# Multi-Modal Masked Pre-Training for Monocular Panoramic Depth Completion

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

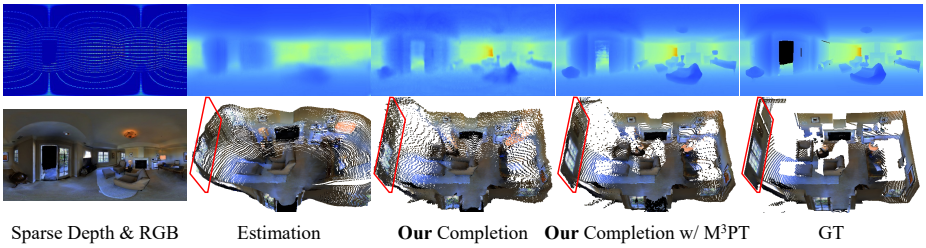


**Fig. 1.** Comparisons of the predicted depth and 3D reconstruction results between panoramic depth estimation (with RGB) and completion (with RGB and sparse depth).

**Abstract.** In this paper, we formulate a potentially valuable panoramic depth completion (PDC) task as panoramic 3D cameras often produce 360° depth with missing data in complex scenes. Its goal is to recover dense panoramic depths from raw sparse ones and panoramic RGB images. To deal with the PDC task, we train a deep network that takes both depth and image as inputs for the dense panoramic depth recovery. However, it needs to face a challenging optimization problem of the network parameters due to its non-convex objective function. To address this problem, we propose a simple yet effective approach termed M$^3$PT: multi-modal masked pre-training. Specifically, during pre-training, we simultaneously cover up patches of the panoramic RGB image and sparse depth by shared random mask, then reconstruct the sparse depth in the masked regions. To our best knowledge, it is the first time that we show the effectiveness of masked pre-training in a multi-modal vision task, instead of the single-modal task resolved by masked autoencoders (MAE). Different from MAE where fine-tuning completely discards the decoder part of pre-training, there is no architectural difference between the pre-training and fine-tuning stages in our M$^3$PT as they only differ in the prediction density, which potentially makes the transfer learning more convenient and effective. Extensive experiments verify the effectiveness

of M$^3$PT on three panoramic datasets. Notably, we improve the state-of-the-art baselines by averagely 29.2% in RMSE, 51.7% in MRE, 49.7% in MAE, and 37.5% in RMSElog on three benchmark datasets.

**Keywords:** 360° depth completion, multi-modal masked pre-training, network optimization, shared random mask, 3D reconstruction

## 1 Introduction

Panoramic depth perception (see Table 1) has received increasing attention in both academic and industrial communities due to its crucial role in a wide variety of downstream applications, such as virtual reality [1], scene understanding [46], and autonomous navigation [17]. With the development of hardware devices, panoramic 3D cameras become easier and cheaper to capture both RGB and depth (RGB-D) data with 360° field of view (FoV). Depending on the captured RGB images, all recent perception technologies [37,46,23,31,63,67,3,42,41], to the best of our knowledge, concentrate on panoramic depth estimation (PDE) that predicts dense depth from a single 360° RGB image. In this paper, we focus on exploring the 360° RGB-D pairs for the panorama perception with an effective pre-training strategy. We show our motivations as follows:

**Two Motivations for Panoramic Depth Completion** (PDC). *One is the 360° depth maps with missing areas.* During the collection process, the popular panoramic 3D cameras (*e.g.*, Matterport Pro2[1] and FARO Focus[2]) still produce 360° depth with missing areas when facing bright, glossy, thin or far surfaces, especially indoor rooms in Figure 2. These depth maps will result in a poor panorama perception. To overcome this problem, we consider a new panoramic depth completion task, completing the depth channel of a single 360° RGB-D pair captured from a panoramic 3D camera. *Another is an experimental investigation that in contrast with PDE, PDC is much fitter for the panoramic depth perception.* For simplicity and fairness, we directly employ the same network architectures (*e.g.*, UniFuse [23]) to estimate or complete the depth map. Figure 7 reports that the PDC has much lower root mean square error than PDE. Furthermore, Figure 1 shows that the PDC can recover more precise 360° depth values, leading to better 3D reconstruction. This observation reveals that PDC is more important than PDE for 3D scene understanding.

**One Motivation for Pre-Training**. When using deep networks to perceive the depth information, there is a challenging problem: *how to optimize the parameters of the deep networks*? It is well-known that the objective function is highly non-convex, resulting in many distinct local minima in the network parameter space. Although the completion can lead to better network parameters and higher accuracy on the depth perception, it is still not to satisfy practical needs of the 3D reconstruction. Here, we are inspired by the greedy layer-wise pre-training technology [14,30] that stacks two-layer unsupervised autoencoders

[1] https://matterport.com/cameras/pro2-3D-camera
[2] https://www.faro.com/en/Products/Hardware/Focus-Laser-Scanners

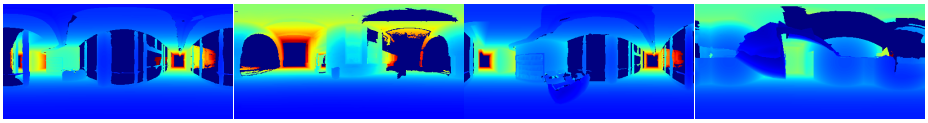

**Fig. 2.** Panoramic depth maps with large missing areas shown in darkest blue color.

to initialize the networks to a point in parameter space, and then fine-tunes them in a supervised setting. This technology drives the optimization process more effective, achieving a 'good' local minimum. Recently, the single-modal masked autoencoders [19,58] are also applied into object detection and semantic segmentation, achieving amazing improvements on their benchmarks. These interpretations and improvements motivate us to explore a new pre-training technology for the multi-modal panoramic depth completion.

In this paper, we propose a Multi-Modal Masked Pre-Training (M$^3$PT) technology to directly initialize all parameters of deep completion networks. Specifically, the key idea of M$^3$PT is to employ a shared random mask to simultaneously corrupt the RGB image and sparse depth, and then use the invisible pixels of the sparse depth as supervised signal to reconstruct the original sparse depth. After this pre-training, no-masked RGB-D pairs are fed into the pre-trained network supervised by dense ground-truth depths. Different from the layer-wise pre-training [14], M$^3$PT is to pre-train all layers of the deep network. Compared to MAE [19], M$^3$PT has no architectural difference between the pre-training and fine-tuning stages, where they differ in only the prediction density of target depth. This characteristic probably makes it convenient and effective for the transfer learning, including but not limited to the multi-modal depth completion, denoising, and super-resolution guided by RGB images. In summary, our contributions are as follows:

- We introduce a new panoramic depth completion (PDC) task that aims to complete the depth channel of a single 360° RGB-D pair captured from a panoramic 3D camera. To the best of our knowledge, we are the first to study the PDC task to facilitate 360° depth perception.
- We propose the multi-modal masked pre-training (M$^3$PT) for the multi-modal vision task. Different from the layer-wise pre-training [14] and MAE [19], M$^3$PT is to pre-train all layers of the deep network, and does not change the network architecture in the pre-training and fine-tuning stages.
- On three benchmarks, *i.e.*, Matterport3D [1], Stanford2D3D [2], and 3D60 [68], extensive experiments demonstrate that (i) PDC achieves higher accuracy of panoramic depth perception than PDE, and (ii) our M$^3$PT technology achieves the state-of-the-art performance.

## 2 Related Work

Since this paper aims to learn the new task of monocular panoramic depth completion, we report three related but different topics whose detailed differences

| Task | Depth Estimation | Depth Completion | Panoramic Depth Estimation | Panoramic Depth Completion |
|---|---|---|---|---|
| New task | No | No | No | Yes |
| Data FoV | <180° | <180° | 360° | 360° |
| Data modal | RGB | RGB-D | RGB | RGB-D |
| Camera | perspective | perspective | panoramic | panoramic |
| Target | depth | depth | depth & 3D reconstruction | depth & 3D reconstruction |

**Table 1.** Comparisons of different depth related tasks. FoV denotes the field of vision.

are listed in Table 1. First, we review depth completion approaches that input single RGB-D pair with limited FoV. Second, we elaborate on panoramic depth estimation works which predict 360° depths from panoramic color images. At last, we introduce the masked image encoding technology.

### 2.1 Monocular Depth Completion with Limited FoV

Existing monocular depth completion methods primarily focus on sparse depths and color images with a narrow FoV less than 180°. Up to now, based on the commonly used KITTI [50] and NYUv2 [43] datasets, a great deal of methods have been proposed to tackle the task, which can be broadly divided into depth-only [50,10,35,13] and multi-sensor fusion based [9,8,40,61,18,62,32] categories. For example, the literatures [22,51] take sparse depths as the only input to recover dense ones without using color images. Further, Lu *et al.* [34] use color images as auxiliary supervision during training and is discarded when testing. Recently, as technology quickly develops, multi-modal information can be captured by sensors, which is beneficial for depth completion. For example, S2D [35] directly concatenate RGB-D pairs and feed them into networks, contributing to promising improvement. Li *et al.* [29] propose multi-scale guided cascade hourglass network to handle diverse patterns. PENet [21] proposes to refine depth recovery at three stages. FCFRNet [33] designs channel-shuffle technology to enhance RGB-D feature fusion. GuideNet [47] proposes dynamic convolution to adaptively generate convolution kernels according to color image contents. ACMNet [64] conducts graph propagation to extract multi-modal representations. Furthermore, DeepLiDAR [38] and PwP [59] jointly utilize color images, surface normals, and sparse depths to recover dense depth. FusionNet [51] and Zhu *et al.* [66] present to estimate uncertainty for robust recovery. NLSPN [36] and DSPN [60] introduce recurrent non-local and dynamic spatial propagation networks, which significantly improve depth accuracy nearby object boundaries.

In addition, several unsupervised depth completion works [56,25,55,54,49] also contribute to the development of this domain. For example, KBNet [57] proposes the fantastic calibrated backprojection network which achieves very superior performances. However, as mentioned above these methods are designed for dense depth recovery from FoV-limited sparse depth, whilst we aim to learn 360° depth completion and 3D reconstruction from panoramic RGB-D input.

### 2.2 Monocular Depth Estimation with Full FoV

Given panoramic color images, current monocular panoramic depth estimation works mainly devote into predicting 360° depths and 3D reconstructions. This topic springs up as soon as the large indoor panoramic datasets Matterport3D [5] and Stanford2D3D [2] are constructed in 2017. For this domain in the last five years, supervised methods play a primary role while unsupervised approaches develop slowly. Next we will introduce each of them.

Supervised category: In 2018, OmniDepth [69] synthesizes 360° data with high-quality ground-truth depth annotations by rendering existing datasets. DistConv [48] proposes distortion-aware convolutional filters to address the inherent distortion of equirectangular projection (EPR) panoramic data. In 2019, Eder *et al.* [12] utilize surface normal and plane boundaries to train a plane-aware network to benefit depth estimation. SpherePHD [27,28] explores a new data representation via spherical polyhedron, which resolves the shape distortion of spherical panoramas. In 2020, Jin *et al.* [24] and Feng *et al.* [15] use geometric priors to help with depth estimation. Wang *et al.* [53] adopt a two-branch network leveraging EPR and cubemap projections, which are the two most common data forms. In 2021, PanoDepth [31] develops a two-stage framework containing view synthesis and stereo matching. UniFuse [23] further improves [53] with better accuracy and fewer parameters. SliceNet [37] transforms the EPR data into slice-based representation, which can tackle the inherent distortion. Sun *et al.* [46,45] focus on horizontal and vertical contents of a scene for 3D reconstruction. 360MonoDepth [39] projects the high-resolution spherical image into tangent image for efficient training. In 2022, SegFuse [16] utilizes geometric and temporal consistency to constraint depth recovery. GLPanoDepth [3] employs vision transformer and CNNs to encode cubemap and spherical images respectively, obtaining global-to-local representation of panoramas. ACDNet [67] designs adaptively combined dilated convolution to extend receptive field in the EPR and achieves state-of-the-art performances.

Unsupervised category: In 2019, Nikolaos *et al.* [68] explore spherical view synthesis for monocular 360° depth estimation in a self-supervised manner. In 2021, OlaNet [26] adopts the distortion-aware view synthesis, atrous spatial pyramid pooling, and L1-norm regularized smooth term to effectively and robustly deal with self-supervised panoramic depth estimation. Zhou *et al.* [65] combine supervised and unsupervised learning methods to facilitate network training. In 2022, Yun *et al.* [63] propose a self-supervised method based on gravity-aligned videos. Similarly, they also utilize the complementarity of supervised and self-supervised learning to improve their models' robustness.

Different from them that only utilize 360° color image, our goal is to recover dense depth and 3D reconstruction from the aligned 360° color image and sparse depth, which could help improve the accuracy with large margins.

### 2.3 Masked Image Encoding for Vision Tasks

Recently, several Transformer [52] based approaches [7,11,4,58,19] have proved it effective to learn representations from masked images. Specifically, iGPT [7]

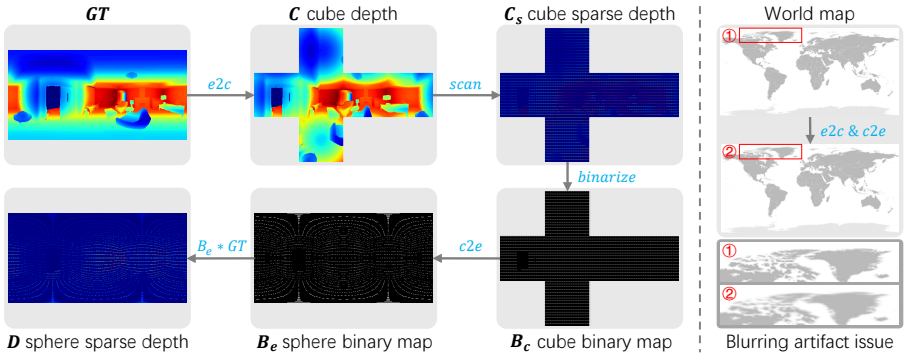


**Fig. 3.** A flowchart of data synthesis (left) and an example of the blurring artifact issue (right) that has negative effects on data processing. Best color of view.

trains a sequence Transformer to auto-regressively predict unknown pixels. ViT [11] conducts masked patch prediction to learn mean color. BEiT [4] presents to predict tokenization. SimMIM [58] and MAE [19] propose to recover raw pixels of randomly masked patches by a lightweight one-layer head and an asymmetric decoder, respectively. In contrast to them, our M$^3$PT is technically designed for multi-modal vision tasks instead of the single-modal image-based recognition.

## 3 Method

In this section, we first introduce how to synthesize sparse depth data and then elaborate on the multi-modal masked pre-training strategy.

### 3.1 Data Synthesis

All existing panoramic datasets do not provide sparse depth for 360° depth completion task. However, the sparse depth data can be possibly captured by some actual products such as Matterport Pro2 and FARO Focus 3D cameras. Limited by the lack of these hardware devices, in this paper, we imitate the principle of laser scanning to produce 360° sparse depth sampled from the dense ground-truth depth annotation, aiming at synthesizing the sparse depth data that matches the actual products as much as possible. The sampling principle is similar to that of KITTI benchmark [50] which provides depth with about 7% density captured by 64-line LiDAR scanning.

As illustrated in the left of Figure 3, the ground-truth depth $\boldsymbol{GT}$ is stored in spherical view by equirectangular projection, which brings inherent distortion. Hence, it's inaccurate to produce sparse depth directly based on $\boldsymbol{GT}$ in scanning mode. As an alternative, we first project the equirectangular $\boldsymbol{GT}$ into cubical map $\boldsymbol{C}$ by *e2c* function, ignoring the inherent distortion. Next, we generate cube sparse depth $\boldsymbol{C_s}$ via imitating the laser scanning, *e.g.*, taking one pixel for every

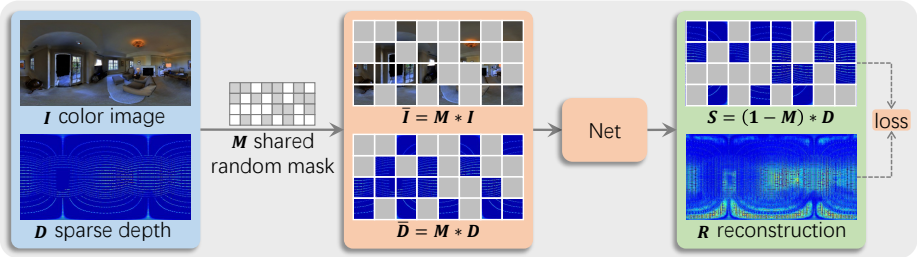


**Fig. 4. Our M$^3$PT pipeline**. During pre-training, the input color image $\boldsymbol{I}$ and sparse depth $\boldsymbol{D}$ are masked out by the shared random mask $\boldsymbol{M}$, obtaining $\bar{\boldsymbol{I}}$ and $\bar{\boldsymbol{D}}$ respectively. Then $\bar{\boldsymbol{I}}$ and $\bar{\boldsymbol{D}}$ are fed into a network to predict the depth reconstruction $\boldsymbol{R}$, supervised by the signal $\boldsymbol{S}$ which is the complementary set of $\bar{\boldsymbol{D}}$ in $\boldsymbol{D}$. After pre-training, with $\boldsymbol{I}$ and $\boldsymbol{D}$ as input, the network with learned initial weights is applied to recover target depths supervised by dense ground-truth depth annotations.

eight pixels horizontally and one pixel for every two pixels vertically. Then $\boldsymbol{C_s}$ is binarized to obtain cube binary map $\boldsymbol{B_c}$. $\boldsymbol{B_c}$ is thus converted into $\boldsymbol{B_e}$ by *c2e* function. Finally, we acquire the desired sparse depth $\boldsymbol{D}$ multiply $\boldsymbol{B_e}$ by $\boldsymbol{GT}$. The process can be simply defined as:

$$\boldsymbol{D} = \boldsymbol{B_e} * \boldsymbol{GT}, \tag{1}$$

where $\boldsymbol{B_e} = f(\boldsymbol{B_c}|\boldsymbol{C_s}, \boldsymbol{C}, \boldsymbol{GT})$, $f(\cdot)$ refers to the combination of *e2c*, *scan*, *binarize*, and *c2e*. The details of *e2c* and *c2e* functions refer to this project[3].

Note that, it is theoretically possible to use *c2e* to directly project the cubical $\boldsymbol{C_s}$ into the equirectangular $\boldsymbol{D}$. However, this would lead to blurring artifacts in the polar region, as evidenced in the right part of Figure 3. Instead, we choose to project a binary map and then use it to accurately sample valid points from $\boldsymbol{GT}$. In this way, our method can reduce error pixels as much as possible.

### 3.2 Multi-Modal Masked Pre-Training

As shown in Figure 4, our multi-modal masked pre-training (M$^3$PT) for 360° depth completion is a simple strategy that reconstructs the sparse depth signal given partial observations of the RGB-D pair under shared random mask. Here we introduce the key components of M$^3$PT and explicitly analyze the differences between recent visual masked pre-training approaches (*e.g.*, MAE [19], SimMIM [58]) and M$^3$PT.

**Shared random mask.** Different from MAE where masking is performed on single RGB data, we propose to mask both RGB image and sparse depth with the *shared random mask* to produce incomplete RGB-D pair as input for pre-training. In fact, there are other options of masking strategies for PDC task, including (i) only mask the RGB image, (ii) only mask the sparse depth, and

[3] https://github.com/sunset1995/py360convert

(iii) mask both RGB image and sparse depth but with different random masks. Unfortunately, all of these strategies have a risk of leaking information from another modality, preventing the pre-training task from learning robust semantics based on the multi-modal context. We will show the comparisons between different masking strategies later in the experimental part.

**Backbone.** Our method is flexible and can be applied to any existing approach which receives the RGB-D pair as input. In this paper, we mainly choose GuideNet [47] as backbone for the majority of our experiments. In addition, we also test the effectiveness of M$^3$PT using UniFuse [23] and HoHoNet [46]. Note that there is no need to design extra modules (e.g., a decoder) for the architectures of these existing approaches, even when they have an additional pre-training stage in M$^3$PT. It is because that the regression targets are physically similar between pre-training and fine-tuning stages. See more details in the 'Reconstruction target' part as follows.

**Reconstruction target.** The reconstruction target of M$^3$PT is the sparse depth on the masked regions. It is quite different from the popular masked pre-training methods [19,58] in vision where the missing image pixels are predicted. Compared to the vision pre-training counterparts [19,58], this design has two obvious advantages: (i) it closes the gap between pre-training and fine-tuning tasks, as they differ only in the prediction density; (ii) it leads to *no architectural modification* between pre-training and fine-tuning stages, which can potentially make the transfer learning more smooth and effective.

## 4 Experiments

Here, we first report datasets and metrics. Next, extensive ablation studies are conducted to verify the effectiveness of the proposed M$^3$PT. Then, we compare against other state-of-the-art (SoTA) works on three datasets. At last, we validate the generalization capability of M$^3$PT on KITTI benchmark [50].

### 4.1 Datasets

We conduct our experiments on three commonly used benchmark datasets of real world, *i.e.*, Matterport3D[4] [1], Stanford2D3D[5] [2], and 3D60[6] [68]. Matterport3D is a scanned dataset collected by Matterport's Pro 3D camera. The latest Matterport3D ($512 \times 256$) consists of 7,907 panoramic RGB-D pairs, of which 5636 for training, 744 for validating, and 1527 for testing. Stanford2D3D is composed of 1,413 panoramic color images and corresponding depth maps, whose training and testing splits contain 1,040 and 373 RGB-D pairs, respectively. We resize them to $512 \times 256$. 3D60 is initially made up of Matterport3D, Stanford2D3D, and SunCG [44]. But now it skips the entire SunCG dataset considering legal matters. As a result, the latest 3D60 ($512 \times 256$) consists of 6,669 RGB-D pairs for training, 906 for validating, and 1831 for testing, 9,406 in total.

[4] https://vcl3d.github.io/Pano3D/download/
[5] http://buildingparser.stanford.edu/dataset.html
[6] https://vcl3d.github.io/3D60/

| Masked Data | Shared Mask | Mask Ratio | Mask Size | | | |
|---|---|---|---|---|---|---|
| | | | 4 | 8 | 16 | 32 |
| RGB | - | 0.75 | 195.7 | 198.1 | 194.0 | 196.2 |
| D | - | | 190.7 | 177.6 | 186.1 | 203.3 |
| RGB-D | No | | 183.4 | 178.5 | 182.2 | 193.0 |
| RGB-D | Yes | | **166.8** | **168.9** | **167.6** | **169.9** |

**Table 2.** Ablation on different masked input data on Stanford2D3D dataset, where the metric is RMSE (mm). The error of baseline without pre-training is **196.7**.

| Mask Size | 4 | | | 8 | | | 16 | | | | | | 32 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mask Ratio | 0.45 | 0.6 | 0.75 | 0.45 | 0.6 | 0.75 | 0.15 | 0.3 | 0.45 | 0.6 | 0.75 | 0.9 | 0.45 | 0.6 | 0.75 |
| RMSE | 172.4 | 170.3 | 168.8 | 171.2 | 169.3 | 168.9 | 173.5 | 169.4 | 168.9 | 169.7 | **167.6** | 169.3 | 173.7 | 172.4 | 169.9 |

**Table 3.** Ablation on different mask sizes and mask ratios on Stanford2D3D dataset.

### 4.2 Metrics

Following previous works [46,37,63,67], we use five common and standard metrics to evaluate our methods, including MRE, MAE, RMSE, RMSElog, and $\delta_i$ ($i = 1.25, 1.25^2, 1.25^3$). Please refer to our appendix for more details.

### 4.3 Ablation Studies

**Settings**: We employ GuideNet [47] as the default backbone. The model is pre-trained for 300 epochs and fine-tuned for 100 epochs on every dataset. The mask is randomly generated following [58,19] with different sizes and ratios.

**(1) Masking strategy.**

**(i)** We explore how to corrupt RGB-D data during pre-training in Table 2. We can find that shielding only RGB, only Depth, or RGB-D without shared mask, all of which lead to worse performances because these operations destroy the model's learning of unknown areas. In contrast, the model achieves the best results when employing the proposed shared random mask, indicating that corrupting the same areas can contribute to improvement for multi-modal vision tasks. The following experiments are based on the random shared mask.

**(ii)** We study the effect of different mask sizes and ratios on the model's representation learning in Table 3. First, the model performs better when the mask size is changed from 4 to 16. We hold the opinion that the larger mask urges the model to learn long-range dependency between invisible and visible pixels. However, when setting the mask size to 32, the model has a degraded performance as it is too large to establish remote dependency. Second, when the mask size is set to 16, the model tends to perform better from 15% to 90% ratios, which could enforce the model to predict more unseen areas and acquires representation that is closer to the real domain.

**(2) Number of pre-training epochs and data amounts.**

**(i)** The left of Figure 5 demonstrates the influence of different pre-training epochs on fine-tuning, and the right shows the loss of pre-training. We can find

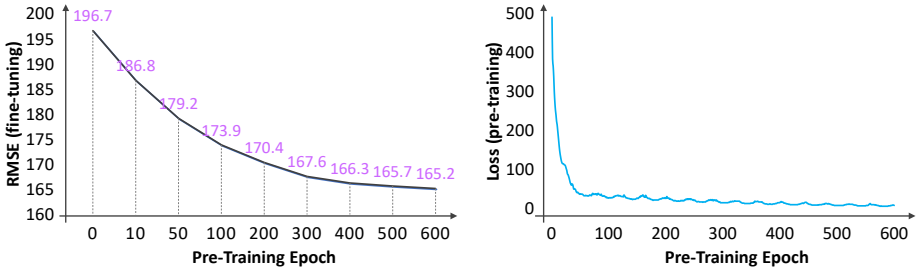


**Fig. 5.** Ablation on different pre-training epochs with mask size 16 and mask ratio 0.75 on Stanford2D3D dataset. The loss value is magnified by $10^4$ for clear visualization.

| Dataset | Data amount | w/o Pret. (100 epochs) | w/o Pret. (400 epochs) | Pret. on Self Data (300+100 epochs) | Pret. on All Data (300+100 epochs) |
|---|---|---|---|---|---|
| Matterport3D | 5.6k | 168.1 | 168.5 | 146.2 | **138.9** |
| Stanford2D3D | 1k | 196.7 | 196.3 | 167.6 | **149.0** |
| 3D60 | 6.7k | 159.9 | 160.2 | 142.3 | **127.2** |

**Table 4.** Ablation on different data amounts used during pre-training (Pret.). The mask size and ratio are 16 and 0.75, respectively. "300+100 epochs" denotes 300 pre-training epochs and 100 fine-tuning epochs.

that the model's error gradually decreases with the increase of pre-training epochs. This is because the model can learn better representation with more epochs, which is also reflected in the lower loss in the right of Figure 5. For the trade-off between speed and accuracy, unless otherwise stated, we pre-train the model for 300 epochs by default.

**(ii)** Table 4 explores the effect of different data amounts for pre-training on fine-tuning. For a fair comparison, we report the results of the 400th epoch without pre-training, whose cost roughly aligns with the setting of pre-training for 300 epochs and fine-tuning for 100 epochs. It can be found that without pre-training, the performance of 400 epochs has no improvements over 100 epochs. However, when conducting M$^3$PT just on single dataset, it leads to 12.9% improvements averagely on three datasets, demonstrating the significant effectiveness of M$^3$PT. Further, when pre-training on all the data of these three datasets, the performances are always superior to that only using a single dataset. Therefore, it is concluded that more data involved in M$^3$PT can consistently prevents the overfitting risks during fine-tuning.

### 4.4 Comparisons with SoTA Methods

In this subsection, we compare with recent SoTA works, including UniFuse [23], HoHoNet [46], PENet [21], and GuideNet [47]. HoHo-R and HoHo-H severally refer to using ResNet [20] and HardNet [6] as its backbone. Table 5 and Figure 6 demonstrate the quantitative and qualitative results, respectively. Based on

| Dataset | Method | Error Metric ↓ | | | | Accuracy Metric ↑ | | |
|---|---|---|---|---|---|---|---|---|
| | | MRE | MAE | RMSE | RMSElog | $\boldsymbol{\delta_{1.25}}$ | $\boldsymbol{\delta_{1.25^2}}$ | $\boldsymbol{\delta_{1.25^3}}$ |
| Matterport3D | UniFuse [23] | 0.0475 | 95.2 | 229.1 | 0.0381 | 0.9710 | 0.9924 | 0.9970 |
| | HoHo-R [46] | 0.0355 | 75.0 | 199.2 | 0.0311 | 0.9806 | 0.9945 | 0.9977 |
| | HoHo-H [46] | 0.0406 | 85.7 | 215.5 | 0.0337 | 0.9772 | 0.9938 | 0.9975 |
| | PENet [21] | 0.0493 | 91.5 | 248.0 | 0.0350 | 0.9728 | 0.9935 | 0.9970 |
| | GuideNet [47] | 0.0438 | 87.2 | 192.9 | 0.0327 | 0.9806 | 0.9948 | 0.9981 |
| | **M$^3$PT** | **0.0164** | **36.2** | **138.9** | **0.0193** | **0.9927** | **0.9976** | **0.9990** |
| Stanford2D3D | UniFuse [23] | 0.0489 | 93.4 | 216.2 | 0.0392 | 0.9661 | 0.9919 | 0.9973 |
| | HoHo-R [46] | 0.0677 | 123.9 | 242.5 | 0.0478 | 0.9463 | 0.9862 | 0.9959 |
| | HoHo-H [46] | 0.0695 | 127.9 | 254.8 | 0.0497 | 0.9434 | 0.9852 | 0.9957 |
| | PENet [21] | 0.0530 | 95.9 | 200.6 | 0.0404 | 0.9694 | 0.9934 | 0.9981 |
| | GuideNet [21] | 0.0506 | 92.1 | 196.7 | 0.0380 | 0.9689 | 0.9926 | 0.9978 |
| | **M$^3$PT** | **0.0274** | **52.9** | **149.0** | **0.0263** | **0.9859** | **0.9963** | **0.9988** |
| 3D60 | UniFuse [23] | 0.0446 | 94.1 | 215.6 | 0.0342 | 0.9749 | 0.9947 | 0.9984 |
| | HoHo-R [46] | 0.0338 | 75.6 | 196.9 | 0.0294 | 0.9818 | 0.9954 | 0.9983 |
| | HoHo-H [46] | 0.0376 | 81.9 | 205.8 | 0.0317 | 0.9788 | 0.9947 | 0.9981 |
| | PENet [21] | 0.0680 | 120.3 | 233.9 | 0.0321 | 0.9743 | 0.9926 | 0.9980 |
| | GuideNet [21] | 0.0689 | 144.2 | 239.3 | 0.0418 | 0.9711 | 0.9954 | 0.9987 |
| | **M$^3$PT** | **0.0144** | **34.1** | **127.2** | **0.0165** | **0.9944** | **0.9985** | **0.9995** |

**Table 5.** Quantitative comparisons of panoramic depth completion on three datasets. The best and the second best results are highlighted in bold and underline, respectively.

different baselines, Figure 7 further shows the influence of additional sparse depth information and proposed M$^3$PT on the recovery of panoramic depth.

**(1) Quantitative results.**

Overall, as illustrated in Table 5, the proposed M$^3$PT is consistently superior to other methods in all metrics on three datasets.

**(i)** On Matterport3D dataset, M$^3$PT greatly exceeds the second-best HoHo-R by 53.8%, 51.7%, and 37.9% in MRE, MAE, and RMSElog, severally. Compared with the suboptimal GuideNet in RMSE, the error is reduced from 192.9mm to 138.9mm, improving the performance nearly by 28.0%. Besides, M$^3$PT achieves the highest accuracies in $\delta_i$ with different thresholds, outperforming the second-best method by 1.21, 0.29, and 0.09 percent point in $\delta_1$, $\delta_2$, and $\delta_3$, respectively.

**(ii)** On Stanford2D3D dataset, M$^3$PT is superior to the suboptimal UniFuse with 44.0% improvement in MRE. Also, the MAE, RMSE, RMSElog is severally reduced by 42.6%, 24.3%, and 30.8% when comparing M$^3$PT with the second-best GuideNet. In addition, the accuracy metric verifies the effectiveness of M$^3$PT again, which plays a prominent role in all approaches.

**(iii)** On 3D60 dataset, M$^3$PT surpasses the second-best HoHo-R with large margins, improving it by 57.4% in MRE, 54.9% in MAE, 35.4% in RMSE, and 43.9% in RMSElog, severally. Furthermore, M$^3$PT is more accurate than other

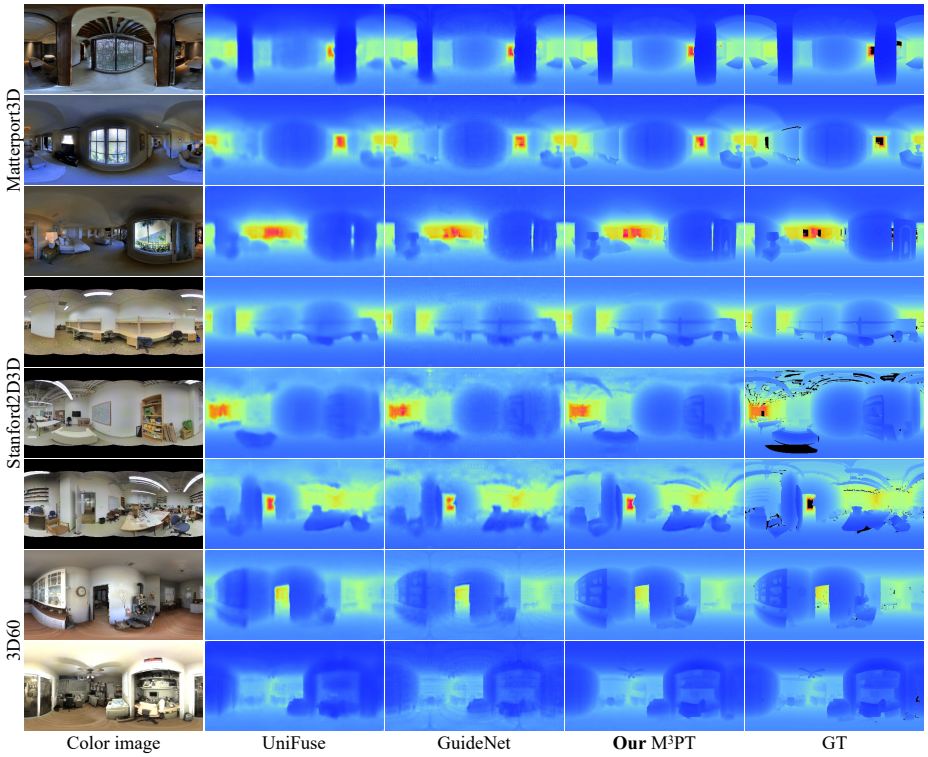


**Fig. 6.** Qualitative comparison of different methods, including UniFuse [23], GuideNet [47], and our M$^3$PT. More visualizations can be found in our supplementary material.

approaches and prevail over the suboptimal methods with 1.26, 0.31, and 0.08 percent point in $\delta_1$, $\delta_2$, and $\delta_3$, respectively.

**(iv)** Last but not least, apart from GuideNet that has been reported in Table 5, we further employ UniFuse, HoHo-R, and HoHo-H as baselines to see the influence of the additional sparse depth data and proposed mask strategy on the recovery of panoramic depth. As shown in Figure 7, gray bar: only using RGB, orange bar: only using sparse depth, light orange bar: using both RGB and its sparse depth, and blue bar: using RGB and sparse depth with the proposed mask strategy M$^3$PT. We can find that the error of only using sparse depth is much lower than that of only using RGB. Also, adding sparse depth data can benefit models with very large margins. Specifically, comparing light orange bar with gray bar, the errors of UniFuse, HoHo-R, and HoHo-H are severally reduced by 55.8%, 41.5%, and 50.5% on average on three datasets. What's more, adopting M$^3$PT contributes to their significant improvements of 27.0%, 25.4%, and 26.3% on average compared with the orange bars on three datasets. These facts indicate sparse depth information has great reference value for depth recovery, and also prove that the panoramic depth completion is a potentially valuable task.

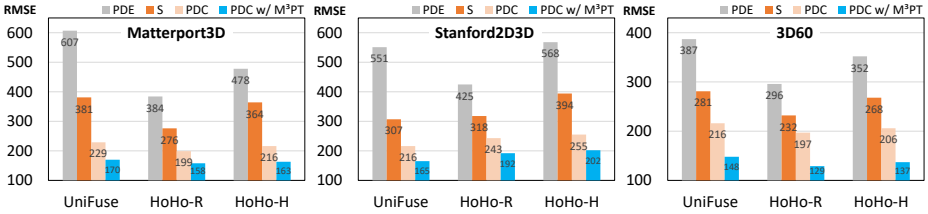


**Fig. 7.** Comparisons of different baselines with different-modal input data and M$^3$PT. PDE: panoramic depth estimation only from color images. PDC: panoramic depth completion from not only color images but the corresponding sparse depths.

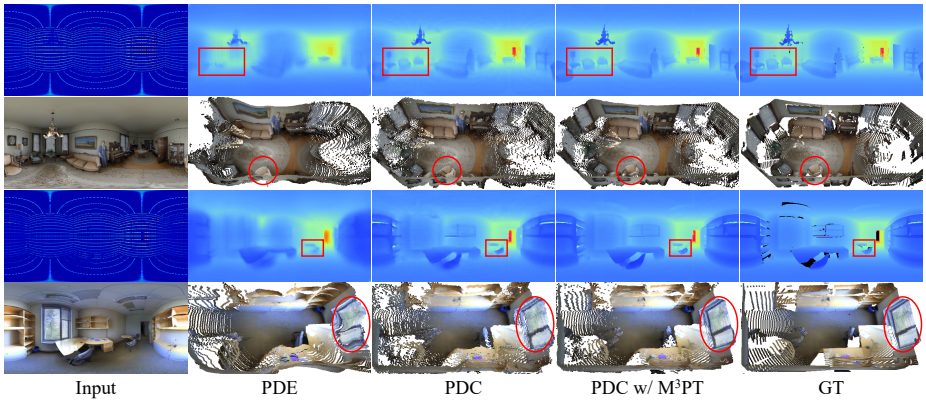


**Fig. 8.** Visual results of UniFuse [23] with different-modal input data and M$^3$PT.

**(2) Qualitative results.**

**(i)** As shown in Figure 6, our M$^3$PT can recover more detailed objects and precise depth with reasonable visual effect. For example, for one thing, as illustrated in the first, fourth, and seventh rows, M$^3$PT succeeds in predicting clearer edges of doors, tables, chairs, windows, *etc.* For another thing, as shown in the fifth and sixth rows, although the color images of Stanford2D3D do not have pixels at both the top and bottom, M$^3$PT can still predict more accurate depth values in the invisible areas. It strongly demonstrates the effectiveness and generalization of the proposed masked pre-training strategy via corrupting RGB-D data. In addition, M$^3$PT is also good at distinguishing from background and foreground, *e.g.*, the furniture can be clearly discriminated from the wall.

**(ii)** As illustrated in Figure 8, based on only color images (PDE), the depth predicted by UniFuse is extremely blurry that the corresponding 3D reconstruction introduces plenty of wrong location information, which causes negative deformation, especially nearby walls. By contrast, adding sparse depth data (PDC) vastly improves the visual effect of both depth recovery and 3D reconstruction. Furthermore, when deploying the proposed M$^3$PT with RGB-D data as input, both objects' structures and details tend to be more clear and abundant.

| Method | RMSE | MAE | iRMSE | iMAE |
|---|---|---|---|---|
| S2D [35] | 858.02 | 311.47 | 3.07 | 1.67 |
| +M$^3$PT | **844.16** | **267.64** | **3.01** | **1.51** |
| GuideNet [47] | 777.78 | 221.59 | 2.39 | **1.00** |
| +M$^3$PT | **761.57** | **217.68** | **2.26** | **1.00** |
| ACMNet [64] | 789.72 | 216.65 | 2.32 | 0.96 |
| +M$^3$PT | **774.63** | **209.31** | **2.25** | **0.93** |

**Table 6.** Performances of M$^3$PT with different baselines on KITTI validation split.

### 4.5 Generalization Capability

In this subsection, we further verify the generalization capability of M$^3$PT on KITTI depth completion benchmark, whose sparse depth data is obtained by a 64-line LiDAR, and the RGB-D pairs have limited field of vision. As reported in Table 6, M$^3$PT consistently improves the performances of S2D, GuideNet, and ACMNet. For example, M$^3$PT reduces RMSE/MAE by 15.05mm/18.36mm averagely, indicating that our M$^3$PT possesses robust generalization capability.

## 5 Conclusion

In this paper, we introduced a potentially valuable task, *i.e.*, panoramic depth completion, to help with dense panoramic depth recovery and 3D reconstruction from monocular 360° RGB-D data. Furthermore, we proposed the multi-modal masked pre-training (M$^3$PT) framework to handle this task. It was the first time we showed that the masked pre-training could be very effective in modeling multi-modal tasks for vision, instead of the single-modal image recognition which was popularized by the masked autoencoders (MAE). As a result, comprehensive evaluations demonstrated the superiority of M$^3$PT on three benchmark datasets. At last, we hope our exploration in this paper can facilitate future studies concerned with multi-modal vision tasks. In the future, we are going to extend M$^3$PT to related topics such as depth denoising and super-resolution.

## 6 Acknowledgement

The authors would like to thank reviewers for their detailed comments and instructive suggestions. This work was supported by the National Science Fund of China under Grant Nos. U1713208, 62072242 and Postdoctoral Innovative Talent Support Program of China under Grant BX20200168, 2020M681608. Note that the PCA Lab is associated with, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, Nanjing University of Science and Technology.

## References


1. Albanis, G., Zioulis, N., Drakoulis, P., Gkitsas, V., Sterzentsenko, V., Alvarez, F., Zarpalas, D., Daras, P.: Pano3d: A holistic benchmark and a solid baseline for 360° depth estimation. In: CVPRW. pp. 3722–3732. IEEE (2021) 2, 3, 8
2. Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2d-3d-semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105 (2017) 3, 5, 8
3. Bai, J., Lai, S., Qin, H., Guo, J., Guo, Y.: Glpanodepth: Global-to-local panoramic depth estimation. arXiv preprint arXiv:2202.02796 (2022) 2, 5
4. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021) 5, 6
5. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. In: 3DV (2017) 5
6. Chao, P., Kao, C.Y., Ruan, Y.S., Huang, C.H., Lin, Y.L.: Hardnet: A low memory traffic network. In: ICCV. pp. 3552–3561 (2019) 10
7. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: ICML. pp. 1691–1703. PMLR (2020) 5
8. Cheng, X., Wang, P., Guan, C., Yang, R.: Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In: AAAI. pp. 10615–10622 (2020) 4
9. Cheng, X., Wang, P., Yang, R.: Learning depth with convolutional spatial propagation network. In: ECCV. pp. 103–119 (2018) 4
10. Chodosh, N., Wang, C., Lucey, S.: Deep convolutional compressed sensing for lidar depth completion. In: ACCV. pp. 499–513. Springer (2018) 4
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 5, 6
12. Eder, M., Moulon, P., Guan, L.: Pano popups: Indoor 3d reconstruction with a plane-aware network. In: 3DV. pp. 76–84. IEEE (2019) 5
13. Eldesokey, A., Felsberg, M., Khan, F.S.: Confidence propagation through cnns for guided sparse depth regression. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2423–2436 (2019) 4
14. Erhan, D., Bengio, Y., Courville, A., Manzagol, P.A., Vincent, P., Bengio, S.: Why does unsupervised pre-training help deep learning? Journal of Machine Learning Research **11**, 625–660 (2010) 2, 3
15. Feng, B.Y., Yao, W., Liu, Z., Varshney, A.: Deep depth estimation on 360 images with a double quaternion loss. In: 3DV. pp. 524–533. IEEE (2020) 5
16. Feng, Q., Shum, H.P., Morishima, S.: 360 depth estimation in the wild–the depth360 dataset and the segfuse network. In: VR. IEEE (2022) 5
17. Gordon, A., Li, H., Jonschkowski, R., Angelova, A.: Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In: ICCV. pp. 8977–8986 (2019) 2
18. Gu, J., Xiang, Z., Ye, Y., Wang, L.: Denselidar: A real-time pseudo dense depth guided depth completion network. IEEE Robotics and Automation Letters **6**(2), 1808–1815 (2021) 4
19. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021) 3, 5, 6, 7, 8, 9
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 10

21. Hu, M., Wang, S., Li, B., Ning, S., Fan, L., Gong, X.: Penet: Towards precise and efficient image guided depth completion. In: ICRA (2021) 4, 10, 11
22. Jaritz, M., De Charette, R., Wirbel, E., Perrotton, X., Nashashibi, F.: Sparse and dense data with cnns: Depth completion and semantic segmentation. In: 3DV. pp. 52–60 (2018) 4
23. 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) 2, 5, 8, 10, 11, 12, 13
24. Jin, L., Xu, Y., Zheng, J., Zhang, J., Tang, R., Xu, S., Yu, J., Gao, S.: Geometric structure based and regularized depth estimation from 360 indoor imagery. In: CVPR. pp. 889–898 (2020) 5
25. Krauss, B., Schroeder, G., Gustke, M., Hussein, A.: Deterministic guided lidar depth map completion. arXiv preprint arXiv:2106.07256 (2021) 4
26. Lai, Z., Chen, D., Su, K.: Olanet: Self-supervised 360° depth estimation with effective distortion-aware view synthesis and l1 smooth regularization. In: ICME. pp. 1–6. IEEE (2021) 5
27. Lee, Y., Jeong, J., Yun, J., Cho, W., Yoon, K.J.: Spherephd: Applying cnns on a spherical polyhedron representation of 360deg images. In: CVPR. pp. 9181–9189 (2019) 5
28. Lee, Y., Jeong, J., Yun, J., Cho, W., Yoon, K.J.: Spherephd: Applying cnns on 360° images with non-euclidean spherical polyhedron representation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 5
29. Li, A., Yuan, Z., Ling, Y., Chi, W., Zhang, C., et al.: A multi-scale guided cascade hourglass network for depth completion. In: WACV. pp. 32–40 (2020) 4
30. Li, J., Zhang, T., Luo, W., Yang, J., Yuan, X.T., Zhang, J.: Sparseness analysis in the pretraining of deep neural networks. IEEE transactions on neural networks and learning systems **28**(6), 1425–1438 (2016) 2
31. Li, Y., Yan, Z., Duan, Y., Ren, L.: Panodepth: A two-stage approach for monocular omnidirectional depth estimation. In: 3DV. pp. 648–658. IEEE (2021) 2, 5
32. Lin, Y., Cheng, T., Zhong, Q., Zhou, W., Yang, H.: Dynamic spatial propagation network for depth completion. In: AAAI (2022) 4
33. Liu, L., Song, X., Lyu, X., Diao, J., Wang, M., Liu, Y., Zhang, L.: Fcfr-net: Feature fusion based coarse-to-fine residual learning for depth completion. In: AAAI. vol. 35, pp. 2136–2144 (2021) 4
34. Lu, K., Barnes, N., Anwar, S., Zheng, L.: From depth what can you see? depth completion via auxiliary image reconstruction. In: CVPR. pp. 11306–11315 (2020) 4
35. Ma, F., Cavalheiro, G.V., Karaman, S.: Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In: ICRA (2019) 4, 14
36. Park, J., Joo, K., Hu, Z., Liu, C.K., Kweon, I.S.: Non-local spatial propagation network for depth completion. In: ECCV (2020) 4
37. Pintore, G., Agus, M., Almansa, E., Schneider, J., Gobbetti, E.: Slicenet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In: CVPR. pp. 11536–11545 (2021) 2, 5, 9
38. Qiu, J., Cui, Z., Zhang, Y., Zhang, X., Liu, S., Zeng, B., Pollefeys, M.: Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In: CVPR. pp. 3313–3322 (2019) 4
39. Rey-Area, M., Yuan, M., Richardt, C.: 360monodepth: High-resolution 360° monocular depth estimation. arXiv e-prints pp. arXiv–2111 (2021) 5

40. Schuster, R., Wasenmuller, O., Unger, C., Stricker, D.: Ssgp: Sparse spatial guided propagation for robust and generic interpolation. In: WACV. pp. 197–206 (2021) 4
41. Shen, Z., Lin, C., Liao, K., Nie, L., Zheng, Z., Zhao, Y.: Panoformer: Panorama transformer for indoor 360 depth estimation. arXiv e-prints pp. arXiv–2203 (2022) 2
42. Shen, Z., Lin, C., Nie, L., Liao, K., Zhao, Y.: Distortion-tolerant monocular depth estimation on omnidirectional images using dual-cubemap. In: ICME. pp. 1–6. IEEE (2021) 2
43. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV. pp. 746–760. Springer (2012) 4
44. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: CVPR. pp. 1746–1754 (2017) 8
45. Sun, C., Hsiao, C.W., Wang, N.H., Sun, M., Chen, H.T.: Indoor panorama planar 3d reconstruction via divide and conquer. In: CVPR. pp. 11338–11347 (2021) 5
46. Sun, C., Sun, M., Chen, H.T.: Hohonet: 360 indoor holistic understanding with latent horizontal features. In: CVPR. pp. 2573–2582 (2021) 2, 5, 8, 9, 10, 11
47. 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) 4, 8, 9, 10, 11, 12, 14
48. Tateno, K., Navab, N., Tombari, F.: Distortion-aware convolutional filters for dense prediction in panoramic images. In: ECCV. pp. 707–722 (2018) 5
49. Teutscher, D., Mangat, P., Wasenmüller, O.: Pdc: Piecewise depth completion utilizing superpixels. In: ITSC. pp. 2752–2758. IEEE (2021) 4
50. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity invariant cnns. In: 3DV. pp. 11–20 (2017) 4, 6, 8
51. Van Gansbeke, W., Neven, D., De Brabandere, B., Van Gool, L.: Sparse and noisy lidar completion with rgb guidance and uncertainty. In: MVA. pp. 1–6 (2019) 4
52. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS. vol. 30 (2017) 5
53. Wang, F.E., Yeh, Y.H., Sun, M., Chiu, W.C., Tsai, Y.H.: Bifuse: Monocular 360 depth estimation via bi-projection fusion. In: CVPR. pp. 462–471 (2020) 5
54. Wong, A., Cicek, S., Soatto, S.: Learning topology from synthetic data for unsupervised depth completion. IEEE Robotics and Automation Letters **6**(2), 1495–1502 (2021) 4
55. Wong, A., Fei, X., Hong, B.W., Soatto, S.: An adaptive framework for learning unsupervised depth completion. IEEE Robotics and Automation Letters **6**(2), 3120–3127 (2021) 4
56. Wong, A., Fei, X., Tsuei, S., Soatto, S.: Unsupervised depth completion from visual inertial odometry. IEEE Robotics and Automation Letters **5**(2), 1899–1906 (2020) 4
57. Wong, A., Soatto, S.: Unsupervised depth completion with calibrated backprojection layers. In: ICCV (2021) 4
58. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. arXiv preprint arXiv:2111.09886 (2021) 3, 5, 6, 7, 8, 9
59. Xu, Y., Zhu, X., Shi, J., Zhang, G., Bao, H., Li, H.: Depth completion from sparse lidar data with depth-normal constraints. In: ICCV. pp. 2811–2820 (2019) 4
60. Xu, Z., Yin, H., Yao, J.: Deformable spatial propagation networks for depth completion. In: ICIP. pp. 913–917. IEEE (2020) 4

61. Yan, L., Liu, K., Gao, L.: Dan-conv: Depth aware non-local convolution for lidar depth completion. Electronics Letters **57**(20), 754–757 (2021) 4
62. Yan, Z., Wang, K., Li, X., Zhang, Z., Xu, B., Li, J., Yang, J.: Rignet: Repetitive image guided network for depth completion. arXiv preprint arXiv:2107.13802 (2021) 4
63. Yun, I., Lee, H.J., Rhee, C.E.: Improving 360 monocular depth estimation via non-local dense prediction transformer and joint supervised and self-supervised learning. In: arXiv preprint arXiv:2109.10563 (2021) 2, 5, 9
64. Zhao, S., Gong, M., Fu, H., Tao, D.: Adaptive context-aware multi-modal network for depth completion. IEEE Transactions on Image Processing (2021) 4, 14
65. Zhou, K., Yang, K., Wang, K.: Panoramic depth estimation via supervised and unsupervised learning in indoor scenes. Applied Optics **60**(26), 8188–8197 (2021) 5
66. Zhu, Y., Dong, W., Li, L., Wu, J., Li, X., Shi, G.: Robust depth completion with uncertainty-driven loss functions. arXiv preprint arXiv:2112.07895 (2021) 4
67. Zhuang, C., Lu, Z., Wang, Y., Xiao, J., Wang, Y.: Acdnet: Adaptively combined dilated convolution for monocular panorama depth estimation. In: AAAI (2022) 2, 5, 9
68. Zioulis, N., Karakottas, A., Zarpalas, D., Alvarez, F., Daras, P.: Spherical view synthesis for self-supervised 360 depth estimation. In: 3DV. pp. 690–699. IEEE (2019) 3, 5, 8
69. Zioulis, N., Karakottas, A., Zarpalas, D., Daras, P.: Omnidepth: Dense depth estimation for indoors spherical panoramas. In: ECCV. pp. 448–465 (2018) 5