Supplementary Material for "Confidence Through Parallel Attention for Depth and Uncertainty Estimation in Dynamic Environments"

Onkar Susladkar[1], Rohit Pawar[2], Chirag Sehgal[3], Samaksh Ujjawal[2], Sparsh Mittal[4]
[1]UIUC [2]IIIT-H [3]DTU [4]IIT Roorkee
(onkarsus13, rohitpawar2406, chiragsehgal224, s.k.u.ujjwal)@gmail.com
sparsh.mittal@ece.iitr.ac.in

This appendix provides additional details supporting our main paper. Specifically, we include:

- **Theoretical Justification:** We present a mathematical formulation and proof of our proposed parallel attention mechanism in Section A.
- **Training Setup and dataset details:** Detailed experimental configurations and training hyperparameters used in all evaluations are described in Section B.
- **Additional qualitative results on depth-estimation:** Depth-estimation results on 4K/HD images taken from the internet and a smartphone are shown in Section C.
- **Additional depth-estimation results at different resolutions:** Section D shows depth estimation results at different resolutions.
- **Additional qualitative results on surface normal estimation:** We have shown the Visual results and explanation of Normal Estimation in Section E
- **Additional qualitative comparison results on point cloud reconstruction from depth maps:** Additional side-by-side visual comparisons between **ConFiDeNet** and **Marigold** are provided in Section F, demonstrating improved visual fidelity and structure preservation.
- **Additional qualitative results on video-depth estimation:** We report additional visual results on the video depth estimation dataset viz., **KITTI-360** and **Bonn** in Section G.

A. Mathematical Proof of Parallel Attention Technique

In this section, we present a mathematical proof to establish the efficiency of our proposed parallel attention technique. Specifically, we demonstrate that the memory complexity can be significantly reduced by applying parallel cross-attention operations over a fixed set of query embeddings and multiple independent key-value sets. Unlike the traditional sequential stacking approach, which incurs a memory complexity of $\mathcal{O}(L \times N_q \times d)$, the parallel method achieves a lower complexity of $\mathcal{O}(\max(N_q, N_k) \times d)$, where N_q and N_k are the query and key lengths, respectively, and d is the feature dimension. This establishes that the parallel approach is much more memory-efficient than the sequential setup.

Proposition 1. *Given a fixed set of query embeddings $Q \in \mathbb{R}^{B\times N_q\times d}$ and multiple independent key-value sets $\{(K^{(i)}, V^{(i)})\}_{i=1}^{L}$ where $K^{(i)}, V^{(i)} \in \mathbb{R}^{B\times N_k\times d}$, applying parallel cross-attention operations reduces the peak memory complexity from $\mathcal{O}(L \times N_q \times d)$ (sequential stacking) to $\mathcal{O}(\max(N_q, N_k) \times d)$ (parallel execution).*

Proof. Recall that the scaled dot-product attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

where $Q \in \mathbb{R}^{B\times N_q\times d}$, $K, V \in \mathbb{R}^{B\times N_k\times d}$, and d_k is the feature dimension.

Sequential Cross-Attention: In the sequential setup, attention operations are performed in a cascading manner:

$$Z^{(1)} = \text{Attention}(Q, K^{(1)}, V^{(1)}),$$

$$Z^{(2)} = \text{Attention}(Z^{(1)}, K^{(2)}, V^{(2)}),$$

$$\vdots$$

$$Z^{(L)} = \text{Attention}(Z^{(L-1)}, K^{(L)}, V^{(L)}).$$

At each stage, the output $Z^{(i)}$ depends on the previous output $Z^{(i-1)}$, necessitating that each intermediate activation $Z^{(i)}$ is stored in memory for backward computation.

Thus, the cumulative memory complexity in the sequential case is:

$$\mathcal{M}_{\text{sequential}} = \mathcal{O}\left(\sum_{i=1}^{L} N_q^{(i)} \times d\right) \approx \mathcal{O}(L \times N_q \times d),$$

assuming $N_q^{(i)} \approx N_q$ remains approximately constant across layers.

Parallel Cross-Attention: In the parallel setup, the attention operations are computed independently using the shared query Q:

$$Z^{(i)} = \text{Attention}(Q, K^{(i)}, V^{(i)}), \quad \forall i \in \{1, \ldots, L\}.$$

Since all attention branches are independent, they can be computed simultaneously, and only the resulting outputs $\{Z^{(i)}\}_{i=1}^{L}$ need to be aggregated (via summation or concatenation) after computation.

In this case, the memory requirement is determined by the largest individual operation, resulting in:

$$\mathcal{M}_{\text{parallel}} = \mathcal{O}\left(\max(N_q, N_k) \times d\right),$$

where $N_k = \max_i N_k^{(i)}$.

Comparison: Comparing the two memory complexities:

$$\mathcal{M}_{\text{sequential}} = \mathcal{O}(L \times N_q \times d),$$
$$\mathcal{M}_{\text{parallel}} = \mathcal{O}(\max(N_q, N_k) \times d).$$

Since $L > 1$ and N_q, N_k are of similar orders of magnitude, it follows that:

$$\mathcal{M}_{\text{parallel}} \ll \mathcal{M}_{\text{sequential}}.$$

Thus, the parallel cross-attention mechanism offers significantly lower peak memory usage compared to sequential stacking, completing the proof. □

B. Evaluation platform, training setup and dataset details

B.1. Evaluation platform and training details

To enhance generalization in monocular depth estimation, we create a unified dataset by uniformly aggregating KITTI [7, 15], NYU Depth v2 [16], and Matterport3D [4] datasets, split into 80% training, 10% validation, and 10% testing. This ensures comprehensive exposure to diverse indoor and outdoor geometries, lighting, and spatial layouts. For evaluating depth estimation from [18], we considered Absolute Relative Error (Abs Rel), Root Mean Squared Error (RMSE), and $\delta_{1.25}$. To assess predictive uncertainty, we employ AUSE, which measures the alignment between predicted uncertainty and actual depth error. However, as these are rank-based and lack inter-image interpretability, we also report Absolute Relative Uncertainty (ARU) and Root Mean Squared Uncertainty (RMSU) [14], which provide more interpretable, absolute error-aligned uncertainty estimates.

We trained our model using a distributed setup comprising 16 nodes, each equipped with 8 AMD Instinct MI300X GPUs, providing 192 GB of HBM3 memory per GPU. Training was performed with a batch size of 128 samples per GPU. During training, input images were randomly resized between 256×256 and 2048×2048 pixels to promote scale robustness. To maintain computational feasibility, we utilized eight gradient accumulation steps.

Optimization was conducted using the Adam optimizer, with hyperparameters set to $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The initial learning rate was set to 10^{-4}, scheduled via cosine decay. A linear warm-up phase lasting two epochs was employed, starting from a learning rate of 10^{-6}. The model was trained for a total of 35 epochs.

To improve generalization, identical augmentations were applied to both the input RGB images and the corresponding depth maps. Augmentations included random rotations within $\pm 30°$, random distortions, affine transformations (including scaling, translation, and shearing), and simulated environmental effects such as rain and light optical occlusions. Each augmentation was applied with a probability of 0.4. During the initial warm-up phase, augmentations were frozen to stabilize early-stage training dynamics.

Furthermore, to enhance decoder flexibility without significantly increasing the parameter count, we incorporated Low-Rank Adaptation (LoRA [9]) modules into alternate layers of the VQ-VAE [21] decoder. We set the LoRA rank to 64 and used a scaling factor of $\alpha = 0.8$. This configuration facilitates efficient fine-tuning by enabling localized parameter updates while preserving the expressiveness of the pre-trained model. Together, these strategies improve depth prediction performance, particularly in complex and occluded environments.

B.2. Dataset-related details

We train our model using KITTI [7, 15], NYU-v2 [16], and Matterport3D [4], covering a wide spectrum of scene types across real-world indoor and outdoor environments. These datasets provide complementary challenges, for example, KITTI includes sparse outdoor scenes with dynamic objects and occlusions; NYU-v2 [16] offers dense indoor depth with cluttered and highly varied room layouts; Matterport3D [4] captures large-scale 3D indoor environments with diverse layouts and wide-baseline viewpoints.

For zero-shot generalization, we evaluate on four unseen datasets: DREDS [5] (synthetic driving scenes with noise and weather variation), nuScenes [3] (urban traffic with sensor fusion and low-light conditions), Virtual KITTI [2, 6] (controlled photorealistic variants of KITTI with perturba-

Table S.1. Comparison of depth estimation methods in terms of model size, computational cost, and inference speed at different resolutions.

Method	#Parameter	$\text{Flops}_{\text{HD}} \downarrow$	Native Output Resolution ↑	t_{VGA} (ms) ↓	t_{HD} (ms) ↓	t_{4K} (ms) ↓
DPT [19]	123M	-	384 × 384 (=0.15 MP)	334.43	306.60	27.8
ZoeDepth [1]	340M	-	384 × 512 (=0.20 MP)	235.7	235.1	235.4
UniDepth [17]	347M	630G	462 × 616 (=0.28 MP)	178.50	183.00	198.10
Metric3D [22]	203M	477G	480 × 1216 (=0.58 MP)	217.9	263.80	398.1
Marigold [12]	949M	-	768 × 768 (=0.59 MP)	5174.3	4443.60	4977.60
Metric3D v2 [10]	1.378G	6830G	616 × 1064 (=0.66 MP)	1299.6	1299.70	1390.2
PatchFusion [13]	203M	-	Original (tile-based)	840.12	840.29	844.59
ZeroDepth [8]	233M	10862G	Original	1344.30	8795.74	3492.29
ConFiDeNet	678M	700G	Original	319.23	472.65	672.57

tions), and Middlebury Stereo [20] (high-resolution indoor stereo with fine-grained geometry and reflections). These datasets present significant domain shifts, sensor variations, and environmental diversity.

Our method consistently outperforms all baselines across accuracy metrics (AbsRel, RMSE, $\delta_{1.25}$) and uncertainty-aware measures (RMSU, ARU, AUSE), demonstrating strong generalization and calibration under both synthetic and real-world distribution shifts.

C. Additional qualitative results on depth-estimation on 4K/HD images

Figure S.1 presents additional qualitative results of ConFiDeNet on a diverse set of high-resolution images (taken from the internet) with varying aspect ratios and complex visual content. The model demonstrates strong generalization by producing sharp and semantically consistent depth maps across a wide range of scenes, including natural landscapes, human portraits, animals, and vehicles. These examples highlight ConFiDeNet's ability to preserve fine structural details and maintain depth coherence even under challenging lighting, occlusion, and texture variations. Additionally, Figure S.2 shows the robustness of ConFiDeNet on images with very high resolution (around 4800 × 3400).

D. Depth estimation results at different resolutions

To evaluate the inference efficiency of ConFiDeNet relative to existing monocular depth estimation approaches, we benchmark all methods across three standard image resolutions: VGA (640×480), HD (1920×1080), and 4K (4032×3024). Average runtimes are reported in Table S.1, where all measurements include preprocessing, internal resizing (for fixed-resolution models), and forward inference, ensuring consistent and fair evaluation. Computational cost in terms of FLOPs (at HD resolution) and model parameter counts were obtained using the fvcore library.

ConFiDeNet demonstrates a favorable trade-off between speed and accuracy. It achieves inference times of 319 ms, 472 ms, and 672 ms at VGA, HD, and 4K resolutions, respectively — significantly outperforming computationally heavier baselines such as Marigold, PatchFusion, and Metric3D v2. Despite having 678M parameters, ConFiDeNet maintains a moderate FLOP count (700G at HD), substantially lower than Metric3D v2 [10] (6830G) and ZeroDepth (10862G), and comparable to lighter models like UniDepth (630G). Moreover, methods such as PatchFusion [13] and ZeroDepth suffer extreme slowdowns at higher resolutions due to inefficient tile-based or original-resolution processing, whereas ConFiDeNet remains scalable and practical for real-time and high-resolution deployments. Overall, these results validate that ConFiDeNet produces accurate and uncertainty-aware depth predictions and exhibits the computational efficiency and scalability required for embodied AI and robotics applications.

E. Results on Normal estimation

Surface normal estimation and depth estimation are fundamentally related tasks, as both seek to recover underlying 3D geometry from image data. A single positive scalar represents the depth of each pixel, whereas the surface normals are described as three-dimensional vectors constrained to the unit sphere. In real-world settings, ground-truth surface normals are challenging to acquire outside simulation environments or controlled laboratory conditions. Consequently, surface normals are often derived from available depth measurements; however, this approach introduces notable artifacts, including noise on planar regions and excessive smoothness at depth discontinuities.

To address these challenges, we train our model on three large-scale synthetic datasets that span both indoor and outdoor scenes. Notably, *Hypersim* and *InteriorVerse* offer photorealistic renderings of diverse indoor environments, providing rich supervision for both depth and normal estimation tasks.

Figure S.4 presents the qualitative results of ConFiDeNet on the FFHQ dataset for both depth estimation and surface

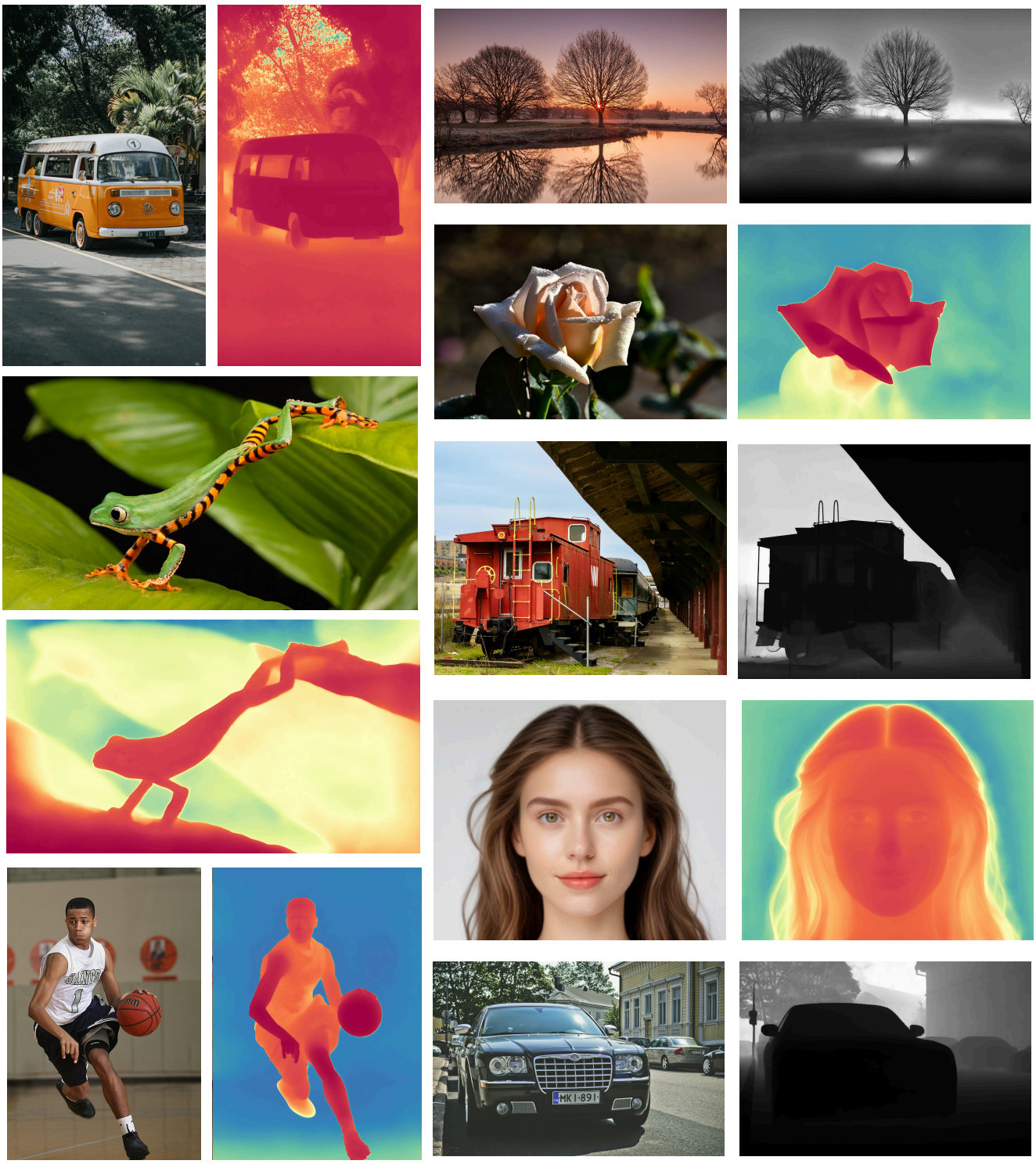

Figure S.1. Additional qualitative results of ConFiDeNet on images taken from the internet

normal prediction, demonstrating the ability of the model to recover accurate geometric information in challenging, real-world images.

Figure S.2. Additional qualitative results of ConFiDeNet on High-Resolution images captured from a smartphone on Historic monuments (Ellora Caves)

F. Additional visual comparison: Point cloud from Depth Map

Figure S.3 demonstrates the superiority of ConFiDeNet over MariGold [12] in the real-world path-planning scenario. While MariGold provides competitive depth predictions, ConFiDeNet demonstrates extended vertical and horizontal horizons in the point cloud, enabling longer and more feasible trajectories via standard planners (e.g., RRT* [11]). ConFiDeNet better preserves critical scene elements, such as right-turn pathways, which enhances downstream planning reliability.

In particular, ConFiDeNet maintains finer geometric fidelity, aligning closer to the actual scene structure visible in RGB images (e.g., vehicles and road turns), while MariGold exhibits greater depth compression in complex regions.

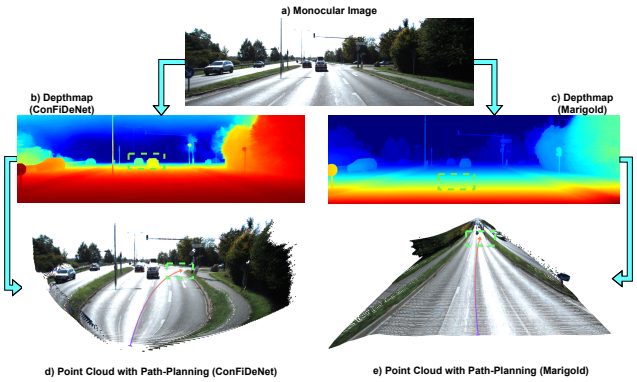

Figure S.3. Visualization of depth estimation and path planning results. **a)** Monocular RGB Image. Depthmap predicted from **b)** ConFiDeNet and **c)** Marigold [12]. Generated point cloud with planned trajectory based on depthmap from **d)** ConFiDeNet and **e)** Marigold. Trajectories are generated using standard planners (e.g., RRT* [11]), and show the scene structure captured in their depth predictions.

ConFiDeNet better preserves critical scene elements, such as right-turn pathways, which enhances downstream planning reliability. This improvement stems from our model's dense and pixel-accurate depth predictions, resulting in smoother color distributions and a more continuous depth representation. These advances contribute to more informative point clouds and enhanced real-time path planning performance. We refer the reader to the Supplementary Material for more point-cloud visualizations.

We conduct a qualitative comparison between our proposed method, ConFiDeNet, and the baseline Marigold [12] on real-world sequences from the KITTI dataset [7, 15], using reconstructed point clouds from estimated depth maps. In the visualizations shown in Figure S.5, red boxes indicate obstacles or elements visible in the RGB image but missing in the point cloud, while green boxes denote elements present in the RGB image but not reconstructed in the point cloud. Across multiple scenes (Examples I–V) in Figure S.5, ConFiDeNet consistently exhibits superior reconstruction quality in both the horizontal and vertical horizons. This expanded perceptual field enables capturing critical environmental cues that are often missing in Marigold's outputs. For instance, in Examples I–III, our model recovers occluded and distant obstacles—such as parked vehicles and roadside barriers— that are completely absent in the Marigold reconstructions. In Example IV, traffic lights are clearly visible in our point cloud but omitted in Marigold's output, underscoring the semantic degradation present in the baseline. Example V further demonstrates our model's ability to generate broader and deeper spatial reconstructions, facilitating high-fidelity visualization of long-range road geometry.

A wider field of horizon plays a pivotal role in safe and efficient autonomous path planning. Horizontally, it provides lateral scene context crucial for lane change decisions, intersection handling, and obstacle circumvention, while vertically, it allows the perception system to anticipate elevation changes, detect overhead objects, and better localize features in urban environments. The extended view captured by ConFiDeNet enhances both spatial coverage and geometric consistency, enabling downstream planners to compute trajectories that are not only smoother and longer-term, but also safer under uncertainty. By revealing obstacles and semantic elements earlier in the planning pipeline, our model allows anticipatory planning, reducing reactive behavior and improving robustness in dynamic scenes. These improvements in spatial perception directly translate into more accurate, informed, and reliable decision-making in autonomous navigation systems.

G. Additional visual comparison for Zero-shot video depth estimation

We present additional zero-shot qualitative comparisons on the *KITTI-360* and *Bonn* datasets, as shown in Figure S.6 and Figure S.7 respectively. Remarkably, our method achieves high spatial accuracy in the zero-shot setting, performing comparably to or even surpassing state-of-the-art models that are either trained or fine-tuned on these datasets. Moreover, inspection of the y-t slices indicates that while most competing approaches introduce high-frequency artifacts, our method yields consistently smoother results, demonstrating superior temporal consistency and generalization capabilities without dataset-specific adaptation.

References

[1] 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. 3

[2] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. 2

[3] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *CoRR*, abs/1903.11027, 2019. 2

[4] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 2

[5] Qiyu Dai, Jiyao Zhang, Qiwei Li, Tianhao Wu, Hao Dong, Ziyuan Liu, Ping Tan, and He Wang. Do-

Figure S.4. Normal and Depth estimation on FFHQ dataset

main randomization-enhanced depth simulation and restoration for perceiving and grasping specular and transparent objects. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[6] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4340–4349, 2016. 2

[7] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 2, 6, 8

[8] 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. 3

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2

[10] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[11] Sertac Karaman, Matthew R. Walter, Alejandro Perez, Emilio Frazzoli, and Seth Teller. Anytime motion planning using the rrt*. In *2011 IEEE International Conference on Robotics and Automation*, pages 1478–1483, 2011. 5, 6

[12] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation, 2024. 3, 5, 6, 8

[13] Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. Patchfusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10016–10025, 2024. 3

[14] Rémi Marsal, Florian Chabot, Angelique Loesch, William Grolleau, and Hichem Sahbi. Monoprob: self-supervised monocular depth estimation with interpretable uncertainty. In *Proceedings of*

Figure S.5. Comparison of Point cloud from Depth map for Marigold [12] vs ConFiDeNet on KITTI dataset [7, 15]

the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 3637–3646, 2024. 2

[15] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 6, 8

[16] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2

[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 (CVPR)*, pages 10106–10116, June 2024. 3

[18] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3227–3237, 2020. 2

[19] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 3

[20] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In Xiaoyi Jiang, Joachim Hornegger, and Reinhard Koch, editors, *Pattern Recognition*, pages 31–42, Cham, 2014. Springer International Publishing. 3

[21] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. 2

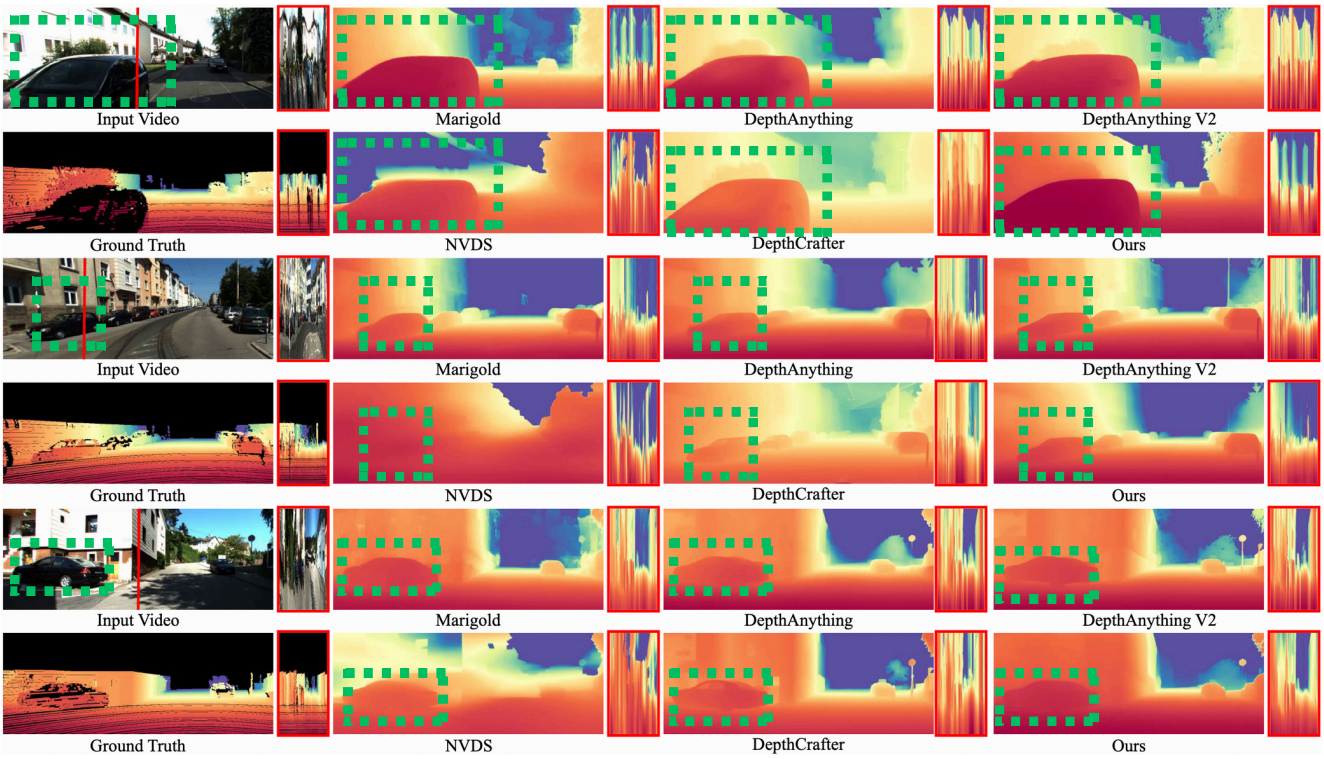

Figure S.6. Zeroshot Qualitative Comparision on **KITTI-360** dataset

[22] 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

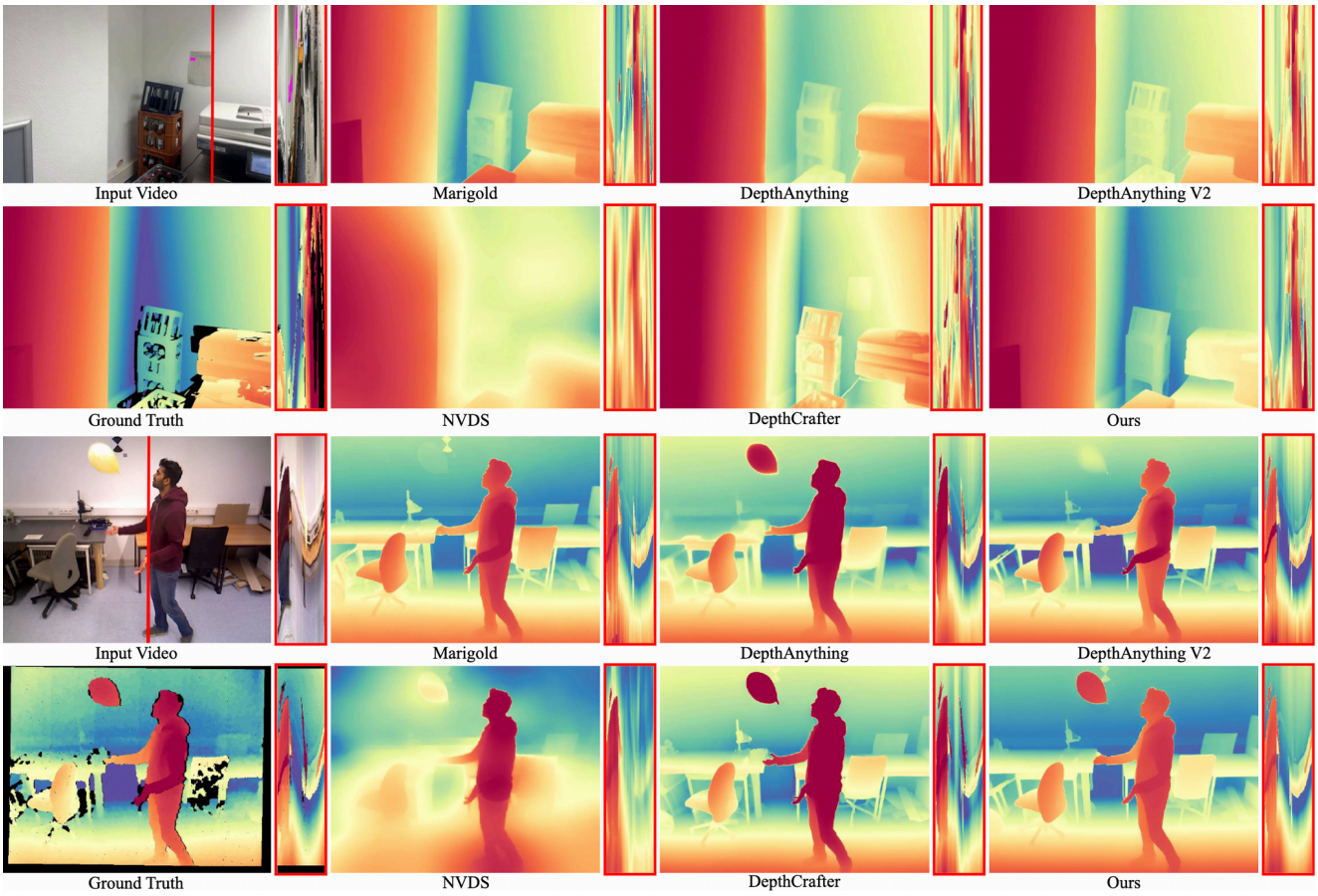

Figure S.7. Zeroshot Qualitative Comparision on **Bonn** dataset