Radar-Guided Polynomial Fitting for Metric Depth Estimation

Supplementary Material

Method	nuScenes MAE	nuScenes RMSE
DPT [16]	5188.2	6884.5
UniDepth [15]	2129.8	4887.7
Depth Anything [25]	2404.9	4851.1
Depth Pro [2]	3835.0	6600.3
RadarCam-Depth w/ DPT	1689.7	3948.0
RadarCam-Depth w/ UniDepth	1872.0	4321.2
RadarCam-Depth w/ Depth Anything	1953.6	5107.8
RadarCam-Depth w/ Depth Pro	3417.1	6462.0
TacoDepth w/ DPT	1492.4	3324.8
POLAR w/ DPT	1447.2	3304.9
POLAR w/ UniDepth	**1407.8**	**3193.5**
POLAR w/ Depth Anything	1515.1	3580.0
POLAR w/ Depth Pro	1461.5	4143.9

Table 1. **MDE backbone** comparative studies on nuScenes.

Method	ZJU MAE	ZJU RMSE
DPT [16]	1885.3	3326.1
UniDepth [15]	1533.0	3188.4
Depth Anything [25]	1943.2	3469.3
Depth Pro [2]	1680.2	3144.9
RadarCam-Depth w/ DPT	1183.5	3229.0
RadarCam-Depth w/ UniDepth	1152.5	3168.6
RadarCam-Depth w/ Depth Anything	1724.4	3661.3
RadarCam-Depth w/ Depth Pro	1490.6	3429.5
TacoDepth w/ DPT	1032.5	2850.3
POLAR w/ DPT	707.1	1216.9
POLAR w/ UniDepth	**629.6**	**1171.3**
POLAR w/ Depth Anything	657.2	1225.4
POLAR w/ Depth Pro	640.3	1174.8

Table 2. **MDE backbone** comparative studies on ZJU.

A. Comparative Studies

MDE Models. We evaluate POLAR using four monocular depth estimation (MDE) backbones: DPT [16], Depth Anything [25], Depth Pro [2], and UniDepth [15]. DPT and Depth Anything infer scaleless inverse depth, and while Depth Pro and UniDepth output metric depth, their raw predictions exhibit significant deviation from metric ground truth (see raw MDE performance in Tabs. 1, 2, 6), and thus we process them the same way as scaleless depth.

Tab. 1 compares raw MDE performance to their performance when used as backbones for RadarCam-Depth and POLAR. Raw MDE predictions show large errors due to scale ambiguity, while RadarCam-Depth provides moderate improvements. In contrast, POLAR consistently reduces error across all backbones, demonstrating the effectiveness of polynomial fitting. For DPT and Depth Anything, outputs are first inverted when used as a backbone, and are further median-scaled for the reported raw results. Among all configurations, POLAR w/ UniDepth achieves the best performance, improving over raw UniDepth predictions by 51.2% and over RadarCam-Depth w/ UniDepth by 37.8%.

For ZJU (see Tab. 2), among all configurations, POLAR w/ UniDepth achieves the best performance, improving over raw UniDepth predictions by 61.1% and over RadarCam-Depth w/ UniDepth by 54.2%.

TacoDepth. Although TacoDepth experiments with only DPT as its MDE backbone, our experiments show that a different backbone provides further performance gains. Nevertheless, POLAR outperforms TacoDepth by over 3% in MAE when both methods use the same DPT backbone. A similar trend holds on ZJU: even when both methods use the same DPT backbone, POLAR outperforms TacoDepth by 31.5% in MAE, further confirming that POLAR's state-of-the-art performance is not due to backbone choice.

B. More on Computational Efficiency

Incremental increase in computation for each additional polynomial degree. Adding a $k+1$-th polynomial term requires minimal computational overhead. For an MDE prediction of shape (H, W), the additional cost consists of three components: predicting another coefficient via the linear layer of $(1 \times 64 \times 2$ FLOPs), computing the $(k+1)$-th power from pre-computed k-th exponentiation ($H \times W$ multiplications), and incorporating this term into the final depth prediction ($2 \times H \times W$ operations for multiplication by the coefficient and addition). For nuScenes, where each image has shape $(900, 1600)$, each additional polynomial term incurs an additional computational cost of 0.0043 GFLOPs, which is a 0.0048% increase.

Cost breakdown. Following the reporting protocol of TacoDepth [23], we measure inference time as the latency added on top of the MDE backbone. Tab. 3 shows that, even when considering the combined cost, our method remains state-of-the-art in both inference time and GFLOPs.

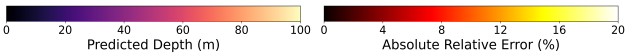

Figure 1. To quantify absolute improvement in our qualitative results, we provide the colorbars that were used in all examples in Figs. 2 and 3, as well as Figs. 3 and 4 in the main paper.

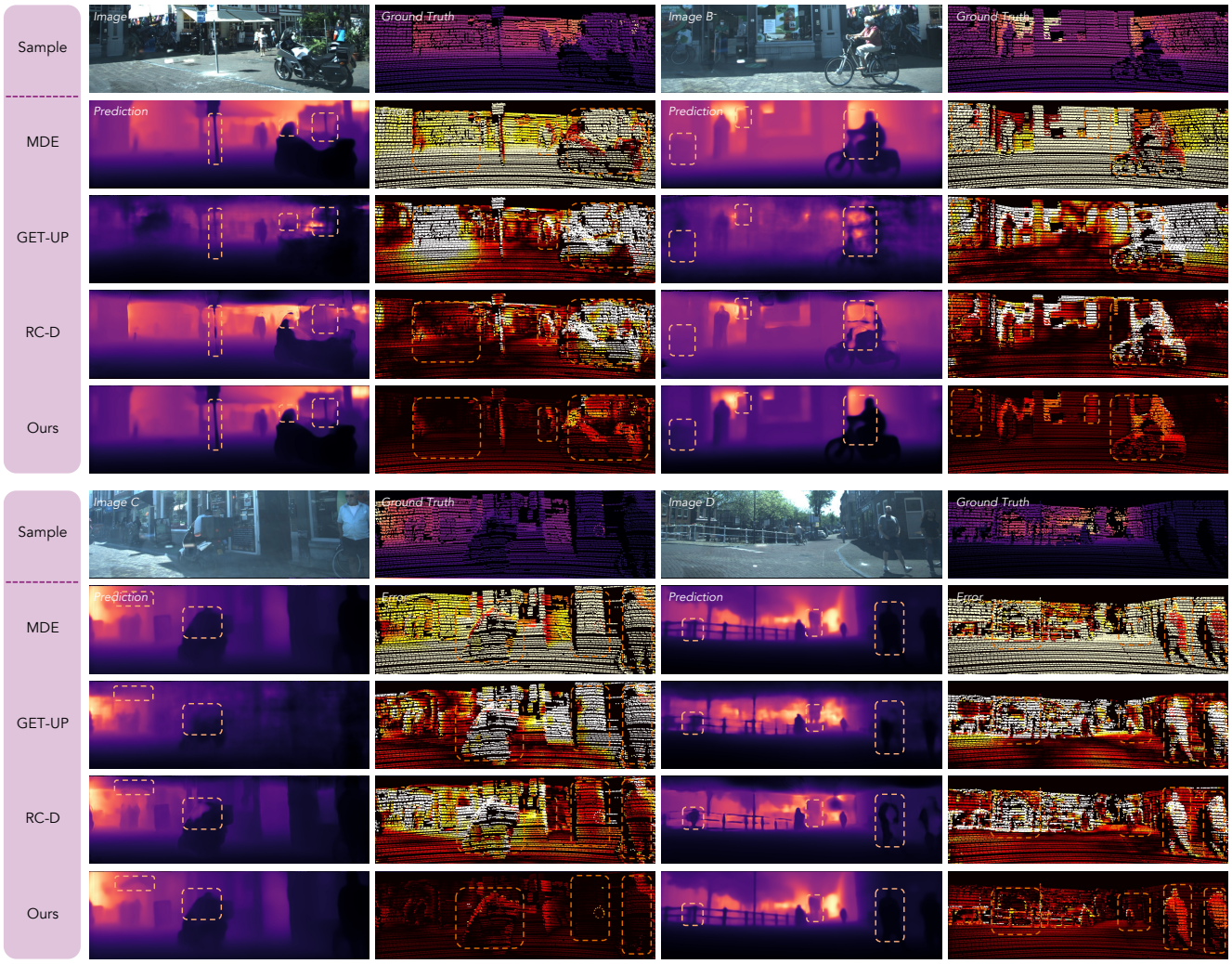

Figure 2. **Additional qualitative comparison** on VoD. POLAR more accurately predicts metric scale, and exhibits fewer global misalignments and geometric artifacts in comparison to other methods. Polynomial fitting, allowing POLAR to non-uniformly scale distinct local regions, results in notable improvements in metric depth estimation over initial MDE predictions. This is apparent in all MDE error maps: though object depth geometry is reasonable, actual metric depth values tend to be inaccurate. GET-UP, while marginally improving metric depth estimation relative to MDE, presents substantial errors in depth structure. For instance, the presence of geometric artifacts or omission of objects or regions entirely can be seen through the following highlighted examples: loss of motorcycle body and visor in Image A, front portion of biker (right) and bike wheel (far-left) in Image B, top of motorcyclist in Image C, and street sign in Image D. RadarCam-Depth (RC-D) similarly exhibits geometric artifacts, present in the sign pole, motorcycle, and fence post in Image A, biker and bicycle wheel in Image B, roof overhang and motorcyclist in Image C, and fence posts and pedestrians in Image D. In contrast, POLAR extrapolates correct structure from photometric priors while also demonstrating clear improvements in metric depth fidelity compared to all other baselines.

C. Dataset Details

The **nuScenes** dataset [3] contains 1,000 scenes, each lasting 20 seconds, collected from a vehicle equipped with Velodyne HDL32E lidar, Continental ARS 408-21 Radar, Basler acA1600-60gc camera with 900×1600, and Advanced Navigation Spatial IMU sensors around Singapore and Boston. This data collection process resulted in 40,000 synchronized keyframes. Each frame has an average of 97 radar point measurements. Additionally, the dataset includes 877,993 3D bounding box annotations about 23 object classes, and is organized with a train-test split of 850 scenes for training and validation, and 150 for testing.

The **ZJU-4DRadarCam** (ZJU) dataset [8] provides lidar, Radar, and camera data, collected through the same method as the nuScenes dataset around Hangzhou, China. The dataset is enhanced with high-density lidar and 4D radar data, utilizing the RoboSense M1 lidar sensor and Oculii's EAGLE 4D radar sensor. Additionally, the vehicle is outfitted with RealSense D455 cameras. The dataset includes a total of 33,409 synchronized keyframes, divided into 29,312

Method	nuScenes Inference Time	GFLOPs
Depth Anything [25]	28.70	201.03
+ Additional over Depth Anything	315.64	619.02
RadarCam-Depth w/ Depth Anything	344.34	820.05
+ Additional over Depth Anything	29.30	139.87
TacoDepth w/ Depth Anything	58.00	340.90
+ Additional over Depth Anything	24.81	89.70
POLAR w/ Depth Anything	53.51	290.73

Table 3. Computational cost breakdown for methods that leverage MDE. Inference time is measured in milliseconds (ms).

frames for training and validation, and 4,097 frames for testing. Each frame has an average of 465 radar point measurements. The original camera resolution was 720×1280 but was cropped to 300×1280 because of the limited presence of reprojected lidar points.

The **View-of-Delft** (VoD) dataset [12] uses similar methods to provide lidar, Radar, and camera data around the city of Delft in the Netherlands. The vehicle was equipped with a Velodyne HDL-64 S3 LIDAR, ZF FRGen 21 3+1D Radar, a stereo camera with 1216×1936 resolution, an RTK GPS, IMU, and wheel odometry. It contains 8,693 frames of synchronized and calibrated keyframes along with 123,106 3D bounding box annotations about 13 road user classes. Each frame has an average of 276 radar point measurements. Similar to the ZJU dataset, the camera resolution was cropped to 608×1936 because of the limited presence of reprojected lidar points.

D. Full nuScenes Benchmark

We present the full set of quantitative results on the nuScenes dataset in Tab. 4. This table includes additional baseline methods that were omitted from the main text for brevity, providing a comprehensive comparison of POLAR against all known existing radar-camera depth estimation methods. As shown, POLAR consistently outperforms all competing methods across all maximum distance thresholds (50m, 70m, and 80m), achieving the lowest mean absolute error (MAE) and root mean squared error (RMSE).

E. Comparison to Depth Completion

One potential idea for radar-camera depth estimation is to apply lidar-camera depth completion methods designed to densify sparse depth maps using surrounding context. One such method, BPNet [21] achieves state-of-the-art performance on the KITTI depth completion benchmark by leveraging bilateral propagation. However, when applied to radar-camera depth estimation, BPNet performs poorly, as shown in Tab. 4. POLAR outperforms BPNet by 37.3% in MAE and 31.2% in RMSE on nuScenes, highlighting the limitations of directly applying lidar depth completion methods to radar data. The key reason for this underperformance lies in the fundamental differences between lidar and radar point clouds. Unlike lidar, radar measurements are orders of magnitude sparser and significantly noisier due to factors such as limited antenna elements, elevation ambiguity (see Fig. 3), and lower range resolution [17]. Lidar depth completion methods assume a relatively dense and structured input [24], leveraging local spatial continuity to propagate depth estimates effectively.

Distance	Method	nuScenes MAE	RMSE
50m	RC-PDA [11]	2225.0	4156.5
	RC-PDA-HG [11]	2210.0	4234.0
	NLSPN [13]	2185.7	4411.6
	BTS [6]	1937.0	3885.0
	DORN [10]	1926.6	4124.8
	RadarNet [17]	1727.7	3746.8
	CaFNet [19]	1674.0	3674.0
	BPNet [21]	1618.9	3596.1
	Lin [9]	1598.2	3790.1
	SparseBeatsDense [7]	1524.5	3567.3
	RadarCam-Depth [8]	1286.1	2964.3
	GET-UP [20]	1241.0	2857.0
	TacoDepth [23]	1046.8	2487.5
	POLAR (Ours)	**1014.4**	**2475.7**
70m	RC-PDA [11]	3326.1	6700.6
	RC-PDA-HG [11]	3485.6	7002.9
	BTS [6]	2346.0	4811.0
	NLSPN [13]	2260.9	4645.0
	DORN [10]	2170.0	4532.0
	RadarNet [17]	2073.2	4590.7
	CaFNet [19]	2010.0	4493.0
	Lin [9]	1897.8	4558.7
	SparseBeatsDense [7]	1822.9	4303.6
	RadarCam-Depth [8]	1587.9	3662.5
	GET-UP [20]	1541.0	3657.0
	TacoDepth [23]	1347.1	3152.8
	POLAR (Ours)	**1286.1**	**2947.3**
80m	RC-PDA [11]	3721.0	7632.0
	RC-PDA-HG [11]	3664.0	7775.0
	AdaBins [1]	3541.0	5885.0
	P3Depth [14]	3130.0	5838.0
	LapDepth [18]	2544.0	5151.0
	PnP [22]	2496.0	5578.0
	NLSPN [13]	2519.3	5033.7
	BTS [6]	2467.0	5125.0
	DORN [10]	2432.0	5304.0
	RadarNet [17]	2179.3	4898.7
	CaFNet [19]	2109.0	4765.0
	Lin [9]	1988.4	4841.1
	SparseBeatsDense [7]	1927.0	4609.6
	RadarCam-Depth [8]	1689.7	3948.0
	GET-UP [20]	1632.0	3932.0
	TacoDepth [23]	1492.4	3324.8
	POLAR (Ours)	**1407.8**	**3193.5**

Table 4. Full quantitative results (mm) on the nuScenes benchmark.

Metric	Definition
MAE ↓	$\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert\hat{d}(x)-d(x)\rvert$
RMSE ↓	$\left(\frac{1}{\lvert\Omega\rvert}\sum_{x\in\Omega}\lvert\hat{d}(x)-d(x)\rvert^2\right)^{1/2}$

Table 5. **Error metrics for depth estimation.** These metrics compute the error between predicted depth $\hat{d}(x)$ and ground truth depth $d(x)$.

In contrast, radar points are too sparse for such methods to infer meaningful local depth relationships, leading to poor depth reconstruction when attempting direct densification.

In addition, we compare against Non-Local Spatial Propagation Network (NLSPN) [13], which achieves near state-of-the-art performance on the KITTI depth completion benchmark by refining sparse lidar depth with an iterative non-local spatial propagation procedure. NLSPN predicts an initial depth map along with pixel-wise confidences, then refines it by estimating non-local neighbors and their corresponding affinities to selectively propagate depth information. Unlike other approaches that rely on fixed local neighbors, NLSPN adaptively determines relevant non-local neighbors, improving depth completion accuracy, especially near depth boundaries. However, again, when applied to radar-camera depth estimation, NLSPN performs poorly, as shown in Tab. 4. POLAR outperforms NLSPN by 44.1% in MAE and 36.6% in RMSE on nuScenes.

Instead of relying on depth completion-style densification, POLAR directly learns a transformation function from radar to metric depth by leveraging polynomial fitting. Our method avoids the pitfalls of propagating unreliable local depth information by refining MDE predictions with learned polynomial coefficients, enabling flexible, scene-adaptive depth corrections that effectively capture object relationships and global scene structure.

F. Evaluation Metrics and Implementation Details

The evaluation metrics used in our study, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), are formulated in Tab. 5. Lower values equal better performance for both MAE and RMSE. Unless specified otherwise, all reported values are in millimeters (mm). We train for 60 epochs using a cosine decay learning rate scheduler with learning rate of 5×10^{-5}, and use weighting terms $\lambda_1 = 1.0, \lambda_2 = 0.4, \lambda_m = 0.25$ for our loss function.

G. Additional Comparisons

Full MDE Comparisons. For VoD (see Tab. 6), POLAR w/ UniDepth achieves the best performance in MAE, improving over raw UniDepth predictions by 42.4% and over RadarCam-Depth w/ UniDepth by 32.7%, while POLAR w/ Depth Anything achieves the best performance in RMSE, improving over RadarCam-Depth w/ Depth Anything by 37.6% and over raw Depth Anything predictions (inverted and median scaled) by 10.4%.

Method	View-of-Delft MAE	View-of-Delft RMSE
DPT [16]	4117.9	5498.9
UniDepth [15]	2605.6	5691.0
Depth Anything [25]	3270.5	4411.9
Depth Pro [2]	3275.9	5936.7
RadarCam-Depth w/ DPT	4013.5	5911.9
RadarCam-Depth w/ UniDepth	2227.4	5385.8
RadarCam-Depth w/ Depth Anything	3103.6	6328.7
RadarCam-Depth w/ Depth Pro	2843.4	6082.0
POLAR w/ DPT	1891.4	4252.6
POLAR w/ UniDepth	**1500.1**	3960.5
POLAR w/ Depth Anything	1770.5	**3951.8**
POLAR w/ Depth Pro	1520.2	3987.2

Table 6. **MDE backbone** comparative studies on View-of-Delft.

Method	nuScenes→ZJU MAE	nuScenes→ZJU RMSE	nuScenes→VoD MAE	nuScenes→VoD RMSE
GET-UP (zero-shot)	3845.2	8469.7	4809.1	8653.9
RadarCam-Depth (zero-shot)	5435.9	9785.8	7521.5	9194.8
GET-UP (trained)	1699.7	3882.6	2917.3	6145.1
RadarCam-Depth (trained)	1183.5	3229.0	**2227.4**	5385.8
Ours (zero-shot)	**1147.9**	**3109.5**	2256.2	**4744.2**

Table 7. **Zero-shot** generalization.

% radar kept / removed	RadarCam-Depth MAE	RadarCam-Depth RMSE	Ours MAE	Ours RMSE
25% kept / 75% removed	7969.4	10831.5	2416.4	4836.6
50% kept / 50% removed	4819.8	7077.7	1816.8	3945.7
75% kept / 25% removed	2537.3	4247.4	1575.2	3611.8
100% kept / 0% removed	1689.7	3948.0	**1407.8**	**3193.5**

Table 8. Reduced radar point density.

	MAE	RMSE
Learnable Points	1860.9	4207.1
Isotonic Reg.	2895.2	4340.5
Cubic Hermite	2131.3	4588.0
PCHIP	1809.6	4054.7
LoRA	2030.6	4493.7
ViT-Adapter	1859.6	4280.0
Our Performance	**1407.8**	**3193.5**

Table 9. Additional comparisons of POLAR vs. learnable dataset-specific points in place of radar points, and regression baselines.

Leveraging Radar. Tab. 9 shows that replacing radar points with learnable, dataset-specific points worsens MAE by 28.0% and RMSE by 29.9%, demonstrating that we in-

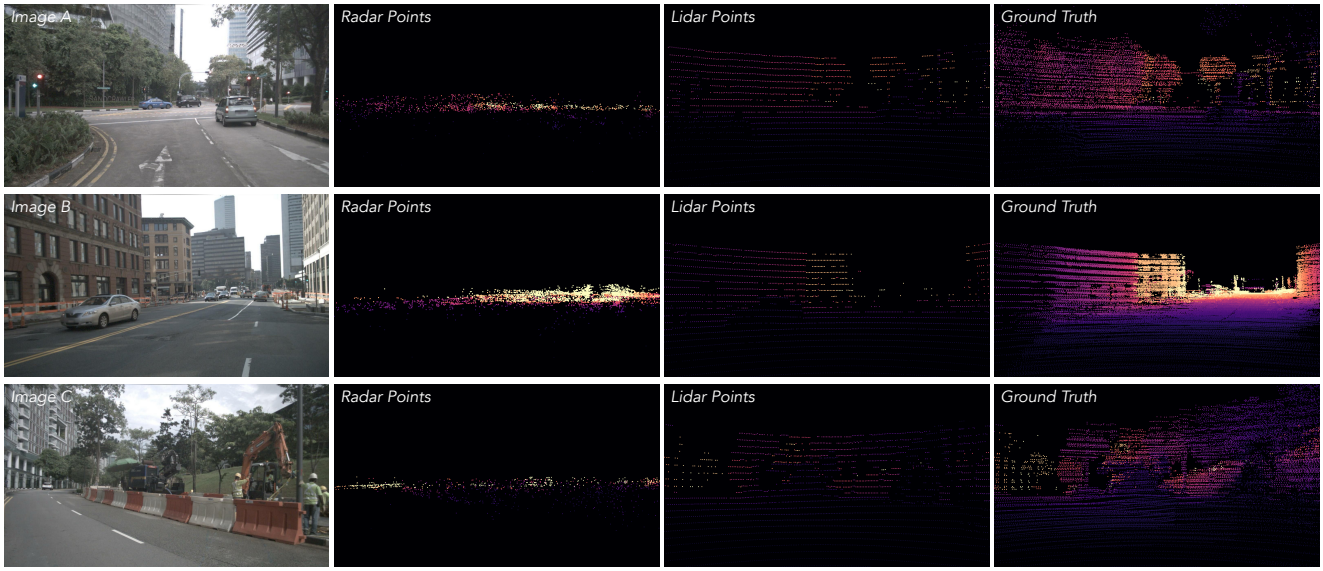

Figure 3. **nuScenes dataset visualization.** The elevation ambiguity of radar points results in erroneous projection onto the image plane that makes it challenging for depth completion methods to infer dense depth. In contrast, lidar points yield a denser, image-aligned projection, which is why accurate 3D scene reconstruction with depth completion methods is possible.

deed leverage the radar inputs effectively. As additional evidence, Tab. 7 shows that our method, evaluated zero-shot cross-dataset, achieves comparable or better performance than baselines trained on the target datasets.

Radar Sparsity. Tab. 8 shows we are more robust to reduced radar point density at inference, i.e., less performance degradation than the baseline method RadarCam-Depth.

Adapters. Tab. 9 shows that recent adapter-based finetuning methods LoRA [5] and ViT-Adapter [4], even with a post-hoc linear fit to projected radar points, do not outperform our method.

Regression Baselines. Isotonic regression and monotone spline fitting methods are natural baselines. Tab. 9 shows these methods for regressing projected radar points on MDE predictions do not outperform us. We hypothesize that this is due to noise in radar points that can be mitigated by learning (Sec. 3.2 from the main paper).

Our learned polynomial fit may, in principle, introduce unwanted inversions of initially correct MDE predictions. POLAR successfully mitigates this effect through the proposed novel first-derivative regularization term (see Sec. 3.4 and Eqs. 6, 7 from the main paper), which effectively constrains such inversions. To quantify this, we compute Kendall's τ coefficient between predicted and ground-truth depths. Our method achieves the highest monotonicity ($\tau = 0.969$) over regression baselines Isotonic Regression ($\tau = 0.871$), PCHIP ($\tau = 0.758$), and Cubic Hermite Spline ($\tau = 0.736$). The raw MDE predictions do exhibit monotonicity with respect to ground truth ($\tau = 0.957$), but our polynomial transformation increases it, indicating that we correct unwanted inversions. To assess statistical significance, we compute Kendall's τ over 30 bootstrap samples for both our method and the raw MDE predictions. A two-sample t-test reveals a statistically significant difference in mean monotonicity ($p = 0.012$).

H. Proof: Limitations of Global Scale and Shift

As further theoretical justification, we prove by construction that an affine scale-and-shift transformation is insufficient to fit MDE predictions to ground truth.

Proposition 1. *There exist infinitely many sets of $k \geq 3$ (*MDE prediction $\hat{d}$*,* ground truth d*) pairs such that no global scale α and shift β satisfy $d = \alpha\hat{d} + \beta$ for all k pairs simultaneously.*

Proof by Construction. Consider the following three (MDE prediction $\hat{d}$, ground truth d) pairs from the nuScenes dataset, specifically from the image shown in Fig. 4 from the main paper:

$$(\hat{d}, d) \in \{(5, 7), (10, 9), (15, 44)\}.$$

Assume to the contrary that there exist $\alpha, \beta \in \mathbb{R}$ such that

$$d = \alpha\hat{d} + \beta$$

holds for all pairs.

From the first two pairs, we obtain:

$$7 = 5\alpha + \beta, \quad 9 = 10\alpha + \beta.$$

Subtracting gives $\alpha = 0.4$ and $\beta = 5$ as the *unique* solution for these two pairs.

Applying this solution to the third pair yields:

$$\alpha \cdot 15 + \beta = 0.4 \cdot 15 + 5 = 11,$$

which contradicts the required equality with the ground truth value $d = 44$, since the residual error equals $44 - 11 = 33$ and not zero.

Hence no global scale α and shift β exist that can satisfy all three pairs simultaneously. Moreover, scaling each pair by any nonzero constant produces infinitely many distinct 3-sets of $(\hat{d}, d)$ pairs for which no α and β exists. Then, for any such 3-set, appending $k - 3$ arbitrary pairs yields an infinite family of k-sets ($k > 3$) that likewise admit no solution.

□

Corollary 1. *It is therefore a misconception that the scale ambiguity in MDE can be resolved solely by a global scale and shift. In contrast, any smooth relationship between $\hat{d}$ and d can be locally approximated by a polynomial via Taylor expansion, giving our polynomial fitting formulation the theoretical capacity to approximate d as a function of $\hat{d}$ arbitrarily well.*

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.

[2] 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, 2024.

[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. *arXiv preprint arXiv:1903.11027*, 2019.

[4] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

[5] 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.

[6] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

[7] Huadong Li, Minhao Jing, Wang Jin, Shichao Dong, Jiajun Liang, Haoqiang Fan, and Renhe Ji. Sparse beats dense: Rethinking supervision in radar-camera depth completion. In *European Conference on Computer Vision*, pages 127–143. Springer, 2024.

[8] Han Li, Yukai Ma, Yaqing Gu, Kewei Hu, Yong Liu, and Xingxing Zuo. Radarcam-depth: Radar-camera fusion for depth estimation with learned metric scale. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10665–10672. IEEE, 2024.

[9] Juan-Ting Lin, Dengxin Dai, and Luc Van Gool. Depth estimation from monocular images and sparse radar data. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10233–10240. IEEE, 2020.

[10] Chen-Chou Lo and Patrick Vandewalle. Depth estimation from monocular images and sparse radar using deep ordinal regression network. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3343–3347. IEEE, 2021.

[11] Yunfei Long, Daniel Morris, Xiaoming Liu, Marcos Castro, Punarjay Chakravarty, and Praveen Narayanan. Full-velocity radar returns by radar-camera fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16198–16207, 2021.

[12] Andras Palffy, Ewoud Pool, Srimannarayana Baratam, Julian F. P. Kooij, and Dariu M. Gavrila. Multi-class road user detection with 3+1d radar in the view-of-delft dataset. *IEEE Robotics and Automation Letters*, 7(2):4961–4968, 2022.

[13] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *European Conference on Computer Vision*, 2020.

[14] 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.

[15] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: Universal monocular metric depth estimation made simpler, 2025.

[16] 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.

[17] Akash Deep Singh, Yunhao Ba, Ankur Sarker, Howard Zhang, Achuta Kadambi, Stefano Soatto, Mani Srivastava, and Alex Wong. Depth estimation from camera image and mmwave radar point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9275–9285, 2023.

[18] Minsoo Song, Seokjae Lim, and Wonjun Kim. Monocular depth estimation using laplacian pyramid-based depth residuals. *IEEE transactions on circuits and systems for video technology*, 31(11):4381–4393, 2021.

[19] Huawei Sun, Hao Feng, Julius Ott, Lorenzo Servadei, and Robert Wille. Cafnet: A confidence-driven framework for radar camera depth estimation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2734–2740. IEEE, 2024.

[20] Huawei Sun, Zixu Wang, Hao Feng, Julius Ott, Lorenzo Servadei, and Robert Wille. Get-up: Geometric-aware depth estimation with radar points upsampling. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 1850–1860, 2025.

[21] Jie Tang, Fei-Peng Tian, Boshi An, Jian Li, and Ping Tan. Bilateral propagation network for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9763–9772, 2024.

[22] Tsun-Hsuan Wang, Fu-En Wang, Juan-Ting Lin, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. Plug-and-play: Improve depth estimation via sparse data propagation. *arXiv preprint arXiv:1812.08350*, 2018.

[23] Yiran Wang, Jiaqi Li, Chaoyi Hong, Ruibo Li, Liusheng Sun, Xiao Song, Zhe Wang, Zhiguo Cao, and Guosheng Lin. Tacodepth: Towards efficient radar-camera depth estimation with one-stage fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10523–10533, 2025.

[24] Chao Xia, Chenfeng Xu, Patrick Rim, Mingyu Ding, Nanning Zheng, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Quadric representations for lidar odometry, mapping and localization. *IEEE Robotics and Automation Letters*, 8 (8):5023–5030, 2023.

[25] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37: 21875–21911, 2024.