.02**	**82.4**
Waymo	KITTI	15.91	82.1	**8.30**	**92.3**	10.92	89.8
	DStereo	14.21	84.4	**11.40**	86.8	12.78	**87.1**
	Waymo	**3.42**	**99.0**	3.61	98.9	3.51	98.8
	Argo	17.83	83.4	12.21	88.6	**9.62**	**94.1**
	DDAD	19.20	66.6	**11.51**	84.4	13.99	**86.2**

Table 7. **Comparison with self-supervised methods that leverage the known ground plane.** Results for KITTI → DDAD zero-shot transfer.(†): Trained with supervision. (‡): Results taken from corresponding paper.

Method	A.Rel ↓	RMS ↓	$\mathrm{RMS}_{\log}$ ↓	δ_1 ↑
PlaneDepth† [22]	30.28	12.01	0.499	37.8
PlaneDepth [22]	32.71	14.77	0.529	37.9
GroCo‡ [1]	42.40	15.37	-	-
Vertical†	14.26	9.39	0.271	73.6
Fusion†	**11.81**	**8.35**	**0.251**	**82.5**

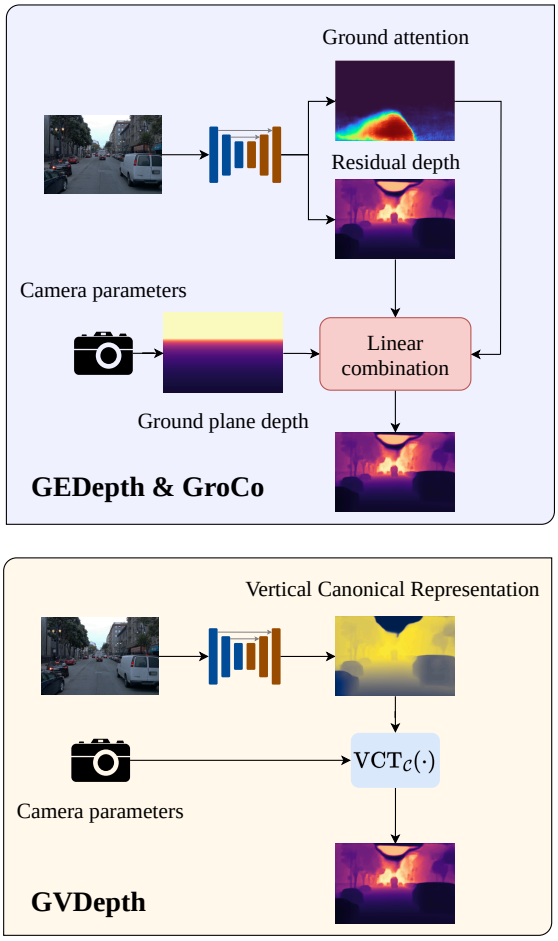

Figure 9. **Different approaches for incorporation of ground plane constraint.** In our approach, VCT$_{\mathcal{C}}(\cdot)$ serves as a *point of disentanglement*, which enables learning of *Vertical Canonical Representation* that is invariant to specific perspective geometry. Linear combination used in GEDepth and GroCo does not hold such properties, leading to limited generalization.

geometry disentanglement for all objects, resulting in better out-of-distribution generalization.

While our extensive evaluation confirms this, here we also provide insights into why GEDepth and GroCo fail to generalize across all objects in the scene.

Both GEDepth and GroCo formulate the problem of depth regression $\mathbf{D} \in [0, D_{max}]^{H \times W}$ as a linear combination of the known ground depth $\mathbf{D_G} \in [0, D_{max}]^{H \times W}$ and the learnable residual depth $\mathbf{D_R} \in [0, D_{max}]^{H \times W}$, weighted by the learnable *"ground attention"* $\mathbf{A} \in [0, 1]^{H \times W}$:

$$\mathbf{D} = \mathbf{A} \odot \mathbf{D_G} + (1 - \mathbf{A}) \odot \mathbf{D_R}. \quad (1)$$

Through a straightforward analysis, we can notice why this formulation enhances depth estimation for road pixels only. First, any part of the scene appearing above the horizon can not benefit from the induced perspective geometry constraint, as $\mathbf{D_G}$ is invalid for those regions of the image. Moreover, ground attention $\mathbf{A}$ is usually estimated with the $\mathrm{Sigmoid}(\cdot)$, which has a natural tendency to converge to either 0 or 1 during optimization, leading to $\mathbf{A}$ effectively being equivalent to road segmentation, as visualized in [1, 25]. Even for rare cases where this is not the case, the resulting linear combination has an ambivalent geometric interpretation chosen arbitrarily by the model. Our proposed VCT$_{\mathcal{C}}(\cdot)$ is a simple and elegant solution for all of these issues; it enables effective utilization of the ground plane constraint for all image regions while preserving a clear geometric interpretation.

The same fundamental principles can be applied to perspective geometry disentanglement. While $\mathbf{D_G}$ clearly disentangles depth from camera parameters, the estimated $\mathbf{D_R}$ does not hold such properties, inducing generalization errors due to the ambiguity of depth and camera parameters. Since the final depth $\mathbf{D}$ is calculated as a linear combination, it is only fully disentangled where $\mathbf{A}$ is 1, which is valid only for road pixels. On the other hand, VCT$_{\mathcal{C}}(\cdot)$ is inherently designed to incorporate perspective geometry disentanglement, regardless of the specific image region.

GEDepth Convergence Issues. GEDepth optimizes $\mathbf{D}$ from Eq. (1) without additional regularization factors or external segmentation modules, claiming that the learned attention map can automatically separate ground and other regions. Unfortunately, in our experiments this occurred rather inconsistently, and the learned attention map converged to adequate ground segmentation only on the 19th training experiment. In most cases, the model resorted to estimating $\mathbf{A} = \mathbf{0}^{H \times W}$, thus ignoring the provided ground plane information and resulting in $\mathbf{D} = \mathbf{D_R}$.

Final Remarks. We conducted a thorough evaluation that highlights the extremely limited generalization capabilities of models that rely on ground plane constraints [1, 22, 25]. While one might assume that these issues arise from limited reproducibility or incorrect convergence, the analysis we presented in previous sections suggests that the root cause lies in inherent design choices. These models are optimized to perform well only on narrow training distribution. Even within these constraints, improvements in accuracy are confined to road pixels, which are of limited relevance for safety-critical applications. In contrast, the proposed VCT$_{\mathcal{C}}(\cdot)$ elegantly addresses all these limitations.

B. Qualitative Results

Qualitative results for different model configurations are shown in Fig. 10. All models demonstrate comparable capabilities in reconstructing semantic objects, effectively cap-

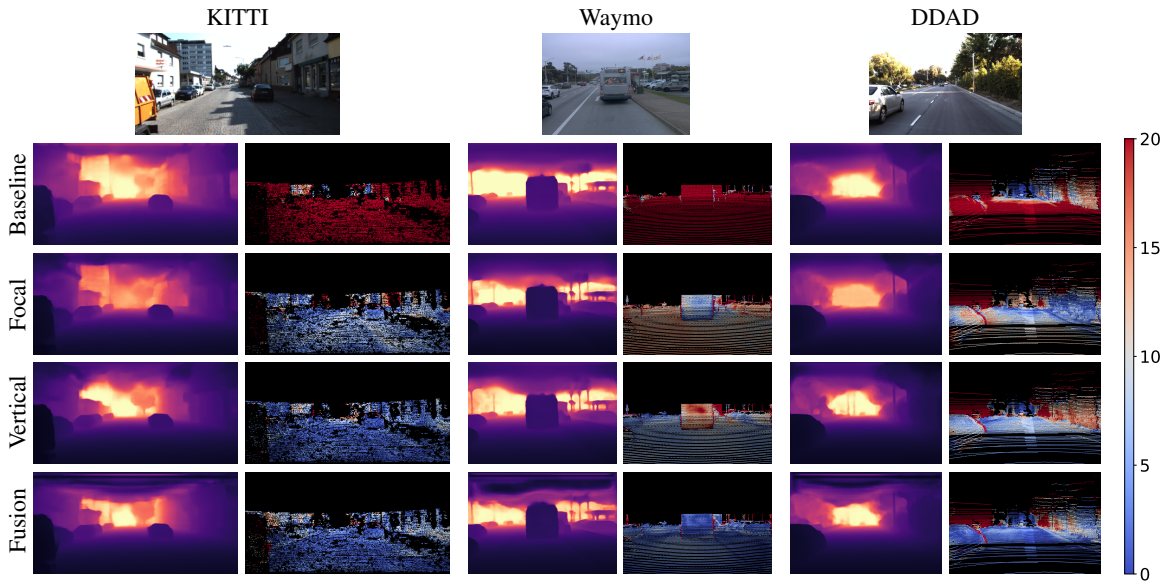

Figure 10. **Qualitative ablation.** Predicted depths and errors for DrivingStereo → {KITI, Waymo, DDAD} models. Baseline – standard depth regression with geometric augmentations. Focal – depth regression via Focal Canonical Transform - $\mathrm{FCT}_{\boldsymbol{\mathcal{C}}}(\cdot)$. This is equivalent to Metric3D[27], but trained on our setup. Vertical – depth regression via Vertical Canonical Transform - $\mathrm{VCT}_{\boldsymbol{\mathcal{C}}}(\cdot)$. Fusion – depth regression with uncertainty-based fusion model.

turing object boundaries and overall scene layout. However, error maps reveal that the Baseline model struggles to handle the domain gap introduced by varying camera parameters in the zero-shot transfer setting, leading to reduced accuracy. Both the Focal and Vertical models perform reasonably well across most image regions, though they exhibit noticeable modeling errors in certain areas. In contrast, the Fusion model attains the highest accuracy by adaptively integrating the depth predictions from both Focal and Vertical models.

C. Model Architecture

In this work, we employ a fully convolutional encoder-decoder architecture with skip connections. While vision transformer (ViT)-based encoders, such as DINOv2 [15], often achieve superior accuracy, their advantages typically rely on large-scale training. A similar fully convolutional design is adopted in Metric3D [27], demonstrating the effectiveness of such models for generalizable monocular depth estimation (MDE). Given our focus on single-dataset training, we prioritize convolutional backbones to reduce computational complexity while maintaining robust generalization performance.

Decoder Architecture Details. Our decoder is designed similarly as in Metric3D [27], consisting of four blocks that progressively upsample and fuse encoder features from a resolution of $(\frac{H}{32}, \frac{W}{32})$ to $(\frac{H}{2}, \frac{W}{2})$. Decoder channels dimensions are $\{756, 512, 256, 128\}$. Our fusion module processes $(\frac{H}{2}, \frac{W}{2})$ resolution feature maps $\mathbf{F}$ with two compact

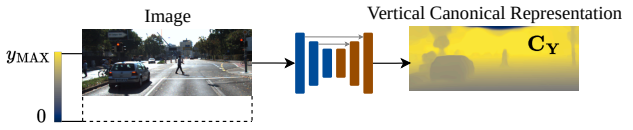

Figure 11. **Vertical Canonical Representation regression.** Visualization of regression boundaries of proposed *Vertical Canonical Representation*. We predict vertical image position of ground plane projection in range $[0, y_{MAX}]$. Image is ***virtually extended*** beyond bottom boundary to enable depth estimation for pixels with ground-plane projection below camera field-of-view. For stability reasons, regression is bounded to y_{MAX}, which corresponds to $d_{MAX} = 80$.

U-Net-like networks [19], each containing 3 downsampling and upsampling blocks, producing two final feature representations. These feature representations are processed with two convolutional blocks which predict our canonical representations $(\mathbf{C_F}, \mathbf{C_Y})$ and accompanying uncertainty estimates $(\mathbf{\Sigma_F}, \mathbf{\Sigma_Y})$. After uncertainty-based fusion, final depth map is bilinearly upsampled to (H, W) resolution.

Vertical Canonical Representation details. For convenience, we restate the equation used in Vertical Canonical Transform $\mathrm{VCT}_{\boldsymbol{\mathcal{C}}}(\cdot)$:

$$d = \frac{f_y h}{(H - c_y - y)\cos(\theta) - f_y \sin(\theta)}. \tag{2}$$

Upon closer inspection, it becomes evident that mapping $y \mapsto d$ can lead to unstable training if not properly regular-

ized. Specifically, at y corresponding to the horizon level of a given camera setup, the depth d approaches infinity. Furthermore, for any y above the horizon, the depth d becomes negative. To address this, as visualized in Fig. 11, we bound the regression of y to y_{MAX}, derived from the inverse mapping from $d_{MAX} = 80$. Additionally, to accommodate depth regression for objects with ground-plane projections below the camera's field of view, we *virtually extend* the image. In practice, this corresponds to increasing H in Eq. (2) accordingly.

D. Model Complexity

Table 8. **Model architecture details.** Encoder architectures number of parameters and execution time for models used in this work. Execution time is measured for resolution 416x640, on a single NVIDIA RTX A6000 GPU.

Method	Encoder	Params #	Execution time
Monodepth2 [9]	ResNet-50 [12]	34M	12.70 ms
DIFFNet [29]	HRNet-18 [21]	11M	29.42 ms
NeWCRFs [28]	Swin-L [13]	270M	62.56 ms
iDisc [16]	Swin-L [13]	208M	120.51 ms
PlaneDepth [22]	ResNet-50 [12]	39M	15.21 ms
MiDaS [18]	ResNeXt-101 [23]	105M	26.51 ms
LeReS [26]	ResNet-50 [12]	52M	20.26 ms
ZeroDepth [11]	ResNet-18 [12]	232M	238.48 ms
Metric3D [27]	ConvNext-L [14]	203M	10.74 ms
UniDepth [17]	ConvNext-L [14]	239M	55.34 ms
GVDepth	ConvNext-L [14]	228M	21.24 ms

In Tab. 8 we present the architectural details of the models used in this work. While most results align with expectations, there are a few notable outliers worth discussing. First, ZeroDepth exhibits the highest inference times, despite utilizing the least complex backbone among all methods. Essentially, ZeroDepth shifts complexity from the pretrained backbone to its proprietary self-attention layers in the decoder. This trade-off limits its ability to fully leverage the benefits of large-scale pretraining, which may explain its subpar accuracy compared to UniDepth, Metric3D, and GVDepth, even though it uses significantly more data during training.

On the other hand, Metric3D achieves remarkably low latency in depth prediction despite employing a relatively complex backbone. This efficiency likely stems from predicting depths at a reduced $(\frac{H}{4}, \frac{W}{4})$ resolution, which is subsequently upsampled to the original resolution Lastly, GVDepth stands out for its lightweight design and low inference time, even though it uses a relatively complex backbone. Its impressive generalization performance is attributed to our novel methodology rather than reliance on model or data scaling. In future work, we will investigate more complex transformer architectures, which were costly to train on our current computational setup.

Table 9. **Ablation of fusion strategies.** Mean – fusion via mean of $\text{FCT}_{\mathcal{C}}(\cdot)$ and $\text{VCT}_{\mathcal{C}}(\cdot)$. L1 – Uncertainty loss without aleatoric uncertainty weighting in $\mathcal{L}_{unc}$. All models are trained on DrivingStereo.

Configuration	DrivingStereo		KITTI		Waymo	
	A.Rel ↓	δ_1 ↑	A.Rel ↓	δ_1 ↑	A.Rel ↓	δ_1 ↑
Mean	**3.05**	99.4	8.21	91.1	14.01	82.8
L1	3.09	99.4	8.02	91.4	13.23	82.9
Fusion	3.07	**99.5**	**6.96**	**92.7**	**12.15**	**83.1**

E. Ablation of Fusion Strategy.

In Tab. 9 we examine the efficacy of different fusion strategies. Zero-shot transfer results demonstrate that our adaptive fusion weighted by aleatoric uncertainties leads to superior generalization performance.

F. Dataset Details

In this work, we use KITTI [8], DDAD [10], DrivingStereo [24], Waymo [20] and Argoverse Stereo [3] datasets, both for training and evaluation. For KITTI dataset, we evaluate all models on commonly used Eigen split [6] with Garg crop [7], resulting in 23158 training images and 652 testing images. On DDAD dataset we use the official training and validation split, with 12650 and 3950 images, respectively. Since Waymo, DrivingStereo, and Argoverse Stereo are not widely used for MDE evaluation, we simplify the process by creating custom dataset splits. The resulting training splits consist of {156K, 168K, 5K} samples, while the corresponding testing splits contain {5K, 5K, 500} samples, respectively.

G. Camera setup calibration

In this section, we provide additional details about our camera calibration procedure. Our proposed Vertical Canonical Transform $\text{VCT}_{\mathcal{C}}(\cdot)$, as indicated in Eq. (2), requires the knowledge of camera parameters $\mathcal{C} = \{f_y, c_y, h, \theta\}$. Here, for all datasets, f_y and c_y are usually known up to the reasonable error induced by the calibration procedure. However, for certain datasets, camera height h and camera pitch θ are either unknown, or not properly calibrated. To remain consistent throughout this work, we recalibrate the extrinsic parameters for each dataset, with details provided in Algorithm 1. For semantic segmentation of the road plane we use the DeepLabv3 model [4].

Calibration results. Estimated camera height h and camera pitch θ for each dataset are provided in Tab. 10. Moreover, in Fig. 12 we visualize the histograms obtained with our calibration procedure. Since DrivingStereo [24] and

Algorithm 1 Estimate Camera Height h and Pitch θ.

Require: RGB images $\{I_i\}_{i=1}^{N}$, ground-truth depth maps $\{D_i\}_{i=1}^{N}$

Ensure: Median camera height h_{median} and pitch θ_{median}

1: Initialize empty sets $\mathcal{H} = \emptyset, \Theta = \emptyset$

2: **for** $i = 1$ to N **do**

3: Run semantic segmentation model on I_i to acquire road pixels $\mathcal{P}_i$

$$\mathcal{P}_i = \{(u, v) \mid (u, v) \text{ belongs to the road in } I_i\}$$

4: Extract depths $\{D_i(u, v) \mid (u, v) \in \mathcal{P}_i\}$

5: Filter road pixels based on depth:

$$\mathcal{P}_i^{\text{filtered}} = \{(u, v) \in \mathcal{P}_i \mid D_i(u, v) < 20\}$$

6: Project filtered road pixels to 3D points:

$$\mathcal{Q}_i = \left\{ \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \middle| \begin{array}{c} X = \dfrac{(u - c_x)Z}{f_x}, \\ Y = \dfrac{(v - c_y)Z}{f_y}, \forall (u, v) \in \mathcal{P}_i^{\text{filtered}} \\ Z = D_i(u, v) \end{array} \right\}$$

7: Fit $\mathbf{R}_i(\theta_i)$ (rotation matrix) and h_i (camera height) to $\mathcal{Q}_i$ with RANSAC plane estimation

8: Append h_i to $\mathcal{H}$ and θ_i to Θ

9: **end for**

10: Compute median camera height and pitch:

$$h_{\text{median}} = \text{median}(\mathcal{H}), \quad \theta_{\text{median}} = \text{median}(\Theta)$$

11: **return** $h_{\text{median}}, \theta_{\text{median}}$

Table 10. **Camera calibration**. Camera extrinsics estimated by the calibration procedure described in Algorithm 1.

Dataset	h[m]	θ[°]
Argoverse Stereo [2]	1.678	0.021
DDAD [10]	1.459	-0.519
DrivingStereo [24]	1.739	-0.561
KITTI [8]	1.659	-0.664
Waymo [20]	2.145	-0.331

KITTI [8] do not report extrinsic parameters, we use the estimated values in our proposed canonical transform. Furthermore, for the DDAD dataset [10], we observe a significant discrepancy between the official and estimated extrinsics. Therefore, we use our calibrated parameters, as they align more closely with the ground-truth depth. In contrast, for Argoverse Stereo [2] and Waymo [20], the estimated values are consistent with the official calibration.

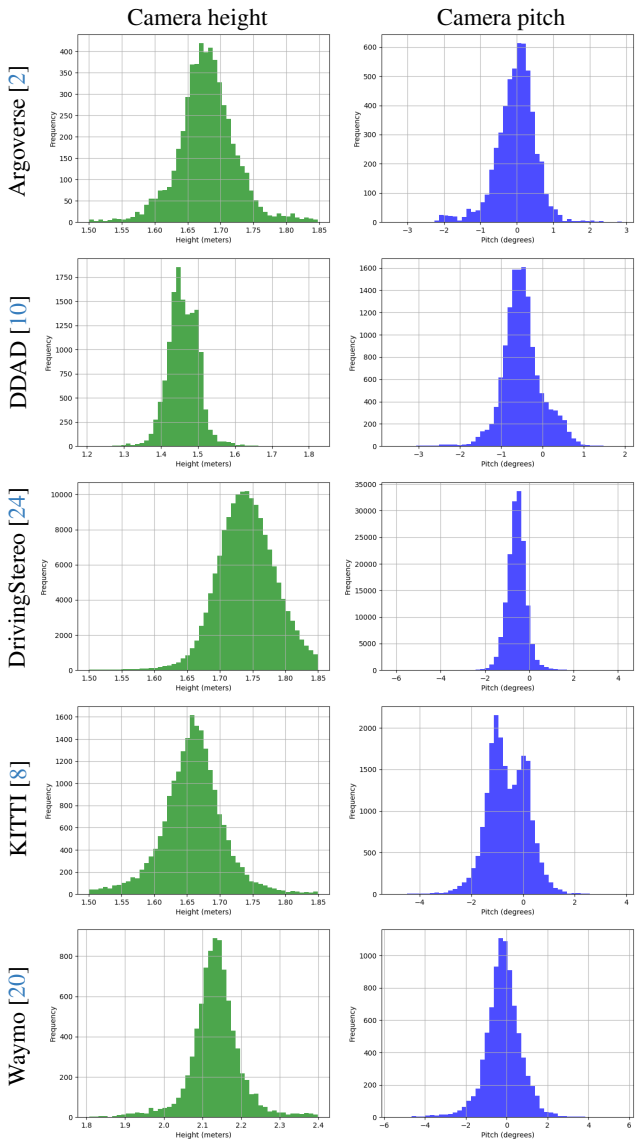

Figure 12. **Calibration histograms.** Height and pitch histograms acquired by calibration procedure described in Algorithm 1. Best viewed zoomned in.

KITTI Calibration Issues. We would also like to address the problems with KITTI calibration. As shown in Fig. 12, the pitch histogram for the KITTI dataset does not follow an unimodal distribution. We believe that this highlights inconsistencies in the official KITTI intrinsics calibration across different recording sequences. While this observation is not novel [5], it is rarely, if ever, discussed in the context of MDE. Given that the KITTI dataset remains a cornerstone for MDE evaluation, we believe that this issue should be given increased attention.

Evaluation with Per-frame Calibration. Our per-frame calibration procedure offers potentially higher accuracy than median-filtered results. This is because it better cap-

Table 11. **Evaluation with per-frame extrinsics calibration.** Vertical – depth regression via Vertical Canonical Transform - $\mathrm{VCT}_{\mathcal{C}}(\cdot)$. Fusion – depth regression with fusion model. Vertical+ and Fusion+ indicate usage of per-frame camera extrinsic calibration during evaluation. Training datasets on rows, testing datasets on columns. Best results are **bolded**. In-domain evaluation results are shaded.

	Representation	KITTI			DrivingStereo		
		A.Rel ↓	RMS ↓	δ_1 ↑	A.Rel ↓	RMS ↓	δ_1 ↑
KITTI	Vertical	5.70	**2.58**	95.5	10.43	**5.42**	87.3
	Vertical+	5.72	2.62	95.6	10.45	5.46	87.1
	Fusion	5.67	2.61	95.7	10.24	5.66	87.4
	Fusion+	**5.61**	2.60	**95.8**	**10.22**	5.60	**87.5**
DStereo	Vertical	7.52	3.33	92.6	3.07	1.75	99.5
	Vertical+	7.31	3.28	92.7	**2.78**	**1.61**	**99.6**
	Fusion	6.96	3.17	92.7	3.01	1.76	99.5
	Fusion+	**6.85**	**3.15**	**92.7**	2.91	1.74	99.6

tures ground plane perturbations and vehicle dynamics, which cause slight variations in camera pitch. In Tab. 11, we present evaluation results using per-frame calibration. While this approach leads to slight performance improvement, the gain is marginal, possibly due to the noise in our calibration process. Although such additional information about vehicle dynamics and perturbation could theoretically be integrated into a real system using localization with visual odometry or sensor fusion, we exclude these results from the main paper since competing methods do not leverage this information.

Calibration Sensitivity. Models involving canonical mapping via $\mathrm{VCT}_{\mathcal{C}}(\cdot)$ are inherently sensitive to inaccuracies in camera extrinsic parameters. In Fig. 13, we evaluate the A.Rel metric for our Vertical and Fusion model configurations under varying levels of Gaussian noise in the extrinsic parameters.

While both models exhibit sensitivity to calibration noise, the Fusion model demonstrates a smaller error increase due to its uncertainty-based fusion with depth from $\mathrm{FCT}_{\mathcal{C}}(\cdot)$. However, the Fusion model's error still increases under noisy conditions, indicating that its uncertainty prediction does not fully compensate for inaccuracies in camera calibration. This is expected, as camera calibration directly affects the canonical mapping performed by $\mathrm{VCT}_{\mathcal{C}}(\cdot)$ and is not internally estimated by the model.

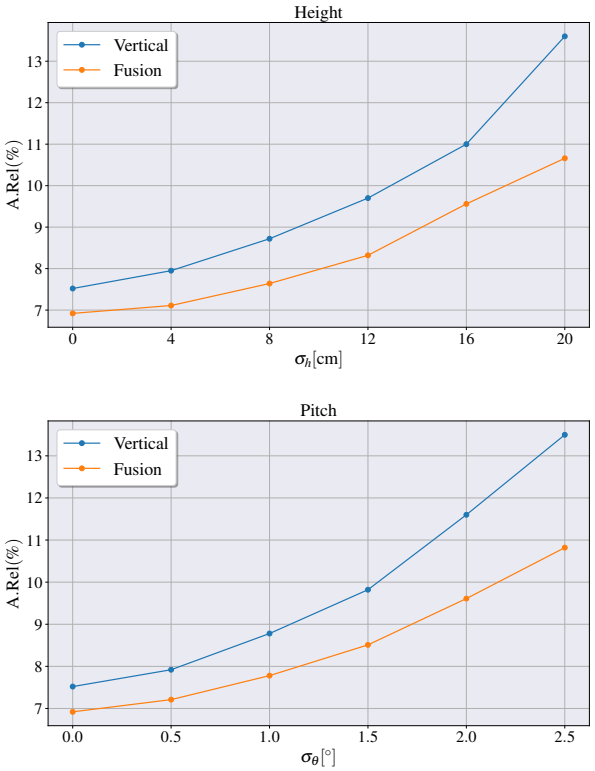

Figure 13. **Calibration sensitivity.** Sensitivity to Gaussian noise perturbations in extrinsic parameters evaluated on KITTI dataset. Vertical – depth regression via Vertical Canonical Transform - $\mathrm{VCT}_{\mathcal{C}}(\cdot)$. Fusion – depth regression with fusion model. Train/test dataset combination for both models is DrivingStereo → KITTI.

H. Additional Considerations

Multi-dataset Training. Almost all challenges in multi-dataset training for MDE arise from inconsistent and ambiguous perspective geometries, making our approach inherently scalable due to the invariance introduced in canonical spaces. While we lacked the computational resources to scale the training to the level of a MDE foundation model, in Tab. 3 in main text we demonstrated our method's ability to leverage diverse perspective geometries in the training data induced by geometric augmentations, especially compared to the *Baseline*.

Flat Ground Assumption. Our method implicitly assumes the ideal ground plane within the $\mathrm{VCT}_{\mathcal{C}}(\cdot)$. Unlike in GEDepth [25], which explicitly models the ground slope, we choose to grant the model greater flexibility. Motivation behind this choice is straightforward; our proposed canonical transformation is designed to assist the model in resolving ambiguities that diverse training data alone cannot, such as the entanglement of depth and camera parameters. For other adversarial perturbations, like ground plane imperfections, we do not explicitly model them. Instead, we allow the model to internally adjust values in canonical space when it detects these perturbations within highly diverse training data. Moreover, the uncertainty-based fusion provides the model with additional capabilities to adaptively weight cues based on aleatoric uncertainty, enabling the down-weighting of specific cue in highly uncertain regions.

References

[1] 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. 1, 2

[2] Ming-Fang Chang, John 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[3] 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. 4

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 4

[5] Igor Cvišić, Ivan Marković, and Ivan Petrović. Recalibrating the kitti dataset camera setup for improved odometry accuracy. In *2021 European Conference on Mobile Robots (ECMR)*, pages 1–6. IEEE, 2021. 5

[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. 4

[7] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 740–756. Springer, 2016. 4

[8] 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. 4, 5

[9] 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. 4

[10] 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. 4, 5

[11] 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. 4

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4

[14] 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. 4

[15] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[16] 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. 4

[17] 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. 4

[18] 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. 4

[19] 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. 3

[20] 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. 4, 5

[21] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 4

[22] 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, 2, 4

[23] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 4

[24] 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. 4, 5

[25] 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. 1, 2, 6

[26] 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. 4

[27] 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. 3, 4

[28] 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. 4

[29] Hang Zhou, David Greenwood, and Sarah Taylor. Self-supervised monocular depth estimation with internal feature fusion. In *British Machine Vision Conference (BMVC)*, 2021. 4