

Supplementary Materials: SM⁴Depth: Seamless Monocular Metric Depth Estimation across Multiple Cameras and Scenes by One Model

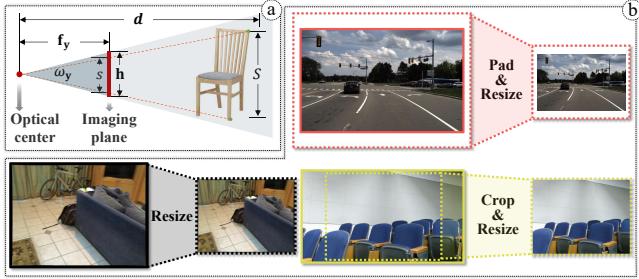


Figure 1: Principle and examples of FOV Alignment. (a) Side view of the pinhole camera model. (b) Cases of FOV alignment.

1 MORE DETAILS OF SM⁴DEPTH

1.1 Role of FOV in reducing metric ambiguity

According to Metric3D [24], due to different intrinsic parameters, two cameras produce different projections when observing an object at the same distance, which is well known as “**metric ambiguity**”. Next, we investigate the key of eliminating metric ambiguity by Fig. 1 that illustrates the imaging process of the pinhole camera model. Assuming that d denotes the depth of the object, and f_y denotes the focal length in the Y direction of the camera, measured in pixels. According to the similarity principle, there is an equation:

$$\frac{d}{S} = \frac{f_y}{s} \quad (1)$$

where S and s are the actual height (in millimeters) and the imaging height (in pixels) of the object respectively. On the basis of Eq. (1), the object’s depth can be formulated as $d = S \left[\frac{f_y}{s} \right]$. Therefore, a fixed value of $\left[\frac{f_y}{s} \right]$ is crucial for a consistent depth d between different cameras. In practice, all images need to be resized into the same resolution before being fed into the deep network:

$$d = S \left[\frac{(f'_y/f_y)f_y}{(h'/h)s} \right] \quad (2)$$

where f'_y and h' are the focal length and height of the network input, h is the original height of the image, and $(f'_y/f_y) = (h'/h)$. Note that, since f'_y and h' are two pre-set values, the consistency of $\frac{h}{f_y}$ ensures a consistent depth d across different cameras. Furthermore, the value $\frac{h}{f_y}$ follows an arc-tangent function relationship with the camera’s vertical FOV denoted as ω_y :

$$\omega_y = 2 \arctan\left(\frac{h}{2f_y}\right) \quad (3)$$

Thus, the consistency of ω_y is essential for consistent depth and eliminating metric ambiguity across different cameras. The same applies to the horizontal FOV indicated as ω_x . Unlike transforming

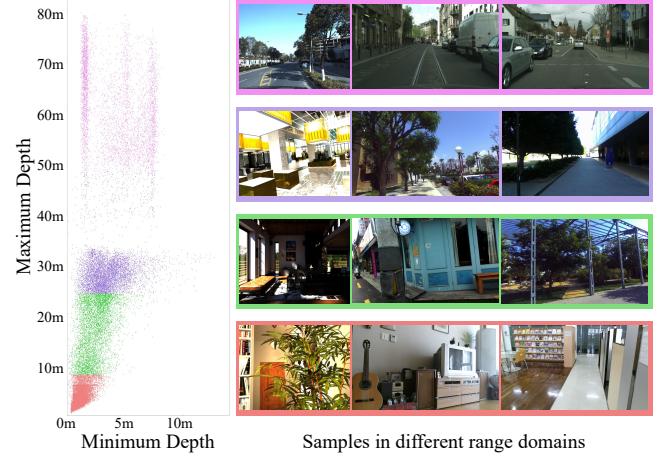


Figure 2: Range domain (RD) visualization. Different colors on the scatter plot represent images within different RDs.

images to the same intrinsic parameter [24], our method ensures consistent inputs by unifying the FOV of images.

1.2 Visualization of images in different RDs

Fig. 2 visualizes the RDs in different colors with $K = 4$. It can be observed that the images of the same RD exhibit different appearances but similar depth ranges.

1.3 Loss Function of SM⁴Depth

Our network is supervised by multiple loss functions. As defined in the main paper, D represents the ground truth depth map, while D' signifies the predicted depth map. D'_s denotes the depth map of the s^{th} stage in the HSC-decoder. Additionally, $c \in \mathbb{R}^{N \times 1}$ refers to the combined metric bin centers, and $y \in \mathbb{R}^{K \times 1}$ signifies the generated probabilities for K RDs. At the pixel level, we employ scale-invariant logarithmic (Silog) loss [8] to minimize per-pixel depth errors:

$$\mathcal{L}_{silog}(D, D') = \sqrt{\frac{1}{T} \sum_{i=1}^T \left(\ln \frac{D(i)}{D'(i)} \right)^2 - \frac{\lambda}{T^2} \left(\sum_{i=1}^T \ln \frac{D(i)}{D'(i)} \right)^2} \quad (4)$$

where T denotes the number of pixels with valid ground truth values, and we set $\lambda = 0.85$. Then, we employ a multi-scale gradient matching term [16] to supervise the discontinuities between pixels in the depth map:

$$\mathcal{L}_{grad}(D, D') = \frac{1}{T} \sum_{m=1}^M \sum_{i=1}^T (|\nabla_x G_i^m| + |\nabla_y G_i^m|) \quad (5)$$

where $G_i = D(i) - D'(i)$. G^m denotes the difference in disparity maps at scale m , and $M = 4$ is the scale level. Overall, the pixel-wise

Training Datasets	Scene	Capture	# Img	Range(m)
ScanNet [7]	Indoor	RGB-D	24,834	[0, 10]
Hypersim [17]	Indoor	Synthetic	15,229	[0, 80]
DIML Sample [5]	Indoor	RGB-D	1,609	[0, 6]
DIML Indoor [5]	Indoor	RGB-D	26,039	[0, 8]
DIML Outdoor [5]	Outdoor	Stereo* [#]	24,031	[0, 35]
UASOL [1]	Outdoor	Stereo* [#]	36,386	[0, 35]
ApolloScape [11]	Outdoor	LiDAR [#]	14,908	[0, 80]
Cityscapes [6]	Outdoor	Stereo* [#]	11,486	[0, 80]
<i>total</i>			154,522	
Validation Datasets	Scene	Capture	# Img	Range(m)
NYUD [14]	Indoor	RGB-D	654	[0, 10]
KITTI [20]	Outdoor	LiDAR	652	[0, 80]
<i>total</i>			1,306	
Test Datasets	Scene	Capture	# Img	Range(m)
SUN RGB-D [19]	Indoor	RGB-D	4,395	[0, 8]
iBims-1 [12]	Indoor	LiDAR	100	[0, 50]
ETH3D Indoor [18]	Indoor	LiDAR	219	[0, 56]
DIODE Indoor [21]	Indoor	LiDAR	325	[0, 50]
nuScenes-val [3]	Outdoor	LiDAR	1,138	[0, 140]
DDAD [10]	Outdoor	LiDAR	3,950	[0, 80]
ETH3D Outdoor [18]	Outdoor	LiDAR	235	[0, 73]
DIODE Outdoor [21]	Outdoor	LiDAR	446	[0, 80]
BUPT Depth	Both-continuous	Stereo*[#]	14,932	[0, 30]
<i>total</i>			25,740	

Table 1: The metric depth datasets used in this paper. * denotes the re-generation of metric depth by CRESTereo [13], and # indicates the removal of sky regions in ground truth by ViT-Adapter [4].

loss function can be formulated as follows:

$$\mathcal{L}_{pixel}(D, \mathbf{D}) = \alpha L_{silog}(D, \mathbf{D}) + \beta L_{grad}(D, \mathbf{D}) \quad (6)$$

where the coefficients $\alpha = 10$ and $\beta = 0.5$ are used in Eq.(6), learning the depth primarily and recovering the depth boundary secondarily. Then, the virtual normal loss [23] is employed to optimize the 3D structure:

$$\mathcal{L}_{vnl}(D, \mathbf{D}) = \frac{1}{V} \sum_{i=1}^V (\| n_i - \mathbf{n}_i \|_1) \quad (7)$$

where $V = 0.15T$ is the sampling number of virtual planes. n_i is the normal vector of the i^{th} virtual plane in the output D and \mathbf{n}_i corresponds to the normal vector in \mathbf{D} .

At the scene level, the bi-directional Chamfer Loss [9] is employed to optimize the combination of bin centers \mathbf{c} , making them closer to the ground truth \mathbf{D} as shown in Eq.(7) of the main paper: $\mathcal{L}_{bin}(\mathbf{c}, \mathbf{D})$, with $N = 256$ bins. Furthermore, the cross entropy loss is applied on the classification head (CLS):

$$\mathcal{L}_{cel}(y, \mathbf{y}) = - \sum_{k=1}^K (y_k \ln y_k) \quad (8)$$

where $\mathbf{y} \in \mathbb{R}^{K \times 1}$ is the one-hot RD label of the input image.

Finally, the total loss of SM⁴Depth can be formulated as follows:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{pixel}(D, \mathbf{D}) + \sum_{s=2}^4 (\mathcal{L}_{pixel}(D'_s, \mathbf{D})) + \\ &\mu \mathcal{L}_{vnl}(D, \mathbf{D}) + \gamma \mathcal{L}_{bin}(\mathbf{c}, \mathbf{D}) + \mathcal{L}_{cel}(y, \mathbf{y}) \end{aligned} \quad (9)$$

where the coefficients μ and γ are empirically set to 5 and 0.1 respectively.

2 DATASETS DETAIL AND PRE-PROCESSING

Table 1 shows all datasets used for training, validation and testing. We conduct the same pre-processing operations before training as [2, 16].

Depth Re-generation: UASOL [1], CityScapes [6], and DIML [5] provide depth using a hand-crafted stereo matching method, which is not accurate enough. For this reason, we employ an advanced algorithm called CRESTereo [13] to re-generate the ground truth.

Sky Removal: The images of outdoor datasets contain large areas of the sky, such as DIML, UASOL, ApolloScape, and CityScapes. We use ViT-Adapter [4] to extract sky areas and invalidate the depth values within these regions.

3 MORE DETAILS OF BUPT DEPTH

BUPT Depth dataset differs in three ways: **Shooting style**: we employ continuous filming, allowing the camera to traverse various scenes, thus better simulating everyday use. **Scene variety**: besides common scenes like streets and corridors, our dataset includes unique environments like patios, building underpasses, and irregular landscapes, etc. **Evaluation focus**: BUPT Depth prioritizes consistent accuracy across various environments, rather than the precision of individual images.

For reliability, we follow the precedent set by ZoeDepth using the DIML dataset and use the ZED2 camera to capture depth for evaluation. According to its official datasheet, ZED2 maintains an error rate of less than 1% within 3 meters and less than 5% within 15 meters across all scenes, significantly outperforming existing multimodal depth estimation (MMDE) systems. However, ZED2 outputs often exhibit holes and imprecise edges, which we mitigate using CRESTereo.

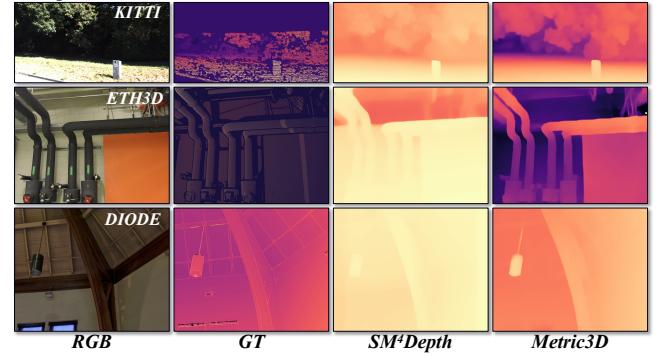


Figure 5: Failure cases sampled from KITTI, ETH3D, DIODE.

RGB-D Pairs	SUN RGB-D	ETH3D	DIODE	DDAD	mRI $_{\eta}$ ↑
50K	0.391	2.527	5.655	6.159	-11.9%
100K	0.333	2.480	5.766	5.374	-3.62%
150K	0.310	2.373	5.605	5.390	0.00%
300K	0.313	2.607	5.946	5.502	-4.74%

Table 2: RMSE of models trained on different numbers of images and datasets. Gray indicates the setting used for our model.

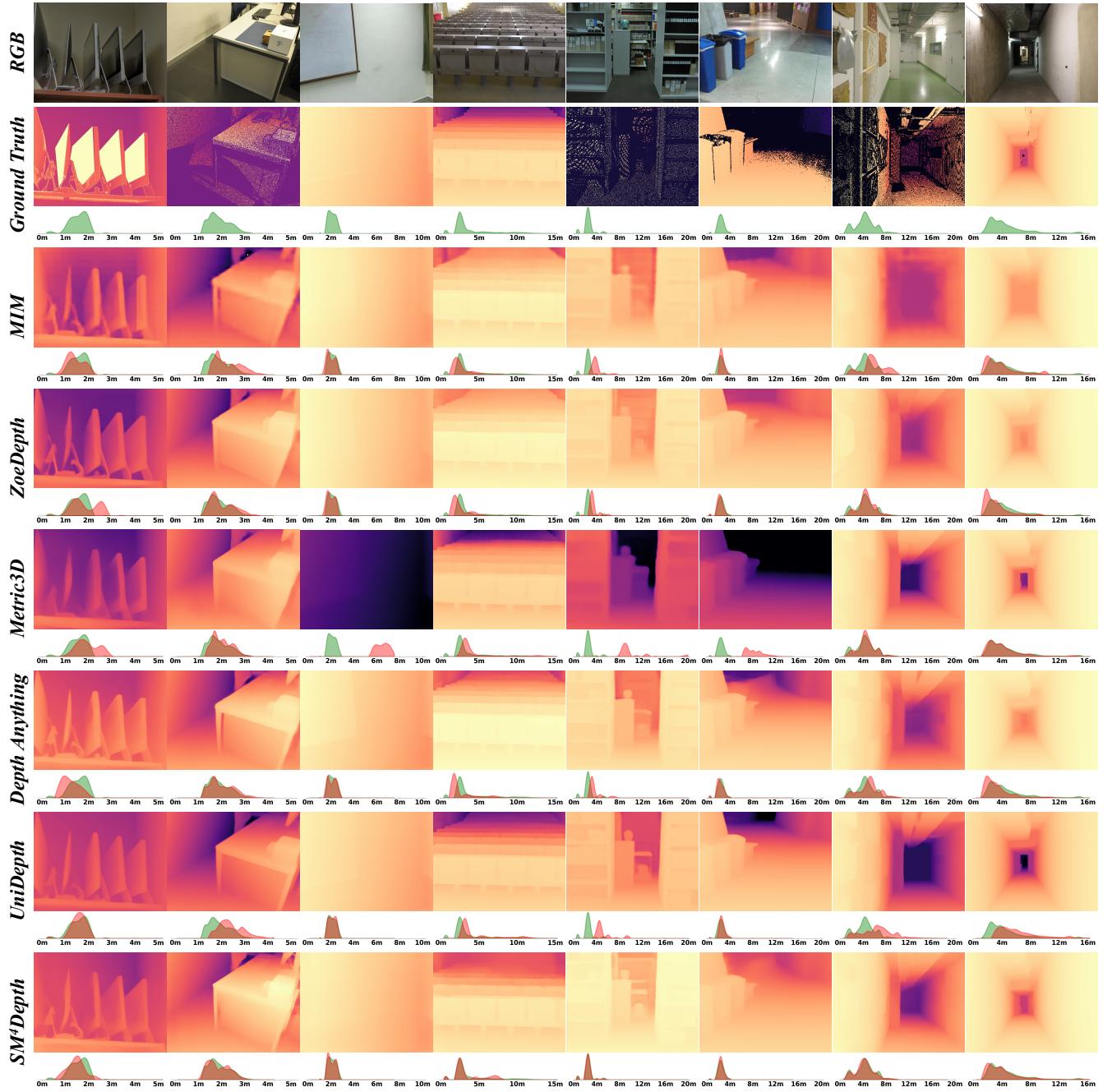


Figure 3: Qualitative results in indoor scenes. The depth distribution is under the depth maps with green for ground truth and red for prediction. Our method performs better on images with multiple viewpoints and diverse scenes.

4 ADDITIONAL EXPERIMENTS

4.1 Ablation about Training Data Amount

Table 2 gives the ablation study on the amount of training data. As the data increases from 50K to 150K, the RMSE decreases, while increases at 300K. The reason behind the increased RMSE is that when the number of images is less than 150K the number of data

in each range domain is balanced, which is broken unexpectedly when the data size is larger than 150K. The indoor datasets lack sufficient data, causing the data amount gap between different RDs to be widened.

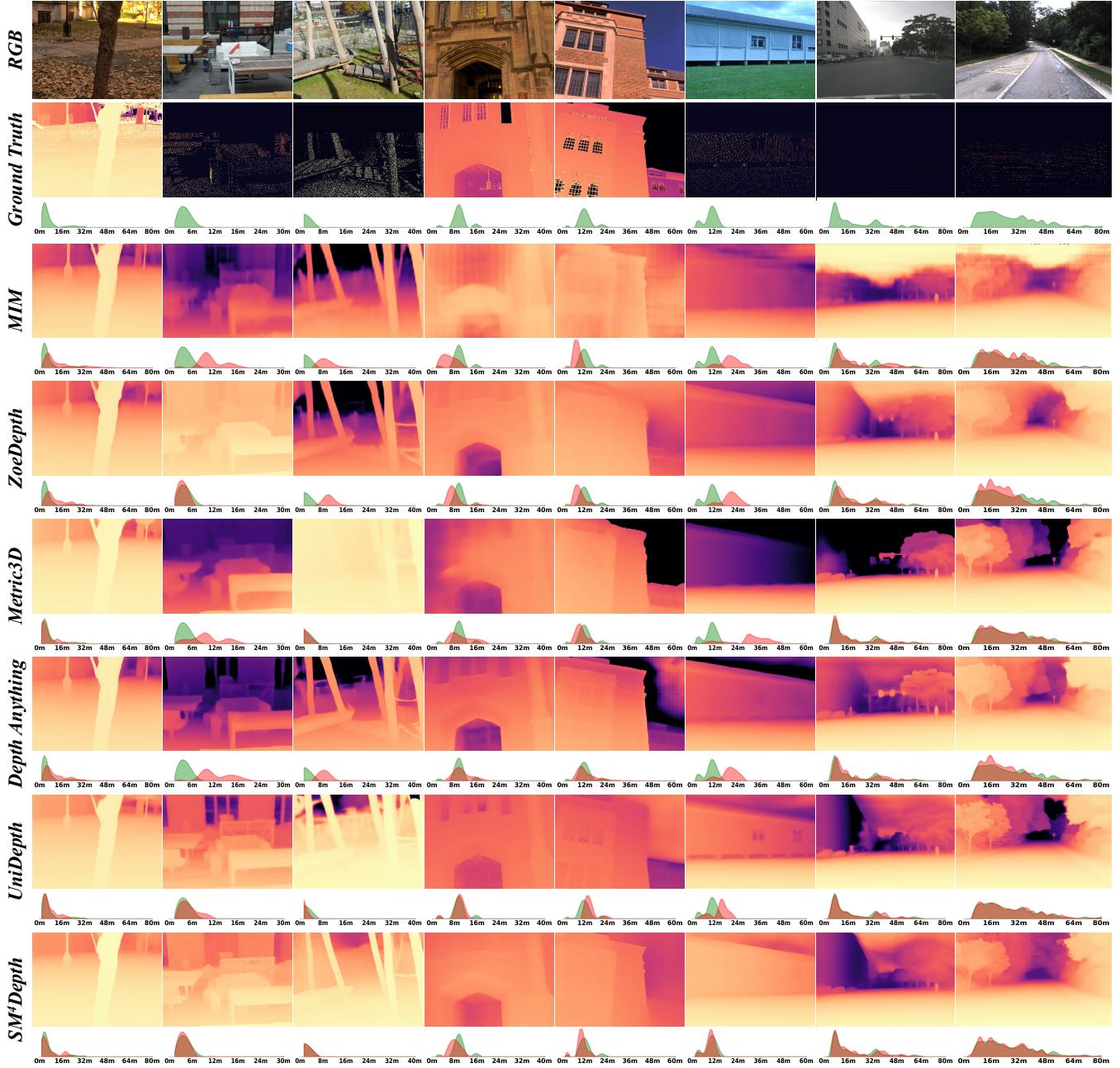


Figure 4: Qualitative results in outdoor scenes. The depth distribution is under the depth maps with green for ground truth and red for prediction. Our method performs better on images with multiple viewpoints and diverse scenes.

4.2 More Visualization Comparison

Fig. 3 and Fig. 4 gives more qualitative comparisons of indoor and outdoor scenes, respectively. Compared with Metric3D [24], DepthAnything [22], and Zoedepth [2], SM⁴Depth can generally obtain more accurate metric scales. Compared with UniDepth [15], our method achieves comparable accuracy in metric scale estimation, but obtains sharper depth.

5 FAILURE CASES

As shown in Fig. 5, describing the limited robustness for turning view in street scenes and rare camera poses. In the ETH3D and DIODE examples, Metric3D, although limited, outperforms SM⁴Depth in depth range due to its higher diversity of camera pose during training.

REFERENCES

- [1] Zuria Bauer, Francisco Gomez-Donoso, Edmanuel Cruz, Sergio Orts-Escobar, and Miguel Cazorla. 2019. UASOL, a large-scale high-resolution outdoor stereo dataset. *Scientific data* 6, 1 (2019), 162.
- [2] Sharif Farooq Bhat, Reiner Birk, Diana Wofk, Peter Wonka, and Matthias Müller. 2023. ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth. <https://doi.org/10.48550/ARXIV.2302.12288>
- [3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liang, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [4] Zhe Chen, Yuchen Duan, Wenhui Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. 2022. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534* (2022).
- [5] Jaehoon Cho, Dongbo Min, Youngjung Kim, and Kwanghoon Sohn. 2021. DIMI/CVL RGB-D dataset: 2M RGB-D images of natural indoor and outdoor scenes. *arXiv preprint arXiv:2110.11590* (2021).
- [6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [7] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [8] David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network. In *Neural Information Processing Systems (NIPS)*.
- [9] Haoqiang Fan, Hao Su, and Leonidas Guibas. 2017. A Point Set Generation Network for 3D Object Reconstruction from a Single Image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [10] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 2020. 3D Packing for Self-Supervised Monocular Depth Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [11] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqiang Lin, and Ruigang Yang. 2018. The apolloscape dataset for autonomous driving. In *IEEE Conference on Computer Vision and Pattern Recognition workshops (CVPR workshop)*.
- [12] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. 2018. Evaluation of CNN-based single-image depth estimation methods. In *European Conference on Computer Vision (ECCV) Workshops*.
- [13] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. 2022. Practical stereo matching via cascaded recurrent networks with adaptive correlation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [14] Pushmeet Kohli, Nathan Silberman, Derek Hoiem, and Rob Fergus. 2012. Indoor Segmentation and Support Inference from RGBD Images. In *European Conference on Computer Vision (ECCV)*.
- [15] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. 2024. UniDepth: Universal Monocular Metric Depth Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [16] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence* 44, 3 (2020), 1623–1637.
- [17] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. 2021. Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. In *International Conference on Computer Vision (ICCV)*.
- [18] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. 2017. A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [19] Shuruan Song, Samuel P Lichtenberg, and Jianxiong Xiao. 2015. SUN RGB-D: A rgb-d scene understanding benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [20] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. 2017. Sparsity Invariant CNNs. In *International Conference on 3D Vision (3DV)*.
- [21] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. 2019. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *CoRR* abs/1908.00463 (2019). <http://arxiv.org/abs/1908.00463>
- [22] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [23] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. 2019. Enforcing geometric constraints of virtual normal for depth prediction. In *International Conference on Computer Vision (ICCV)*, 5684–5693.
- [24] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. 2023. Metric3D: Towards Zero-shot Metric 3D Prediction from A Single Image. In *IEEE International Conference on Computer Vision (ICCV)*.