TacoDepth: Towards Efficient Radar-Camera Depth Estimation with One-stage Fusion

Supplementary Material

This supplement contains the following contents:

- More quantitative and qualitative results.
- More details on experimental settings.
- More implementation details for TacoDepth.

1. More Experimental Results

1.1. Quasi-dense Depth in Prior Arts

As mentioned in Fig. 2 and line 050 of our main paper, previous multi-stage approaches [5, 6, 8, 9, 13, 15] predict intermediate quasi-dense depth, which remains sparse and noisy. We provide more visual results of their intermediate and final depth in Fig. 1. Flawed quasi-dense results lead to blurred details, disrupted structures, and noticeable artifacts in their final predictions, especially on nighttime and glaring scenes, which limits the robustness of their models.

1.2. Model Robustness

We show more results to prove our robustness on both daytime and nighttime scenes (Sec. 4.3, line 469, main paper).

Quantitative Results. As shown in Table 1, on the nuScenes [2] dataset, we compare our model with the state-of-the-art two-stage method of Singh *et al.* [13] on daytime and nighttime scenes separately. For daytime samples, our method reduces MAE and RMSE by 12.9% and 11.0%. On nighttime scenarios, TacoDepth decreases MAE and RMSE by 29.1% and 25.2%. The results can further highlight our superior robustness under challenging nighttime conditions.

Visual Results. We present more visual comparisons for daytime (Fig. 5, Fig. 6) and nighttime samples (Fig. 7, Fig. 8). Without relying on the intermediate quasi-dense depth [5–9, 13–15], TacoDepth robustly predicts accurate depth with more complete structures and meticulous details on daytime and nighttime scenes.

Reasons for Robustness. Our robustness can be attributed to three factors. Firstly, our one-stage framework avoids reliance on intermediate results, thereby preventing the negative impacts of defective quasi-dense depth. Secondly, our graph-based Radar structure extractor captures the informative geometric structures and graph topologies. Compared with the simple point features [13], the overall structures are more robust and resilient [4, 12, 18] against Radar outliers. Furthermore, our pyramid-based Radar fusion module integrates Radar and image information from shallow to deep layers effectively. The Radar-centered flash attention can efficiently build cross-modal correspondences and suppress unreliable Radar points, which will be discussed in Sec. 1.3.

Scene	Method	0 - 50m		0 - 70m		0 - 80m	
		MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
Daytime	Singh *et al.* [13] (CVPR'23)	1618.9	3613.0	1924.7	4359.2	2017.9	4632.5
	TacoDepth (Ours)	**1389.5**	**3227.3**	**1680.9**	**3897.1**	**1782.4**	**4092.3**
Nighttime	Singh *et al.* [13] (CVPR'23)	2340.8	4683.8	2863.9	5935.4	3012.9	6338.3
	TacoDepth (Ours)	**1673.6**	**3631.4**	**1944.8**	**4425.3**	**2207.6**	**4574.8**
Overall	Singh *et al.* [13] (CVPR'23)	1727.7	3746.8	2073.2	4590.7	2179.3	4898.7
	TacoDepth (Ours)	**1423.6**	**3275.8**	**1712.6**	**3960.5**	**1833.4**	**4150.2**

Table 1. **Comparisons on daytime and nighttime scenarios of the nuScenes [2] dataset**. We calculate the average performance improvements across the three different depth ranges. On daytime scenes, compared with Singh *et al.* [13], our method reduces MAE and RMSE by 12.9% and 11.0%. On nighttime scenes, TacoDepth decreases MAE and RMSE by 29.1% and 25.2%. Overall, our model improves the performance by 17.0% and 13.9%. These results further prove our strong robustness on challenging scenarios.

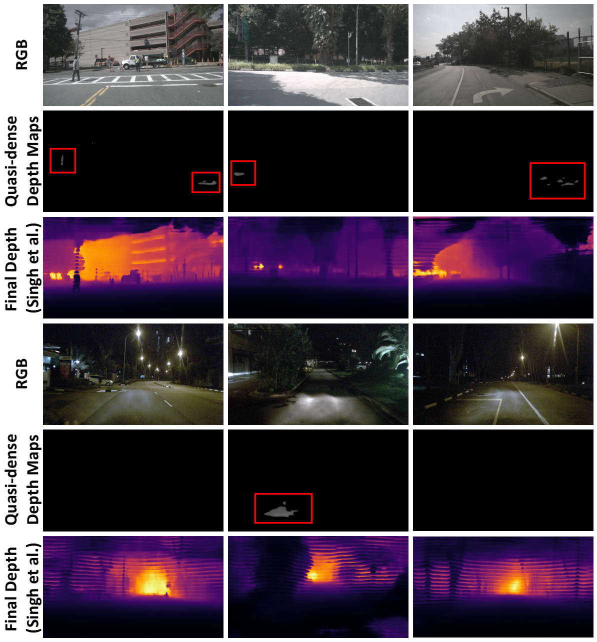


Figure 1. **Intermediate quasi-dense depth [5–9, 13–15] remains sparse and noisy.** We showcase more intermediate and final results from the previous two-stage method of Singh *et al.* [13] (CVPR'23). Pixels with valid depth values in the quasi-dense depth are visualized by gray areas in red rectangular boxes. Only few pixels exhibit valid depth. For some nighttime and glaring conditions, even no pixels are predicted with depth values. Due to the multi-stage frameworks [5–9, 13–15], the defective intermediate depth could lead to blurred details, disrupted structures, and noticeable artifacts in their final predictions. Our TacoDepth does not rely on quasi-dense depth, achieving superior efficiency, accuracy, and robustness with one-stage fusion.

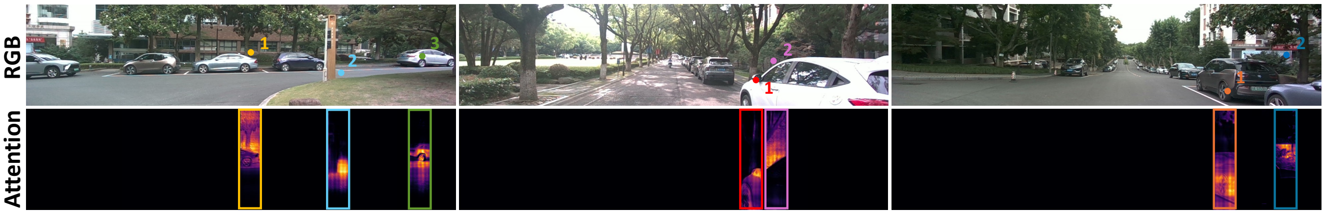


Figure 2. **Visualization of attention maps.** Some accurate Radar points are projected onto the image plane, as indicated by the colored points and indices. For each Radar point, the attention computation is confined to Radar-centered areas to maintain efficiency. In the attention maps, brighter colors represent higher attention scores. Our Radar-centered flash attention can effectively focus on correct corresponding regions and establish cross-modal correspondences. The attention maps also accurately distinguish between foreground and background, *e.g.*, shrub, road surface, car body, side mirror, stone tablet, car rear, and billboard. Best view zoomed in on-screen for details.

Module	MAE↓	RMSE↓	FLOPs (G)↓
Attention [16]	1982.4	4428.3	835.4
Radar-centered flash attention	**1712.6**	**3960.5**	**139.3**

Table 2. **Ablation on the Radar-centered flash attention.** We compare the original attention [16] with our Radar-centered flash attention implemented in TacoDepth. The MAE and RMSE are evaluated on the nuScenes dataset [2] in 0-70 meters. The FLOPs are reported for the whole model to process one 900 × 1600 image and 30 Radar points. Without restricting the Radar-centered areas, the original attention [16] fuses irrelevant image pixels and Radar points, resulting in unacceptable computational overheads. In contrast, our Radar-centered flash attention can effectively establish the cross-modal correspondences and maintain model efficiency.

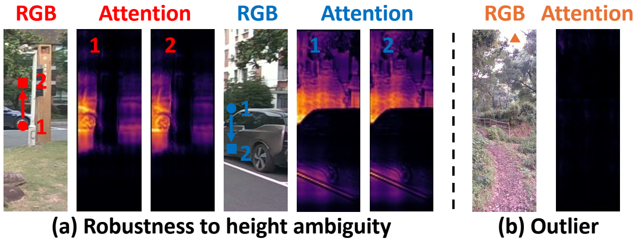
Metric	$a_l = \{32, 16, 8\}$	$\mathbf{a}_l$**={48, 32, 16}**	$a_l = \{64, 48, 32\}$
MAE↓	1811.3	**1712.6**	1785.7
RMSE↓	4179.8	**3960.5**	4082.2

Table 3. **Ablation on the widths $\mathbf{a}_l$ of Radar-centered areas.** The results are evaluated on the nuScenes dataset [2] in 0-70 meters. Reducing a_l could exclude some valid Radar points and image pixels from the fusion process, leading to a decrease in depth accuracy. On the other hand, since the horizontal Radar coordinates are relatively precise, using larger a_l could incorporate some irrelevant points and increase computational costs. Thus, we adopt $a_l = \{48, 32, 16\}$ in our experiments for the three fusion layers.

1.3. Radar-centered Flash Attention

In Sec. 3.2, line 252 of the main paper, we propose the Radar-centered flash attention mechanism in our pyramid-based Radar fusion module to build cross-modal correspondences between Radar points and RGB pixels. Here, we provide additional experiments to demonstrate its efficacy.

Visualization of Attention Maps. In Fig. 2, we visualize our Radar-centered flash attention. Several accurate Radar points are projected onto the image plane. The attention is calculated within Radar-centered areas for efficiency. External pixels and points could not be correlated, since horizontal Radar coordinates are relatively precise. As depicted in Fig. 2, our Radar-centered flash attention can effectively focus on the correct corresponding regions and establish cross-modal correspondences. The attention maps also accurately distinguish between foreground and background.



(a) Robustness to height ambiguity

(b) Outlier

Figure 3. **(a) Robustness to height ambiguity of 3D Radar.** The 3D Radar suffers from height ambiguity [9] due to insufficient antenna elements along the elevation axis. We simulate this issue by manually altering the vertical coordinates of some accurate Radar points (represented by the circles). Even with perturbed Radar inputs (represented by the squares), our attention maps can still identify the correct corresponding regions in the images, *e.g.*, the car wheel and the trees. **(b) Robustness to Radar outliers.** For Radar outliers with inaccurate point coordinates or depth values, *e.g.*, the orange triangle in the sky, TacoDepth suppresses these noisy points with low attention scores, such as the black attention map. Image pixels are only integrated with reliable Radar points.

Robustness to Height Ambiguity. Due to insufficient antenna elements along the elevation axis, 3D Radar suffers from height ambiguity [2, 9, 13] with unreliable vertical coordinates. We simulate this issue by manually altering the height dimensions of some accurate Radar points. As shown in Fig. 3(a), even with perturbed Radar inputs, our attention maps still identify correct corresponding regions.

Robustness to Radar Outliers. The accuracy of Radar is generally lower than LiDAR. When faced with Radar outliers, as shown in Fig. 3(b), our Radar-centered flash attention can suppress the noisy points with low attention scores. Image pixels will only be integrated with reliable Radar points, which can further enhance the model robustness.

Ablation on Radar-centered Flash Attention. Following Sec. 4.5, line 510 of our paper, we conduct an ablation on the Radar-centered flash attention. In Table 2, we compare the original attention [16] with our Radar-centered flash attention implemented in TacoDepth. Without restricting the Radar-centered areas, the original attention [16] fuses irrelevant image pixels and Radar points with heavy computa-

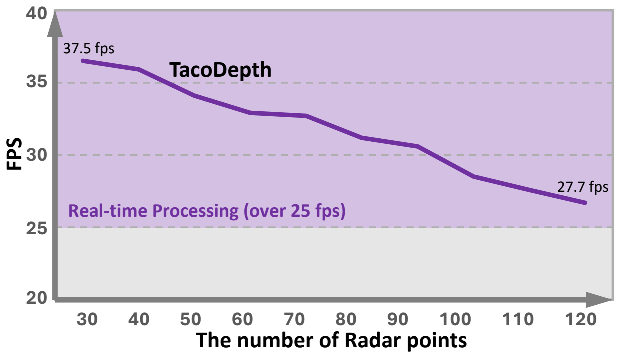


Figure 4. **Inference speed and Radar point numbers.** We evaluate the Frames Per Second (FPS) of our TacoDepth with different amounts of input Radar points. For the nuScenes dataset [2], the average and maximum numbers of Radar points per sample are 96.84 and 125. Our model achieves real-time processing (*e.g.*, over 27.7 fps) across typical Radar point numbers on nuScenes [2].

tional overheads. In contrast, our Radar-centered flash attention reduces the MAE, RMSE, and FLOPs by 13.6%, 10.6%, and 83.3%, which effectively establishes the cross-modal correspondences and maintains model efficiency.

Ablation on the widths $\mathbf{a}_l$ of Radar-centered areas. As noted in Sec. 4.2, line 399 of the main paper, we set $a_l = \{48, 32, 16\}$. Here, in Table 3, we ablate this specific choice. Reducing a_l could exclude some valid Radar points and image pixels from the fusion process, leading to a decrease in depth accuracy. On the other hand, since the horizontal Radar coordinates are relatively precise, using larger a_l could incorporate some irrelevant points and increase computational costs. Therefore, we adopt $a_l = \{48, 32, 16\}$ in all other experiments for the three fusion layers.

1.4. Model Efficiency

In Table 1 and Table 3 of the main manuscript, we compare the model efficiency under one 900×1600 image and 30 Radar points. Here, in Fig. 4, we further evaluate the inference speed of our TacoDepth with varying numbers of Radar points as input. From 30 to 120 Radar points, our model achieves real-time processing across the typical range of Radar point numbers on the nuScenes [2] dataset.

Multi-stage methods [5–7, 13, 15] are complex and inefficient. Two-stage independent models [6–9, 13–15] use two separate networks for intermediate and final depth. The four-stage plug-in RadarCam-Depth [5] employs least-squares optimization or RANSAC [3] for global alignment. TacoDepth noticeably outperforms these models in efficiency with one-stage fusion.

1.5. More Qualitative Results

We show more visual comparisons in Fig. 5, 6, 7, and 8. The Fig. 5 and Fig. 6 contain daytime samples, while Fig. 7 and Fig. 8 present nighttime scenarios.

2. More Details on Experimental Settings

2.1. Depth Metrics

As described in Sec. 4.1 of the main paper, following prior arts [5, 6, 9, 13], we adopt the commonly-applied depth metrics MAE, RMSE, iMAE, iRMSE, δ_1, and Rel for comparisons. Their definitions are specified in this section.

For Radar-Camera depth estimation, most previous works [5–9, 13, 15] utilize MAE and RMSE for evaluations. Besides, we also follow RadarCam-Depth [5] to report iMAE and iRMSE on ZJU-4DRadarCam [5]. The iMAE and iRMSE measure errors of inverse depth (*i.e.*, disparity), which are less sensitive to varied depth ranges (50, 70, or 80 meters).

To compare the plug-in models [5] with different depth predictors [1, 10, 11, 19] (Table 4, main paper), we adopt the δ_1 and Rel. These metrics are formulated as follows:

- Mean Absolute Error (MAE):
$$\frac{1}{|\Omega_{gt}|}\sum_{\Omega_{gt}}|D - D_{gt}|\,;$$
- Root Mean Square Error (RMSE):
$$(\frac{1}{|\Omega_{gt}|}\sum_{\Omega_{gt}}|D - D_{gt}|^2)^{\frac{1}{2}}\,;$$
- Inverse Mean Absolute Error (iMAE):
$$\frac{1}{|\Omega_{gt}|}\sum_{\Omega_{gt}}|\frac{1}{D} - \frac{1}{D_{gt}}|\,;$$
- Inverse Root Mean Square Error (iRMSE):
$$(\frac{1}{|\Omega_{gt}|}\sum_{\Omega_{gt}}|\frac{1}{D} - \frac{1}{D_{gt}}|^2)^{\frac{1}{2}}\,;$$
- Mean Relative Error (Rel):
$$\frac{1}{|\Omega_{gt}|}\sum_{\Omega_{gt}}\frac{|D - D_{gt}|}{D_{gt}}\,;$$
- δ_1 Threshold: pixel percentage of prediction D such that $max(\frac{D}{D_{gt}}, \frac{D_{gt}}{D}) = \delta < 1.25$, where D denotes the predicted depth. D_{gt} represents the depth ground truth. Ω_{gt} depicts the mask of D_{gt} with valid depth values.

2.2. Data Processing

We further illustrate the data processing procedures for the nuScenes [2] and ZJU-4DRadarCam [5] datasets (Sec. 4.2, line 390, main paper). Following prior arts [5–7, 9, 13], on the nuScenes [2] dataset, we accumulate 80 future and 80 past LiDAR frames to generate D_{acc}. Dynamic objects annotated by bounding boxes are removed before the projection. On ZJU-4DRadarCam [5], since it contains denser LiDAR returns and depth maps, we directly interpolate D_{gt} to obtain D_{acc} as the RadarCam-Depth [5].

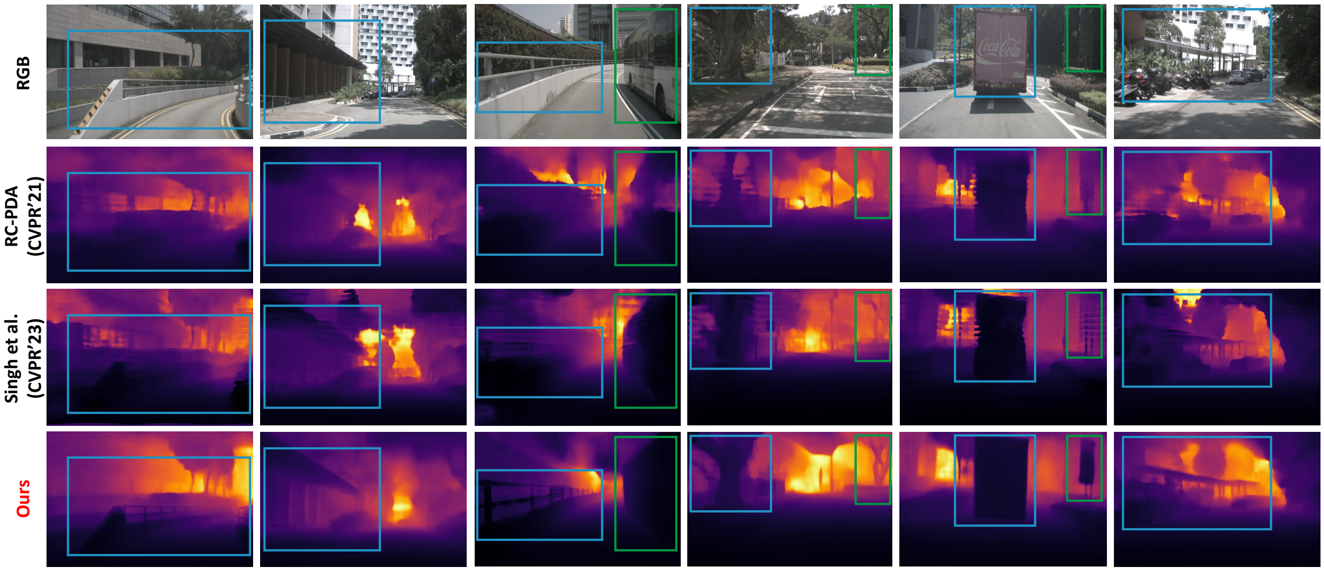


Figure 5. **Visual results on daytime scenes.** Previous multi-stage methods [9, 13] exhibit disrupted structures, blurred details, or noticeable artifacts. Our TacoDepth can predict accurate depth with complete structures and fine details. Best view zoomed in on-screen.

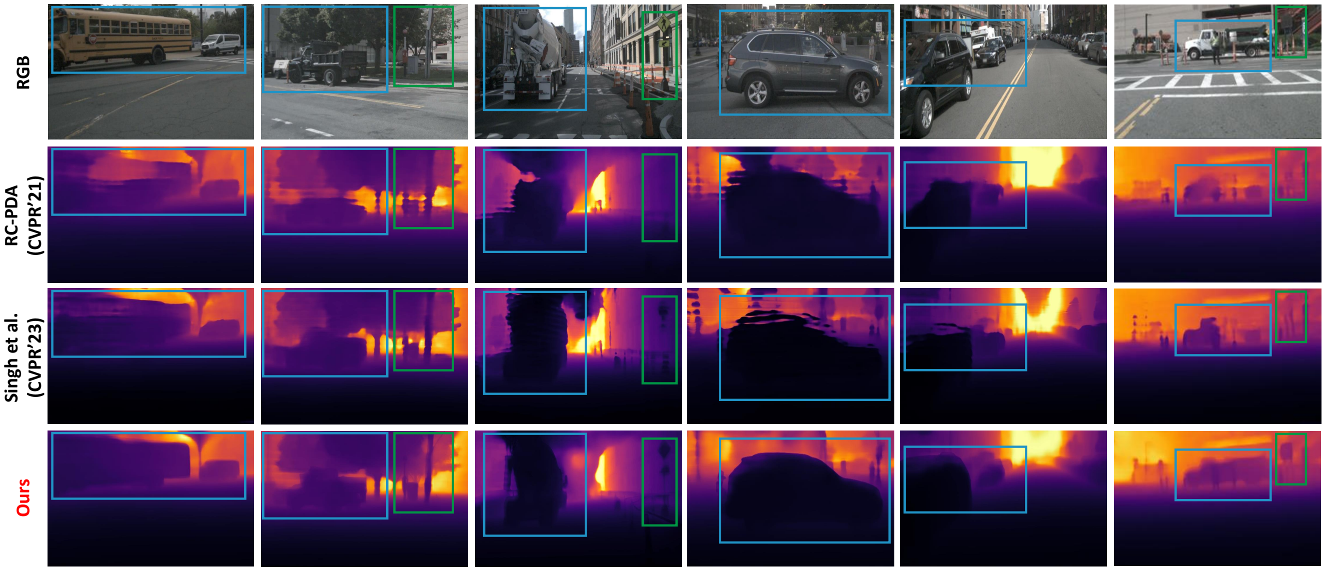


Figure 6. **Visual results on daytime scenes.** Previous multi-stage methods [9, 13] exhibit disrupted structures, blurred details, or noticeable artifacts. Our TacoDepth can predict accurate depth with complete structures and fine details. Best view zoomed in on-screen.

3. More Implementation Details for TacoDepth

Following Sec. 4.2, line 400 of the main paper, we present more detailed descriptions regarding our implementations.

3.1. Graph-based Radar Structure Extractor

As mentioned in Sec. 3.1, line 190 of the paper, our graph-based Radar structure extractor captures the geometric structures of Radar point clouds, involving a lightweight GNN [4, 12, 18] architecture with $L = 3$ layers. Each layer comprises a node and an edge generator. For one Radar point, the node generator extracts local node features from K-nearest neighboring points using MLPs, maxpooling, and concatenation. With the node features, the edge generator builds a soft adjacency matrix of Radar points as the edge feature via MLPs and attention [4, 16]. Node features can then be aggregated along edges by PCA-GM [17]. Thus, from shallow to deep layers, graph-based Radar structure extractor captures detailed coordinates and overall topologies, which are more robust to outliers [4, 12, 18].

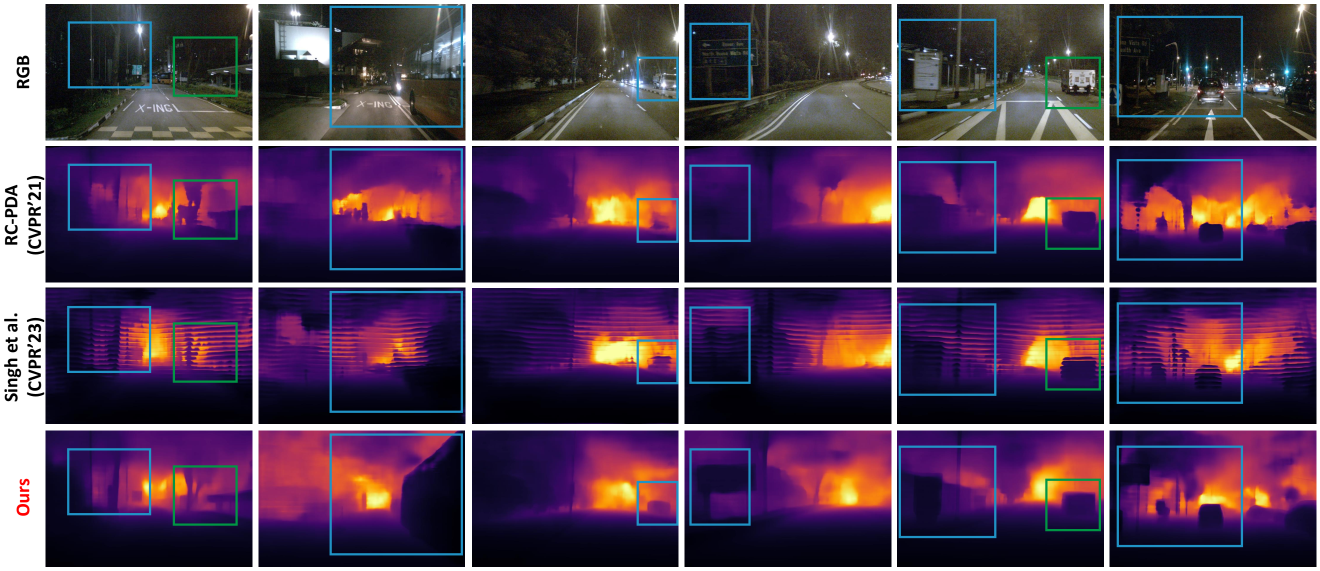


Figure 7. **Visual results on nighttime scenes.** Previous multi-stage methods [9, 13] rely on the intermediate quasi-dense results, which lack robustness, producing final depth with disrupted structures or even obvious artifacts on nighttime scenes. In contrast, our TacoDepth predicts more accurate depth with complete structures and fine details, showcasing our superior robustness. Best view zoomed in on-screen.

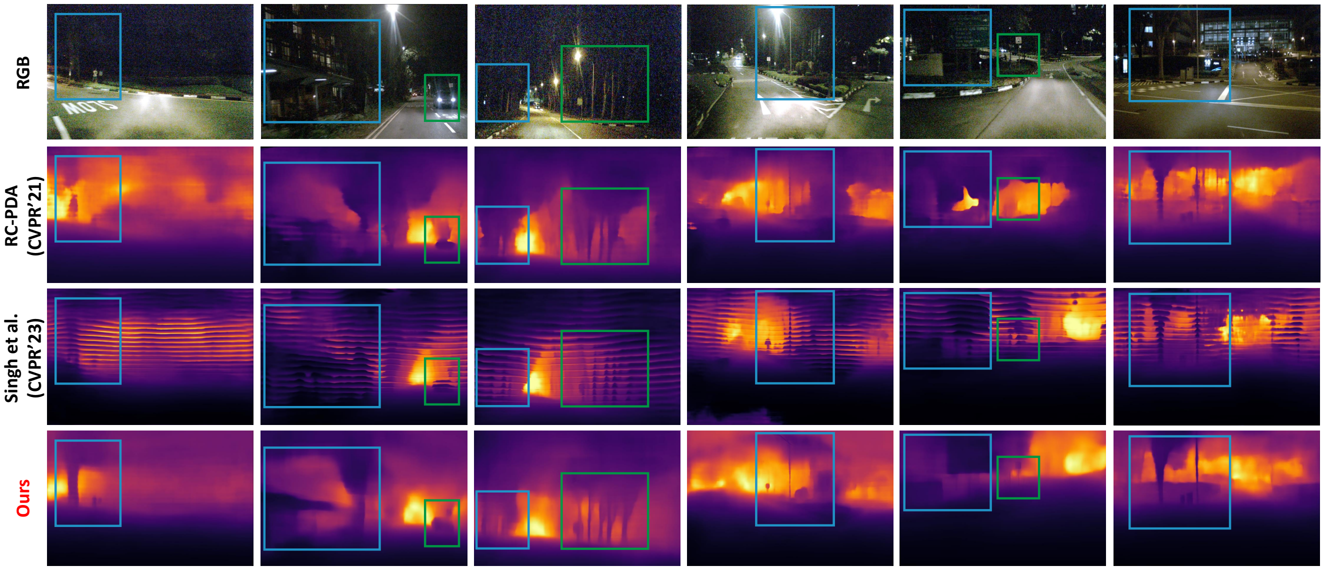


Figure 8. **Visual results on nighttime scenes.** Previous multi-stage methods [9, 13] rely on the intermediate quasi-dense results, which lack robustness, producing final depth with disrupted structures or even obvious artifacts on nighttime scenes. In contrast, our TacoDepth predicts more accurate depth with complete structures and fine details, showcasing our superior robustness. Best view zoomed in on-screen.

3.2. Depth Decoder

With the fused features, a common decoder [1, 5, 9, 10, 13] is employed to produce depth results (line 220, main paper). Specifically, resolutions are gradually increased while channel numbers are decreased. Skip connections are adopted to restore depth details. At last, an adaptive output module [10] adjusts the channel and restores depth maps.

3.3. Auxiliary Input Branch

TacoDepth is flexible for independent and plug-in inference (Sec. 3.3, main paper). For the plug-in mode, an auxiliary branch processes initial relative depth [1, 11, 19]. To be specific, convolution extracts features from initial depth, which are then fused with RGB features by concatenation and convolution. Other steps are identical to the independent mode.

3.4. The Name of TacoDepth

Ultimately, we would like to explain the naming of our proposed framework. We name it TacoDepth for two reasons. Firstly, TacoDepth is an acronym derived from the words in the title of our paper, representing our key objectives and focus, such as efficient, Radar, camera, depth, and one-stage. Moreover, our task shares similarities with the concept of a taco. Just as a taco wraps and blends diverse ingredients to create a new flavor, our framework aims to fuse the information from multiple sensors and modalities, yielding a more accurate, effective, and robust depth estimation model.

References

[1] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 3, 5

[2] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, 2020. 1, 2, 3

[3] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3

[4] Kexue Fu, Shaolei Liu, Xiaoyuan Luo, and Manning Wang. Robust point cloud registration framework based on deep graph matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8893–8902, 2021. 1, 4

[5] 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. *arXiv preprint arXiv: 2401.04325*, 2024. 1, 3, 5

[6] Juan-Ting Lin, Dengxin Dai, and Luc Van Gool. Depth estimation from monocular images and sparse radar data. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10233–10240. IEEE, 2020. 1, 3

[7] Chen-Chou Lo and Patrick Vandewalle. Depth estimation from monocular images and sparse radar using deep ordinal regression network. In *IEEE International Conference on Image Processing (ICIP)*, pages 3343–3347, 2021. 3

[8] Chen-Chou Lo and Patrick Vandewalle. Rcdpt: Radar-camera fusion dense prediction transformer. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[9] Yunfei Long, Daniel Morris, Xiaoming Liu, Marcos Castro, Punarjay Chakravarty, and Praveen Narayanan. Radar-camera pixel depth association for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12507–12516, 2021. 1, 2, 3, 4, 5

[10] 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(03):1623–1637, 2020. 3, 5

[11] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021. 3, 5

[12] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks and Learning Systems*, 20(1):61–80, 2009. 1, 4

[13] 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 (CVPR)*, pages 9275–9285, 2023. 1, 2, 3, 4, 5

[14] Tieshuai Song, Bin Yang, Jun Wang, Guidong He, Zhao Dong, and Fengjun Zhong. mmwave radar and image fusion for depth completion: a two-stage fusion network. In *The 27th International Conference on Information Fusion (FUSION)*, 2024.

[15] Huawei Sun, Hao Feng, Julius Ott, Lorenzo Servadei, and Robert Wille. Cafnet: A confidence-driven framework for radar camera depth estimation. *arXiv preprint arXiv:2407.00697*, 2024. 1, 3

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, 2017. 2, 4

[17] Runzhong Wang, Junchi Yan, and Xiaokang Yang. Learning combinatorial embedding networks for deep graph matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3056–3065, 2019. 4

[18] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 38(5), 2019. 1, 4

[19] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 3, 5