Supplementary Material for M^{3}PT

Zhiqiang Yan*, Xiang Li*, Kun Wang, Zhenyu Zhang,
Jun Li✉, and Jian Yang✉

PCA Lab, Nanjing University of Science and Technology, China
{Yanzq,xiang.li.implus,kunwang,junli,csjyang}@njust.edu.cn
zhangjesse@foxmail.com
* Equal contribution

Config	Value	
	pre-training	fine-tuning
optimizer	AdamW [4]	AdamW
learning rate schedule	CosineAnnealingLR [3]	MultiStepLR
learning rate	0.001	0.001
weight decay	0.01	0.01
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$	$\beta_1, \beta_2 = 0.9, 0.999$
batch size	92	92
augmentation	no	no

Table S1. Pre-training and fine-tuning settings.

A Overview

This document provides additional technical details and more visualization results. Specifically, in section B, we first describe the implementation details of M^{3}PT. Next, we conduct additional experiments in section C. Finally, in section D, we show more visualizations of the depth prediction and 3D reconstruction on Matterport3D, Stanford2D3D, and 3D60 datasets.

B Implementation Details

Backbone architecture. We employ GuideNet [5] as the default architecture. It consists of two hourglass units, one for encoding RGB image input and another for modeling sparse depth input. Synchronized Cross-GPU Batch Normalization [1] is deployed. M^{3}PT does *not* require any network structure adjustment when fine-tuning compared to pre-training.

Pre-training. The default setting is reported in Table S1. Our M^{3}PT does *not* use additional data augmentation, including the color jittering and random horizontal flip conducted in GuideNet. We use cosine annealing learning rate strategy [3] with default 'T_max=20, eta_min=0.00001, last_epoch=-1'.

Fine-tuning. The setting of fine-tuning is very similar with that of pre-training apart from the learning rate strategy with default 'milestons=50, 100, 150, gamma=0.5, last_epoch=-1', which is reported in Table S1.

Dataset		pre-training & fine-tuning		
		Stanford2D3D	Matterport3D	3D60
testing	Stanford2D3D	**167.7**	159.4	248.5
	Matterport3D	254.7	**146.2**	224.3
	3D60	262.6	151.9	**142.3**

Table S2. Generalization capability of M^{3}PT. We train it on one dataset but test on all three datasets. Best RMSE results of each column are marked in bold. The mask size and ratio on three datasets are 16 and 0.75, respectively.

C Additional Experiments

C.1 Testing on Different Datasets

Table S2 lists the generalization capability of M^{3}PT. We pre-train and fine-tune M^{3}PT on single dataset, but test it on all three datasets. In each column of Table S2, the performances of testing on other dataset is worse due to the different data distribution. We can also find that the generalization capability is weaker when training on Stanford2D3D, which is probably caused by the least amount of data and incomplete pixels of color images nearby the top and bottom areas. 3D60 dataset contains parts of Stanford2D3D and Matterport3D, which are potentially hard for models to generalize very well.

C.2 Reconstruction under Different Pre-Training Settings

Figure S2 demonstrates the reconstruction results of M^{3}PT during pre-training on Stanford2D3D. As we can see that the model can partly recover depth clues given masked RGB-D pair and supervised by invisible areas of sparse depth, although the recovery is different from ground-truth depth annotation.

Figure S3 shows the generalization of M^{3}PT which is pre-trained on Stanford2D3D but transferred to 3D60. Though the reconstruction results are far away from the ground-truth, it is probably plausible.

In addition, we also explore the generalization capability of M^{3}PT under different mask settings. As shown in Figure S4, on Stanford2D3D, we test the model pre-trained with mask size 16 and ratio 0.75 on other mask sizes and ratios. We can find that the model can also recover depth clues in certain degree.

C.3 Testing with Different Data Patterns

Table S3 reports the performances of M^{3}PT with different data synthesis patterns. Note that the sparse depths of all patterns contain 6144 valid points, accounting for about 4.69% of the entire maps, shown in Figure S1. As we can see from Table S3, the scanning pattern can help to recover better depth, indicating it's reasonable to synthesize sparse depths by imitating laser scanning.

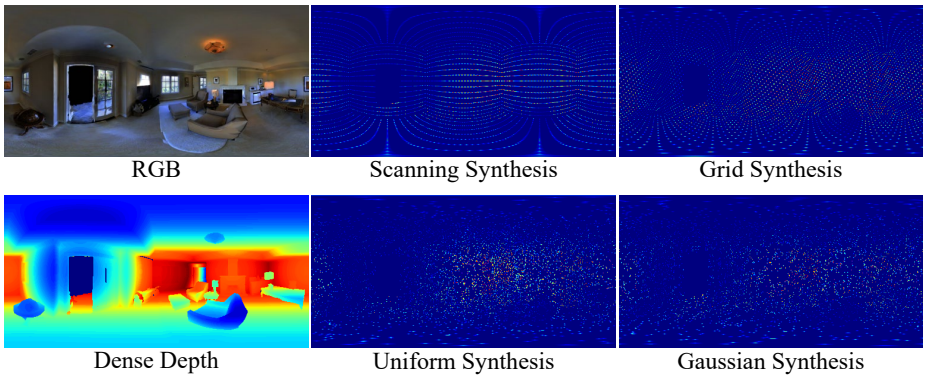

Fig. S1. Different patterns of data synthesis, including scanning, uniform, gaussian, and grid. Note that original cubical sampling results have been projected into equirectangular maps.

Pattern	Error Metric ↓				Accuracy Metric ↑		
	MRE	MAE	RMSE	RMSElog	$\boldsymbol{\delta_{1.25}}$	$\boldsymbol{\delta_{1.25^2}}$	$\boldsymbol{\delta_{1.25^3}}$
Scanning	**0.0274**	**52.9**	**149.0**	**0.0263**	**0.9859**	0.9963	0.9988
Uniform	0.0598	121.3	227.7	0.0398	0.9764	0.9955	0.9984
Gaussian	0.0563	115.4	221.3	0.0384	0.9783	0.9957	0.9984
Grid	0.0511	104.6	195.8	0.0335	0.9851	**0.9972**	**0.9991**

Table S3. Generalization capability of M^{3}PT on Stanford2D3D under different data synthesis patterns, which are shown in Figure S1.

D More Visualizations

This section shows more visualizations on three datasets. As illustrated in Figure S5, it is obviously that our M^{3}PT can produce better depth predictions and 3D reconstructions than other related methods.

References

1. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. pp. 448–456. PMLR (2015) 1
2. Jiang, H., Sheng, Z., Zhu, S., Dong, Z., Huang, R.: Unifuse: Unidirectional fusion for 360 panorama depth estimation. IEEE Robotics and Automation Letters **6**(2), 1519–1526 (2021) 7
3. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. In: ICLR (2017) 1
4. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019) 1
5. Tang, J., Tian, F.P., Feng, W., Li, J., Tan, P.: Learning guided convolutional network for depth completion. IEEE Transactions on Image Processing **30**, 1116–1129 (2020) 1, 7

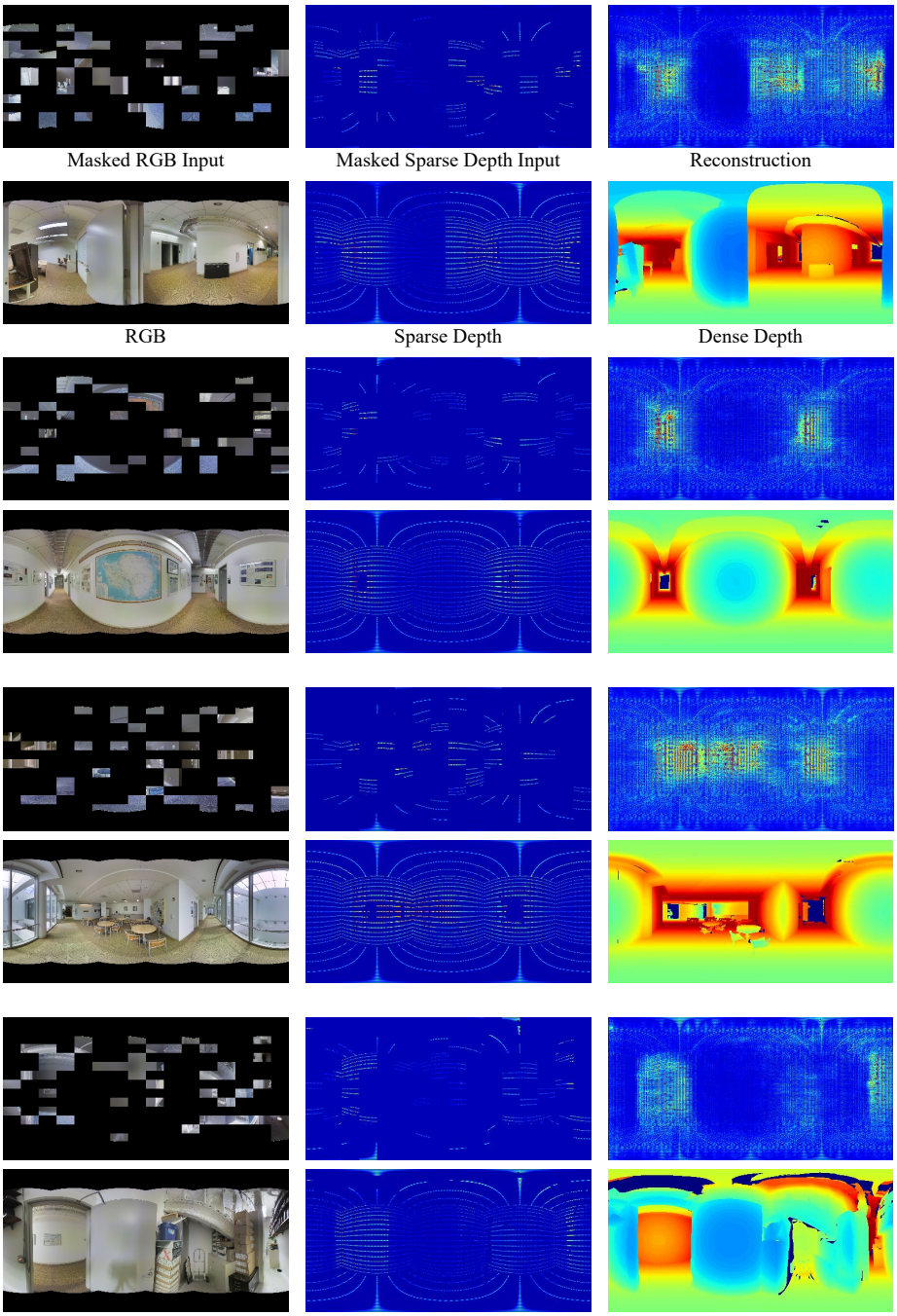

Fig. S2. Example of M^{3}PT reconstructions during pre-training on **Stanford2D3D** testing split. The mask size is 16 and the mask ratio is 0.75. Best color of view.

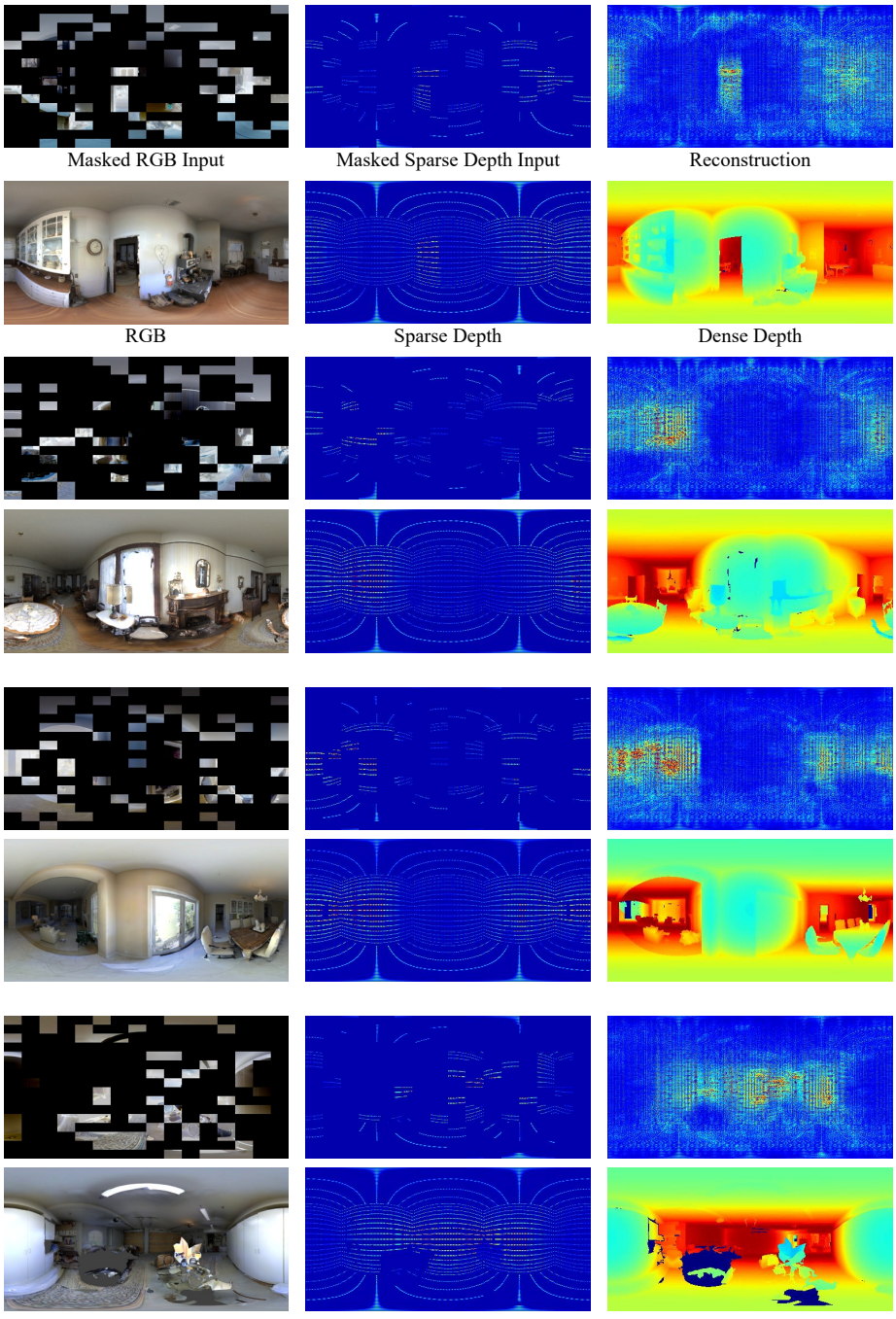

Fig. S3. Example of M^{3}PT reconstructions during pre-training on **3D60** testing split, using the M^{3}PT pre-trained on Stanford2D3D (the same model weights as in Figure S2).

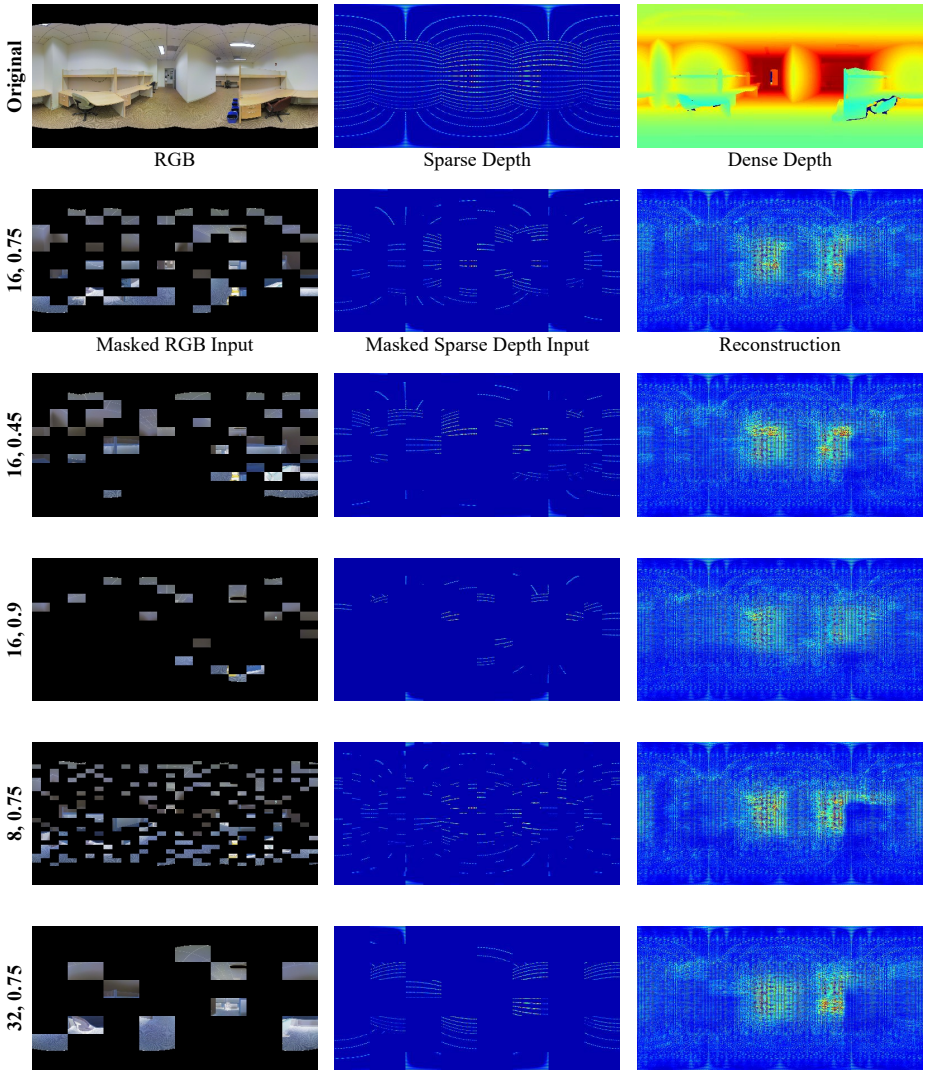

Fig. S4. Example of M^{3}PT reconstructions during pre-training on Stanford2D3D testing split, using the M^{3}PT with mask size 16 and mask ratio 0.75 that is pre-trained on Stanford2D3D (the same model weights as in Figure S2), but applied on inputs with different mask sizes and mask ratios.

Fig. S5. Qualitative comparison of different methods, including UniFuse [2], GuideNet [5], and our M^{3}PT on Matterport3D (the first four rows), Stanford2D3D (the second four rows), and 3D60 (the third four rows) datasets.