

# 1 Appendix

## 2 A. Detailed Structure and Evaluation of the UEM

3 Inspired by the work [1] for the degraded image recognition task, we design a novel uncertainty  
4 estimation module (UEM) for our depth estimation task and further use the output for distillation  
5 in the feature space and result space, which can better produce 3D-aware features and responses  
6 from monocular observations. The details of our proposed UEM for depth estimation are shown  
7 in Fig. 1. Unlike the uncertainty estimation module proposed in [1], which is situated at the end  
8 of the backbone network and uses a transposed convolution layer as a decoder while using residual  
9 blocks to learn the uncertainty, our UEM is located at the end of the encoder-decoder network and  
10 alongside the depth head as an uncertainty regression head to learn the per-pixel uncertainty. The  
11 proposed UEM consists of one convolution layer, one sigmoid activation function, one scaling layer,  
and an upsampling layer for rearranging the uncertainty map to the same size as the input image.

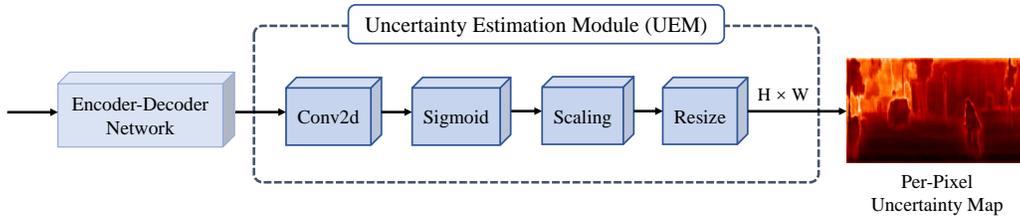


Figure 1: The details of our uncertainty estimation module (UEM) for depth estimation.

12

Model Method	Abs Rel AUSE ↓	RMSE AUSE ↓	$\delta \geq 1.25$ AUSE ↓
Monodepth2[2]	0.044	2.864	0.056
Mono-Uncertainty[3]	0.030	2.009	0.030
Training-Free[4]	0.031	0.871	0.029
GrUMoDepth[5]	0.024	0.494	0.017
ADU-Depth	<b>0.013</b>	<b>0.199</b>	<b>0.004</b>

Table 1: Uncertainty evaluation results on KITTI Eigen Split dataset based on the AUSE metric.

Method	Sq Rel ↓	Abs Rel ↓	RMSE ↓
DORN [6]	0.274	0.068	2.693
Adabins [7]	0.164	0.041	1.981
NeWCRFs [8]	0.157	0.039	1.977
ADU-Depth	<b>0.130</b>	<b>0.035</b>	<b>1.962</b>

Table 2: Quantitative results on the Virtual KITTI 2.

13 To validate how precise our estimated uncertainty values are, we follow [9] to use the Area Under  
14 Sparsification Error Curve (AUSE) metric. Specifically, we evaluate the AUSE in terms of the  
15 depth estimation metrics Absolute Relative Error (Abs Rel), Root Mean Squared Error (RMSE), and  
16 Accuracy ( $\delta \geq 1.25$ ) on the Eigen split of KITTI. Table 1 reports the comparison results, where we  
17 compare the uncertainty values produced by the UEM module with the state-of-the-art uncertainty  
18 estimation methods [2, 3, 4, 5]. Clearly, our UEM module achieves significant improvements over  
19 other competitors.

## 20 B. More Qualitative Results

21 To better understand the generalization capability and robustness of our method, we provide more  
22 qualitative results by comparing our method to state-of-the-art depth estimation methods in various  
23 scenarios on both real and virtual datasets, including the KITTI depth prediction online bench-  
24 mark [10] and Virtual KITTI 2 benchmark dataset [11].

25 **Evaluation on the KITTI Benchmark dataset.** We first note that our ADU-Depth ranked first  
26 on the official testing set of the KITTI depth benchmark. Please refer to [official website](#) for the  
27 leaderboard and Fig. 2 for the leaderboard screenshot. Note that our ADU-Depth was originally  
28 named as ZongDepth when the result was submitted.

	Method	Setting	Code	SILog	sqErrorRel	absErrorRel	iRMSE	Runtime	Environment	Compare
1	ZongDepth			9.69	1.69	7.26	9.61	0.1 s	1 core @ 2.5 Ghz (Python)	<input type="checkbox"/>
2	VA-DepthNet			9.84	1.66	7.96	10.44	0.1 s	1 core @ 2.5 Ghz (Python)	<input type="checkbox"/>
3	iDisc			9.89	1.77	8.11	10.73	0.1 s	1 core @ 2.5 Ghz (C/C++)	<input type="checkbox"/>
4	MG			9.93	1.68	7.99	10.63	0.1 s	1 core @ 2.5 Ghz (C/C++)	<input type="checkbox"/>
5	URCD-Depth		code	10.03	1.74	8.24	10.71	0.1 s	1 core @ 2.5 Ghz (C/C++)	<input type="checkbox"/>
6	PLDepth			10.14	1.75	8.16	10.84	0.1s s	1 core @ 2.5 Ghz (C/C++)	<input type="checkbox"/>
7	BinsFormer		code	10.14	1.69	8.23	10.90	0.1 s	1 core @ 2.5 Ghz (C/C++)	<input type="checkbox"/>
Z. Li, X. Wang, X. Liu and J. Jiang: <a href="#">BinsFormer: Revisiting Adaptive Bins for Monocular Depth Estimation</a> . arXiv preprint arXiv:2204.00987 2022.										
8	TrapNet			10.15	1.66	7.92	10.45	0.1 s	1 core @ 2.5 Ghz (Python)	<input type="checkbox"/>
9	PixelFormer			10.28	1.82	8.16	10.84	0.1 s	1 core @ 2.5 Ghz (Python)	<input type="checkbox"/>
A. Agarwal and C. Arora: <a href="#">Attention Attention Everywhere: Monocular Depth Prediction with Skip Attention</a> . WACV 2023.										
10	RED-T			10.36	1.92	8.11	10.82	0.1 s	GPU @ 2.5 Ghz (Python)	<input type="checkbox"/>
11	NeWCRFs			10.39	1.83	8.37	11.03	0.1 s	1 core @ 2.5 Ghz (Python)	<input type="checkbox"/>
W. Yuan, X. Gu, Z. Dai, S. Zhu and P. Tan: <a href="#">NeWCRFs: Neural Window Fully-connected CRFs for Monocular Depth Estimation</a> . Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2022.										
12	DepthFormer		code	10.46	1.82	8.54	11.17	0.1 s	1 core @ 2.5 Ghz (Python)	<input type="checkbox"/>
Z. Li, Z. Chen, X. Liu and J. Jiang: <a href="#">DepthFormer: Exploiting Long-Range Correlation and Local Information for Accurate Monocular Depth Estimation</a> . arXiv preprint arXiv:2203.14211 2022.										

Figure 2: Quantitative results on the KITTI depth online benchmark. (Initial name of ADU-Depth is “ZongDepth” on the online server.)

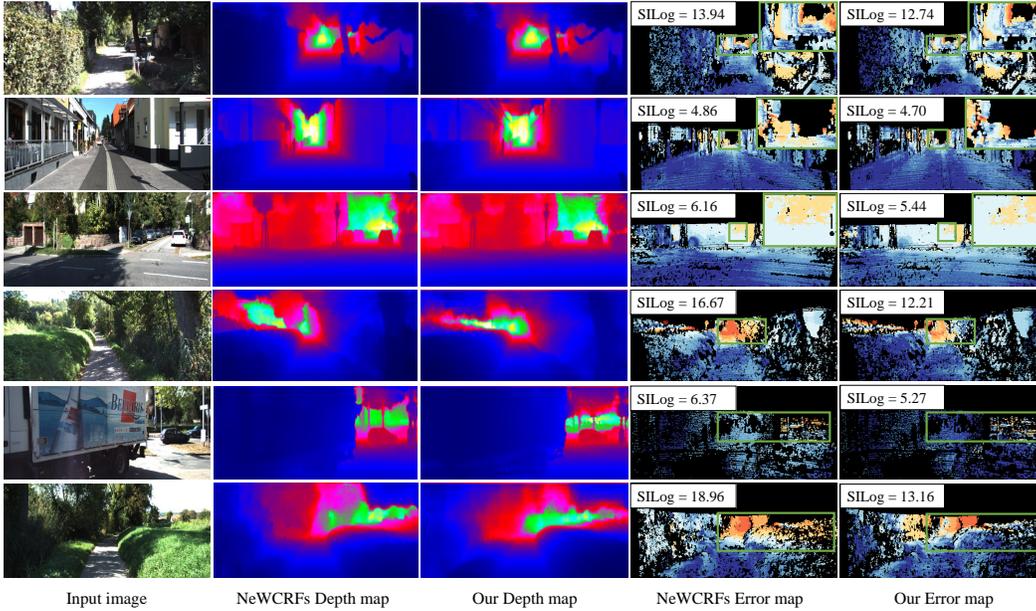


Figure 3: Qualitative results on the test set of KITTI depth online benchmark.

29 We train our model on the Eigen split training data [12], where the left-right image pairs are used to  
30 train the teacher expert model. Only the student network and monocular testing images are employed  
31 during the reference stage. The table from the KITTI evaluation server demonstrates that our method  
32 outperforms NeWCRFs [8] and other public top performers in monocular depth estimation.

33 We also display more qualitative results of NeWCRFs and our method on the KITTI depth bench-  
34 mark, including the colorful visual predictions of the depth map and error map. It can be seen  
35 in Fig. 3 that our method estimates the depth more accurately and noticeably reduces the main  
36 ranking metric “SILog” error, especially for distant and hard-to-predict image regions. Our pro-  
37 posed attention-adapted feature distillation and focal-depth adapted response distillation effectively  
38 transfer the learned 3D-aware knowledge from the teacher network to the student network. Fur-  
39 thermore, the introduction of uncertainty modeling enhances the learning for distant regions and

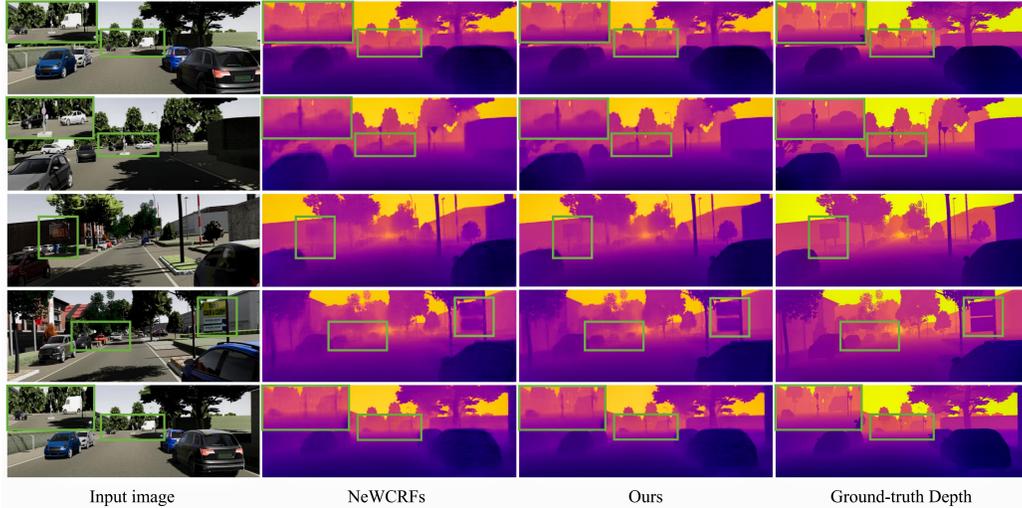


Figure 4: More qualitative results on the Virtual KITTI 2 dataset.

40 hard-to-predict image regions, such as for repetitive textures and low-light conditions, which are  
 41 usually accompanied by high uncertainty. Note that the Eigen split has a cap of 0-80m while the  
 42 KITTI online benchmark extends beyond that with 80m+ distances [13]. As a result, our method  
 43 achieves more performance gains over the KITTI Eigen split on the KITTI online benchmark due to  
 44 better depth prediction in distant image regions.

45 **Evaluation on Virtual KITTI 2.** We further compare our method with several top performers on  
 46 the virtual KITTI 2 dataset. We use a subset of the virtual KITTI 2, which contains 1,700 image  
 47 pairs for training and 193 images only for testing. The quantitative results are shown in Table 2 and  
 48 Fig. 4. Notably, our ADU-Depth achieves significantly better performance on all evaluation metrics  
 49 and estimates more accurate depth for distant regions and object contours compared to the previous  
 50 state-of-the-art method.

51 We further provide qualitative results of NeWCRFs and our method on Virtual KITTI 2 in Fig. 4.  
 52 Our method also achieves better depth estimation than the baseline method on this photo-realistic  
 53 synthetic video dataset. In particular, our ADU-Depth predicts more accurate depth maps and recov-  
 54 ers more object details in distant regions and hard-to-predict regions, such as the outline of distant  
 55 vehicles and road signs.

## 56 References

- 57 [1] Z. Yang, W. Dong, X. Li, J. Wu, L. Li, and G. Shi. Self-feature distillation with uncertainty modeling for  
 58 degraded image recognition. In *European Conf. on Computer Vision (ECCV)*, 2022.
- 59 [2] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow. Digging into self-supervised monocular depth  
 60 estimation. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 3828–3838, 2019.
- 61 [3] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia. On the uncertainty of self-supervised monocular depth  
 62 estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3227–3237, 2020.
- 63 [4] L. Mi, H. Wang, Y. Tian, H. He, and N. Shavit. Training-free uncertainty estimation for dense regression:  
 64 Sensitivity as a surrogate. In *AAAI Conf. on Artificial Intell. (AAAI)*, volume 36, 2022.
- 65 [5] J. Hornauer and V. Belagiannis. Gradient-based uncertainty for monocular depth estimation. In *European  
 66 Conf. on Computer Vision (ECCV)*, 2022.
- 67 [6] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular  
 68 depth estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2002–2011,  
 69 2018.

- 70 [7] S. F. Bhat, I. Alhashim, and P. Wonka. AdaBins: Depth estimation using adaptive bins. In *IEEE Conf. on*  
71 *Computer Vision and Pattern Recognition (CVPR)*, pages 4009–4018, 2021.
- 72 [8] W. Yuan, X. Gu, Z. Dai, S. Zhu, and P. Tan. Newcrfs: Neural window fully-connected crfs for monocular  
73 depth estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- 74 [9] E. Ilg, Çiçek, S. Galesso, A. Klein, O. Makansi, F. Hutter, and T. Brox. Uncertainty estimates and multi-  
75 hypotheses networks for optical flow. In *European Conf. on Computer Vision (ECCV)*, pages 677–693,  
76 2018.
- 77 [10] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark  
78 suite. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012.
- 79 [11] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtual worlds as proxy for multi-object tracking analysis. In  
80 *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4340–4349, 2016.
- 81 [12] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep  
82 network. *Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, 27, 2014.
- 83 [13] V. Patil, C. Sakaridis, A. Liniger, and L. Van Gool. P3depth: Monocular depth estimation with a piecewise  
84 planarity prior. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1610–1621,  
85 2022.