This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Distilling Monocular Foundation Model for Fine-grained Depth Completion

Yingping Liang[1] Yutao Hu[2] Wenqi Shao[3] Ying Fu[1†]
[1]Beijing Institute of Technology [2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications, Southeast University [3]Shanghai AI Laboratory
{liangyingping,fuying}@bit.edu.cn huyutao@seu.edu.cn weqish@link.cuhk.edu.hk

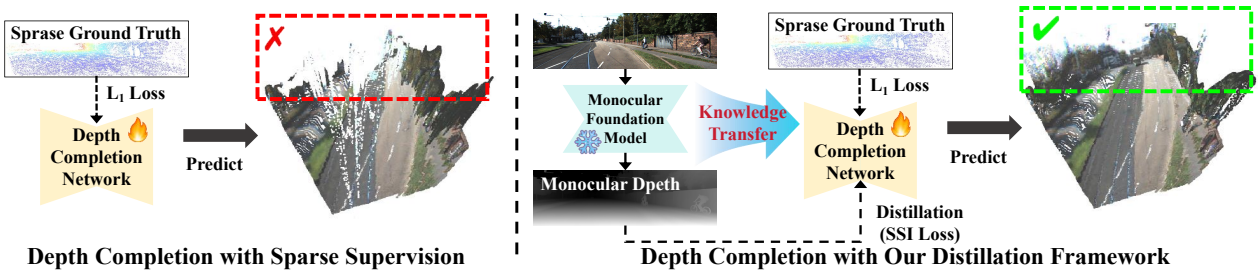


Figure 1. Depth completion models trained solely with L_1 loss and sparse ground truth produce incomplete and fragmented depth predictions. Our framework, however, demonstrates significant improvements by distilling knowledge from monocular foundation models and incorporating a scale- and shift-invariant loss (SSI Loss), resulting in more complete and accurate dense depth completion.

Abstract

Depth completion involves predicting dense depth maps from sparse LiDAR inputs. However, sparse depth annotations from sensors limit the availability of dense supervision, which is necessary for learning detailed geometric features. In this paper, we propose a two-stage knowledge distillation framework that leverages powerful monocular foundation models to provide dense supervision for depth completion. In the first stage, we introduce a pre-training strategy that generates diverse training data from natural images, which distills geometric knowledge to depth completion. Specifically, we simulate LiDAR scans by utilizing monocular depth and mesh reconstruction, thereby creating training data without requiring ground-truth depth. Besides, monocular depth estimation suffers from inherent scale ambiguity in real-world settings. To address this, in the second stage, we employ a scale- and shift-invariant loss (SSI Loss) to learn real-world scales when fine-tuning on real-world datasets. Our two-stage distillation framework enables depth completion models to harness the strengths of monocular foundation models. Experimental results demonstrate that models trained with our two-stage distillation framework achieve state-of-the-art performance, ranking ***first place*** *on the KITTI benchmark. Code is available at https://github.com/Sharpiless/DMD3C*

1. Introduction

Depth completion is a fundamental task in computer vision, where the goal is to generate dense depth maps from sparse depth measurements. This task is particularly important in applications such as autonomous driving [23, 47, 61], robotics [26, 39, 63], and augmented reality [21, 30].

Recent approaches [7, 27, 28, 65, 67, 78] leverage deep neural networks to learn from depth data. Some methods [56, 79] also incorporate RGB images to guide the depth completion process. Despite these advancements, state-of-the-art models [50, 57, 70] still face difficulties in capturing fine-grained geometric details, particularly in complex outdoor scenes where depth annotations are sparse [53, 58, 75].

The challenge arises from the reliance on sparse ground truth for training. The lack of dense ground truth makes it difficult for models to accurately learn depth completion across an entire scene. Monocular depth estimation [25, 46, 71, 72], on the other hand, has the ability to provide dense depth predictions from a single image. Specifically, state-of-the-art monocular foundation models could generate dense depth maps, containing fine-grained details and relative depth relationships, offering valuable guidance for training depth completion networks.

To make full use of the advantages of monocular foun-

† Corresponding author.

dation models, we propose a novel two-stage distillation framework to transfer geometric knowledge from monocular foundation models to depth completion networks. Specifically, in the **first distillation stage**, we generate training data through monocular foundation models and then distill knowledge via the proposed pre-training strategy. Specifically, diverse natural images are used to generate pseudo depth maps. Then, we utilize randomly sampled camera parameters to re-construct the scene with mesh and simulate LiDAR using ray simulation. The generated data trains the depth completion model to learn diverse geometric knowledge from monocular foundation models, enhancing its ability to generalize across different scenes.

However, monocular depth estimation suffers from inherent scale ambiguity [16, 81], resulting in depth predictions that vary greatly in scale. Thus, monocular depth alone cannot serve as a reliable basis for real-world depth. To solve this problem, in the **second distillation stage**, we introduce a scale- and shift-invariant loss (SSI Loss) [46]. Specifically, when fine-tuning on labeled datasets with sparse ground truth, SSI Loss ignores the scale and shift that causes the least loss from the depth prediction and monocular depth, to ensure consistent depth completions aligning monocular depth supervision across varying scales.

By integrating these two distillation stages into the training process, our method achieves state-of-the-art performance on the KITTI benchmark. The contributions of this work are summarized as follows:

- We propose a novel two-stage distillation framework to transfer knowledge from monocular foundation models to depth completion models, which enables the learning of fine-grained depth information from sparse ground truth by providing dense supervision.
- In the first stage, we propose a data generation strategy that uses monocular depth estimation and mesh reconstruction to simulate training data, enabling the model to learn geometric features from diverse natural images without the need for any LiDAR and ground truth.
- In the second stage, we propose a scale- and shift-invariant loss (SSI Loss) to address the scale ambiguity problem in monocular depth estimation, which ensures consistent depth completions across varying scales and focuses on learning real-world scale information.
- Our method achieves first place on the official KITTI depth completion benchmark, demonstrating the effectiveness of the proposed approach with significant gains in both quantitative metrics and qualitative visualizations.

2. Related Work

2.1. Depth Completion Datasets

Depth completion has been widely studied in both indoor and outdoor settings. Indoor datasets, such as NYU Depth V2 [48] and Matterport3D [4], are popular because dense depth measurements are easier to obtain in controlled environments. NYU Depth V2, for instance, provides dense depth maps from a Kinect sensor along with RGB images.

In outdoor environments, however, obtaining dense depth annotations is more challenging due to limitations in sensing technologies like LiDAR. The KITTI dataset [53] is a key benchmark for outdoor depth completion but suffers from sparse annotations, covering only about 5% of the image. To address this, training methods often employ complex post-processing and multi-frame fusion techniques to increase the annotation coverage to around 20% at most. Furthermore, ground truth values for long distances and dynamic objects typically cannot be included. The sparse ground truth raises challenges for training fine-grained depth completion models. Therefore, we employ monocular depth estimation techniques for dense distillation to preserve fine-grained details.

2.2. Depth Completion Methods

Depth completion methods have seen significant advancements [21, 26, 64]. Compared to predicting depth directly from a single RGB image [2, 22, 44, 73], depth completion fuses LiDAR information to obtain more accurate sparse depth cues. Recent methods [1, 6, 8, 9, 11, 14, 29, 34–38, 40–42, 45, 49, 50, 52, 55–57, 62, 66–70, 74, 76, 79, 80, 82, 84] rely on supervised learning, using pairs of sparse depth maps and corresponding dense ground truth. While effective with sufficient labeled data, these methods are limited in real-world applications, where obtaining dense annotations, especially in outdoor environments, is challenging [5, 15, 83]. To reduce the reliance on dense labels, [10, 77] use a teacher-student model, where a teacher network trained on labeled data provides guidance to a student network. However, pseudo-labels in complex or dynamic scenes can be noisy or inaccurate. Therefore, we turn to monocular depth estimation to provide fine-grained supervision for depth completion.

2.3. Monocular Depth Estimation

Monocular depth estimation [25, 46, 71, 72] has emerged as a powerful tool for generating dense depth maps from single RGB images, making it particularly valuable in scenarios where dense depth annotations are sparse or unavailable [18, 19, 32]. Notable models, such as MiDaS [46] and Depth-Anything [71, 72], have demonstrated impressive capabilities in predicting dense depth maps across a wide range of scenes. G2-MonoDepth [54] have proposed a general framework for monocular RGB-X Data with a scale invariant loss. Despite their versatility, monocular depth estimation methods face the challenge of scale ambiguity, which can limit the accuracy of depth completion, especially when integrated with sparse depth measurements

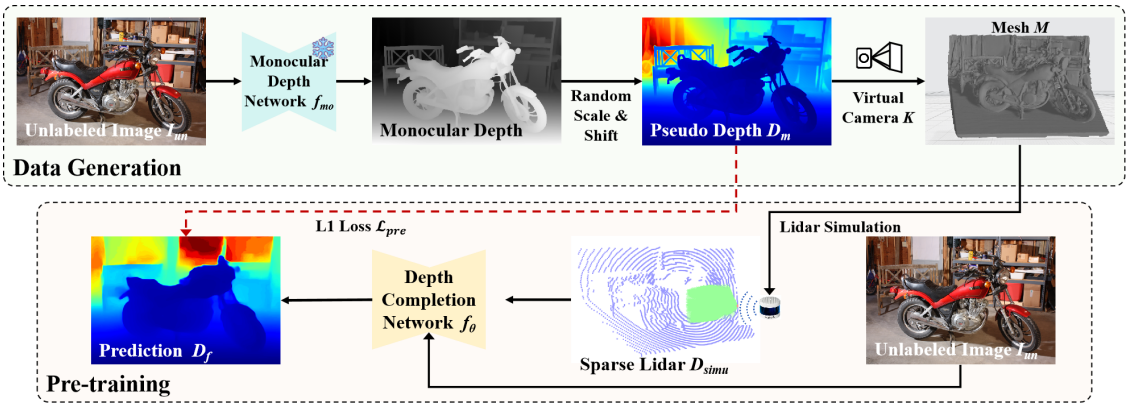


Figure 2. Illustration of our proposed first distillation stage with a data generation strategy to learn geometric features from monocular foundation models, which only requires unlabeled RGB images. We use the estimated monocular depth to re-construct the scene and then simulate the Lidar swap process to generate sparse points for training.

from LiDAR. Therefore, we introduce a scale- and shift-invariant loss inspired by [54], combined with a supervised loss to keep real-world scales when fine-tuning on real-world datasets.

3. Method

In this section, we first provide a brief overview of the motivation and formulation of our method. Then, we introduce our two-stage distillation framework. In the first stage, we propose a data generation strategy for pre-training that leverages large-scale, unlabeled images without ground-truth depth annotations, as illustrated in Figure 2. In the second stage, we detail the fine-tuning process on a labeled dataset with sparse ground truth, using monocular depth in combination with SSI Loss, as shown in Figure 3.

3.1. Motivation and Formulation

In the depth completion task, the input typically consists of an RGB image I and a sparse depth map D_s obtained from a LiDAR sensor. The objective is to produce a dense depth map D_f that provides a complete and detailed representation of the scene's depth. To achieve this, a depth completion model f_θ, parameterized by θ, takes both the RGB image and sparse depth map as inputs:

$$D_f = f_\theta(I, D_s). \tag{1}$$

To guide the model's learning, a sparse supervision loss is defined based on the available sparse depth annotations:

$$\mathcal{L}_{\text{sup}} = M \times |D_f - D_{\text{sparse}}|, \tag{2}$$

where M represents the valid mask for sparse depth ground truth D_{sparse}. This loss aligns the predicted dense depth map D_f with the sparse ground truth, ensuring that the output is consistent with the available depth data. However, this sparse supervision provides limited guidance, particularly in outdoor scenes [53], leading to challenges in achieving fine-grained and consistent depth predictions due to the inherent limitations of sparse supervision.

Foundation models for monocular depth estimation [71, 72] provide an alternative approach by generating fine-grained dense depth from single RGB images, which we use as supervision without relying solely on LiDAR. Therefore, in the first stage, we utilize the monocular foundation models to generate diverse and large-scale training data, to provide dense supervision for pre-training, allowing the model to learn geometric features. However, monocular methods suffer from scale ambiguities, which can lead to inconsistencies when combining dense monocular predictions with sparse depth. To address this, in the second stage, we introduce SSI Loss [46] combined with L1 loss when fine-tuning on labeled datasets with sparse ground truth to mitigate scale inconsistencies and learn real-world scale.

3.2. First Stage: Data Generation and Pre-training

To leverage diverse images without depth annotations, we introduce a pre-training strategy that utilizes synthesized depth data generated through monocular depth estimation on large-scale natural image datasets inspired by [31, 32]. As illustrated in Figure 2, this pre-training phase enables the model to learn robust geometric features from monocular foundation models across diverse scenes, preparing it for subsequent fine-tuning on real-world datasets.

Specifically, we first utilize a pre-trained monocular depth estimation model, such as Depth Anything V2 [72], to predict depth maps for such images. For each natural image I_{un}, the monocular depth estimation model predicts a

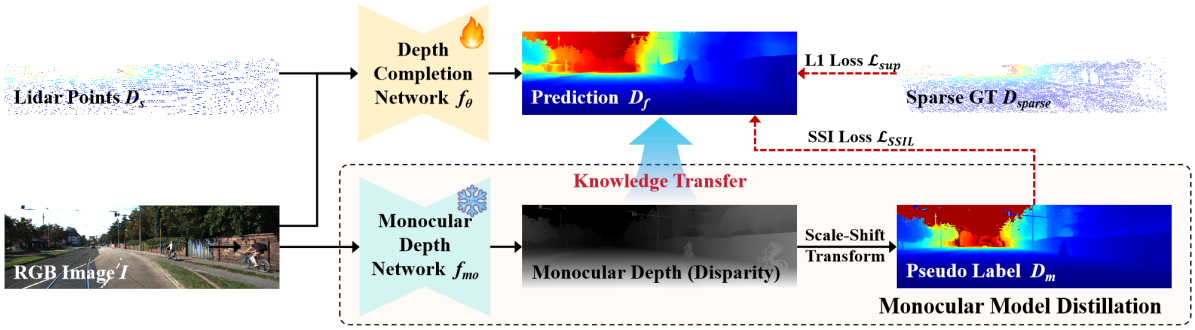


Figure 3. Illustration of our proposed second distillation stage utilizing foundation models for monocular depth estimation when fine-tuning on labeled datasets. Sparse ground truth provides real-world depth scale with L1 loss. Our method enhances this process by incorporating dense monocular depth for fine-grained supervision. However, monocular depth maps come with inherent scale and shift ambiguities. To address these challenges, we employ a Scale- and Shift-Invariant Loss (SSI Loss) that aligns the predictions with the dense monocular depth to match the real-world depth scale, ensuring more accurate depth completion.

corresponding dense depth map D_{syn}:

$$D_{\text{syn}} = f_{\text{mo}}(I_{\text{un}}), \tag{3}$$

where f_{mo} represents the monocular depth estimation model. These synthesized depth maps, though not accurate in metric scale, capture the relative depth relationships and structural details in the scene, providing valuable supervision signals during pre-training.

To simulate LiDAR scanning, we first sample a random camera intrinsic matrix K as:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{4}$$

where f_x and f_y represent the focal lengths along the x- and y-axes, and (c_x, c_y) denotes the center of the image.

For each pixel (u, v) in the depth map, the corresponding 3D coordinates (X, Y, Z) in the camera coordinate system are derived using the sampled camera intrinsic matrix K. This transformation generates a point cloud $P = \{(X, Y, Z)\}$ that represents the 3D spatial locations of each pixel. Leveraging the 3D point cloud P, we reconstruct a mesh M via surface reconstruction techniques, such as Poisson Surface Reconstruction [24], to create a continuous 3D surface model for simulation.

Next, we simulate a LiDAR sensor by generating a set of ray direction vectors $\mathbf{d}$. Each ray is cast from the origin along $\mathbf{d}_{i,j}$, intersecting the mesh M to determine the distance to the intersection point. This distance is recorded as the simulated LiDAR depth reading $D_{\text{simu}}(i, j)$. The resulting sparse depth map D_{simu} emulates LiDAR depth readings from the reconstructed 3D surface, to provide a synthetic depth image comparable to actual LiDAR data.

Then, the depth completion model f_θ takes the RGB image and the sparse depth from simulated scanning as inputs:

$$D_f = f_\theta(I_{\text{un}}, D_{\text{simu}}), \tag{5}$$

where D_{simu} is the sparse input from simulated LiDAR scanning. Afterwards, the depth completion model is pre-trained using L1 Loss:

$$\mathcal{L}_{\text{pre}} = |D_f - D_m|. \tag{6}$$

This pre-training process encourages the depth completion model f_θ to align its predictions with the synthesized depth maps, learning meaningful geometric features from the generated training data using monocular foundation models and diverse natural image data. Additionally, the model learns complex geometric structures and relative depth relationships across varying scenes, which are crucial for accurate depth completion in the fine-tuning phase.

3.3. Second Stage: Fine-tuning

The depth completion model pre-trained on the generated data learns powerful geometric features from the monocular foundation models. However, due to the inherent scale ambiguity of monocular depth, the model prediction is not consistent with the scale in the real world. To learn the real-world scale, we introduce a combined loss for fine-tuning on labeled datasets. To be specific, we first utilize the sparse ground truth D_s for supervised loss $\mathcal{L}_{\text{sup}}$ to focus on real-world depth, as in Equation 2. Then, the dense monocular depth D_m predicted by monocular depth models is also used to provide approximate depth values across the image, especially in regions not covered by the ground truth D_s.

However, the key challenge in using monocular depth estimation as supervision is the inherent scale ambiguity in depth predictions. To address this issue, we incorporate a Scale- and Shift-Invariant Loss (SSI Loss) into our distillation framework. The SSI Loss is designed to align the

Table 1. Unlabeled datasets used for pre-training data generation.

Dataset	Indoor	Outdoor	# Images
COCO [33]	✓	✓	118,287
Google Landmarks [59]		✓	117,576
Nuscenes [3]		✓	93,475
Cityscapes [12]		✓	19,998
DAVIS [43]	✓	✓	10,581

predicted depth map D_f with the dense monocular depth D_m. It remains invariant to differences in scale and shift between D_f and D_m. The SSI Loss is formulated as:

$$\mathcal{L}_{\text{SSIL}} = \min_{s,b} |D_f - (s \cdot D_m + b)|, \tag{7}$$

where s and b are the optimal scale and shift parameters that align the predicted depth D_f with the dense monocular depth D_m. The loss function seeks to find the best alignment between D_f and D_m, effectively normalizing any global differences in scale and offset. By minimizing this loss across all pixels in the image, the model is encouraged to produce depth maps that maintain consistency with the relative depth structure provided by the monocular depth estimates. In addition, we adapt a gradient matching term to preserve sharpness and align with with depth discontinuities. The gradient matching term is defined as:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N} \sum_{k=1}^{K} \left(|\nabla_x R^k| + |\nabla_y R^k| \right), \tag{8}$$

where $R = D_f - D_m$, and R^k denotes the difference of depth maps at scale k. We use $K = 4$ scale levels as in [46], halving the image resolution at each level. The final objective function combines the supervised loss $\mathcal{L}_{\text{sup}}$, the dense distillation loss $\mathcal{L}_{\text{SSIL}}$, and the regularization term $\mathcal{L}_{\text{reg}}$.

4. Experiments

In this section, we first introduce the details of our implementation, as well as the datasets and evaluation metrics for experiments. Then, detailed comparisons are conducted with the state-of-the-art methods. Finally, ablations and discussions are performed to confirm the effectiveness of our proposed main components. Additional analysis is provided in the supplementary materials, along with videos.

4.1. Experimental Pipeline

First stage involves utilizing our proposed data generation strategy to train the depth completion model from scratch, as described in Section 3.2. In this stage, we generate training data using the RGB images in Table 1 and train the depth completion model using SSI Loss, which does not require any ground truth depth. This stage allows the pre-trained model to focus on learning diverse geometric features without being constrained by labeled data, enhancing generalization across different scenes.

Second stage involves adapting the pre-trained model on labeled datasets, using supervised loss (L1 loss) with depth annotations and our proposed SSI Loss for knowledge distillation, as described in Section 3.3. Using L1 loss with sparse ground truth allows the depth completion model to adapt to the real-world depth scale under sparse supervision. Specifically, our proposed monocular model distillation uses SSI Loss to help the model maintain fine-grained detail with dense supervision by distilling monocular foundation models. Fine-tuning and validation are performed on the KITTI and NYU Depth V2 datasets, respectively.

4.2. Implementation Details

We follow the setup of recent work [50, 57, 79] and train our model on 4 NVIDIA RTX A100 GPUs. For the monocular foundation model used in distillation, we adopt Depth Anything V2 [72] due to its robust performance. For the depth completion network architecture, we use BP-Net [50] as our base model, which is primarily composed of ResNet [20] blocks. In the first stage, we utilize a mixed dataset collected from a wide range of resources, as shown in Table 1. We select five large-scale public datasets as our unlabeled sources for their diverse scenes, totaling approximately $360,000$ images with varying scenes and image scales. We adopt AdamW with a weight decay of 0.05 as the optimizer and clip gradients if their norms exceeding 0.1. In the first stage, our model is pre-trained from scratch for 600K iterations. In the second stage, the pre-trained model is fine-tuned on the labeled datasets for 300K iterations using a OneCycle learning rate policy. We set the batch size to 16. The final model is obtained by applying Exponential Moving Average (EMA) with a decay of 0.9999, which helps stabilize the model parameters during training.

4.3. Datasets and Evaluation Metrics

KITTI Depth Completion Dataset [53] serves as a standard benchmark for depth completion in autonomous driving scenarios. It provides over 93,000 training samples, including sparse LiDAR depth maps, their corresponding RGB images, and ground truth depth maps. The sparse LiDAR depth maps are generated from Velodyne LiDAR scans, which typically cover only a small percentage of the image pixels (approximately 5%). The ground truth depth maps are obtained via multi-frame matching and cover approximately 20% image pixels. The dataset also includes 1,000 test samples on the private online benchmark.

NYU Depth V2 Dataset [48] is also a standard benchmark for indoor scene understanding, consisting of RGB-D data captured with a Kinect sensor. It provides 1,449 densely

Table 2. Performance on KITTI and NYUv2 datasets. For the KITTI dataset, results are evaluated by the KITTI testing server and ranked by the RMSE (in mm). For the NYUv2 dataset, we report their performance on the official test set in their papers.

Method	Year	KITTI				NYUv2		
		RMSE ↓	MAE ↓	iRMSE ↓	iMAE ↓	RMSE ↓	REL ↓	$\delta_{1.25}$ ↑
S2D [38]	2018	814.73	249.95	2.80	1.21	0.230	0.044	97.1
CSPN [8]	2019	1019.64	279.46	2.93	1.15	0.117	0.016	99.2
DeepLiDAR [45]	2019	758.38	226.50	2.56	1.15	0.115	0.022	99.3
CSPN++ [9]	2020	743.69	209.28	2.07	0.90	0.101	0.015	99.5
GuideNet [49]	2020	736.24	218.83	2.25	0.99	0.115	–	–
FCFR [35]	2021	735.81	217.15	2.20	0.98	0.106	0.015	99.5
ACMNet [80]	2021	744.91	206.09	2.08	0.90	0.105	0.015	99.4
NLSPN [41]	2020	741.68	199.59	1.99	0.84	0.092	0.012	99.6
PointDC [74]	2023	736.07	201.87	1.97	0.87	0.089	0.012	99.6
RigNet [67]	2022	712.66	203.25	2.08	0.90	0.090	0.013	99.6
DySPN [34]	2022	709.12	192.71	1.88	0.82	0.090	0.012	99.6
BEV@DC [82]	2023	697.44	189.44	1.83	0.82	0.089	0.012	99.6
CFormer [79]	2023	708.87	203.45	2.01	0.88	0.090	0.012	–
LRRU [56]	2023	696.51	189.96	1.87	**0.81**	0.091	0.011	99.6
TPVD [70]	2024	693.97	188.60	**1.82**	**0.81**	0.086	**0.010**	**99.7**
ImprovingDC [57]	2024	686.46	**187.95**	1.83	**0.81**	0.091	0.011	99.6
BP-Net [50]	2024	684.90	194.69	**1.82**	0.84	0.089	0.012	99.6
DMD^{3}C (Ours)	-	**678.12**	194.46	**1.82**	0.85	**0.085**	0.011	**99.7**

labeled images from various indoor scenes. Missing ground truth is mainly due to the difference in perspective between the depth sensor and the camera.

Evaluation Metrics used for KITTI are root mean squared error (RMSE), mean absolute error (MAE), root mean squared error of the inverse depth (iRMSE), and mean absolute error of the inverse depth (iMAE), among which RMSE is used for ranking. These metrics are obtained by submission and evaluation from a non-public test set. For NYU, we follow the previous work and report the RMSE, mean absolute relative error (REL), and δ_θ, which represents the percentage of pixels whose error is less than a threshold θ.

4.4. Comparison with State-of-the-art Methods

Evaluation on Depth Completion Benchmarks. We evaluate our method on the test sets of the NYUv2 dataset [48] and KITTI dataset [53]. Table 2 shows the quantitative comparison of our method and other top ranking published methods. On the KITTI leaderboard, our proposed DMD^{3}C ranks 1st outperforming all other methods under the primary RMSE metric at the time of paper submission. It also has comparable performance under other evaluation metrics. On the NYUv2 dataset, our method achieves the best RMSE and best $\delta_{1.25}$. Additionally, DMD^{3}C maintains competitive results across several other evaluation metrics, further confirming its robustness and generalization.

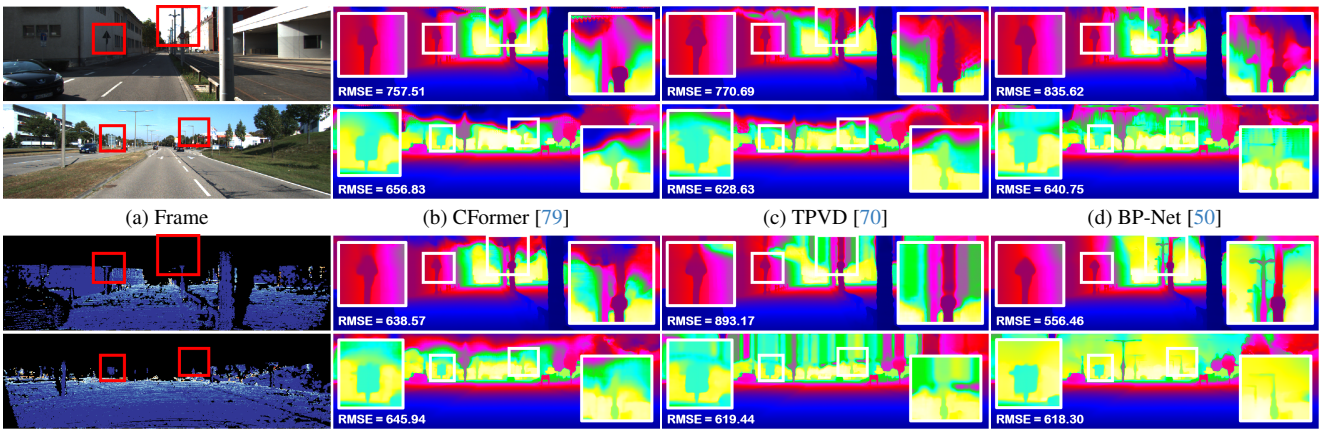
To enhance understanding of DMD^{3}C's superiority over other state-of-the-art methods, we provide visual comparisons. Figure 4 provides qualitative comparisons with other SOTA methods on the KITTI test set using publicly available results. We use results from the official benchmark[1] produced by the best models for method comparison. Our DMD^{3}C model excels in maintaining sharp object boundaries and capturing fine details in areas where other models encounter difficulties, particularly in scenes with complex structures or objects at varying distances. In such challenging regions, other methods fail to estimate accurate depth. Furthermore, as shown in the error maps, our method demonstrates a clear advantage even in regions without ground truth, despite these areas not being factored into the evaluation metrics.

Not only does our model retain the fine details in the depth map, but it also maintains structural integrity in the potential three-dimensional space. Figure 5 illustrates that our proposed two-stage distillation improves performance in the 3D space by visualizing point clouds from the completed depth maps. In comparison to other methods, our DMD^{3}C produces a more coherent and complete 3D structure, further outperforming existing methods.

4.5. Discussions

Effect of Pre-training in the First Stage. First, to evaluate the impact of the proposed pre-training strategy using unlabeled images, we conduct an experiment where the depth completion model is trained from scratch, without any pre-training. As shown in Table 3, removing the pre-training

[1] https://www.cvlibs.net/datasets/kitti/



(a) Frame (b) CFormer [79] (c) TPVD [70] (d) BP-Net [50]

(e) Error Map (f) LRRU [56] (g) ImprovingDC [57] (h) DMD^{3}C (Ours)

Figure 4. Qualitative comparison of our proposed DMD^{3}C with several state-of-the-art methods on the KITTI benchmark, using public test results. Error maps highlight pixels with ground truth. In regions lacking ground truth, our method demonstrates notable improvements in depth completion, even though these areas are excluded from the evaluation metrics.

Table 3. Ablations on the proposed main components.

Method	KITTI			
	RMSE ↓	MAE ↓	iRMSE ↓	iMAE ↓
w/o Pre-train	682.34	194.96	**1.82**	**0.84**
w/o SSI Loss	684.54	195.65	1.86	0.85
DMD^{3}C	**678.12**	**194.46**	**1.82**	0.85

Table 4. Ablations on different network architectures.

Method	KITTI			
	RMSE ↓	MAE ↓	iRMSE ↓	iMAE ↓
LRRU [56]	696.51	189.96	1.87	0.81
+ Ours	**693.17**	**189.60**	**1.85**	**0.80**
CFormer [79]	764.87	183.88	1.89	**0.80**
+ Ours	**760.29**	**183.62**	**1.88**	**0.80**
BP-Net [50]	684.90	194.69	**1.82**	**0.84**
+ Ours	**678.12**	**194.46**	**1.82**	0.85

step results in a performance degradation, with an increase in RMSE by 4.22 mm. This demonstrates that pre-training plays a crucial role in enhancing the model's ability to generalize, enabling it to better capture complex geometric features across diverse scenes.

Effect of SSI Loss in the Second Stage. Next, we show the impact of removing SSI Loss in the second distillation stage, while retaining the standard supervised loss with sparse ground truth. As shown in Table 3, the results indicate that the proposed SSI Loss improves the model's performance. This demonstrates the benefit of aligning monocular depth supervision with real-world scale using SSI Loss.

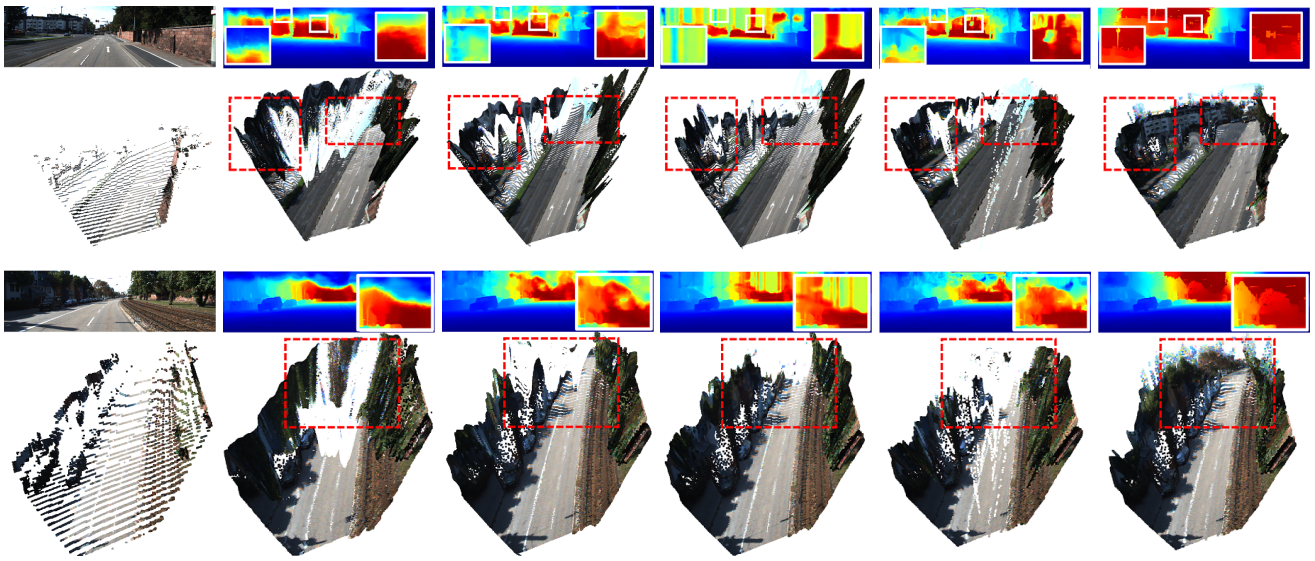
Table 5. Zero-shot performance comparison with other methods on out-of-the-domain datasets.

Method	ScanNet		DDAD		VOID1500	
	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓
CFormer	0.120	0.232	9.606	3.328	0.726	0.261
LRRU	0.132	0.245	9.164	2.738	0.698	0.232
BP-Net	0.122	0.212	8.903	2.712	0.704	0.230
Ours	**0.101**	**0.210**	**7.766**	**2.498**	**0.676**	**0.225**

Network Architectures. We evaluate the compatibility of our method with different network architectures, as shown in Table 4. Since we focus on training strategies, we have the flexibility to various models, demonstrating the robustness of our approach across different network designs. Specifically, we evaluate BP-Net [50], LRRU [56], and CFormer (L1 Loss) [79]. BP-Net is one of the most advanced methods, while LRRU and CFormer are considered representative architectures. Our approach consistently improves performance across all three models. Notably, BP-Net with our method achieves the best performance in terms of RMSE, the primary evaluation metric.

Generalization to Out-of-Domain Datasets. To evaluate the generalization ability of our method, we conduct zero-shot testing on unseen datasets, including ScanNet [13], DDAD [17], and VOID1500 [60], as shown in Table 5. Our approach consistently achieves the best performance across all datasets, demonstrating lower RMSE and MAE compared to prior methods. Notably, our method outperforms BP-Net by a significant margin on DDAD, reducing RMSE from 8.903 to 7.766, highlighting its effectiveness in handling unseen scenarios.

Application on Dense SLAM. Figure 6 compares a representative SLAM method, Droid-SLAM [51], with sparse LiDAR points (left) and our method with dense depth com-



(a) Input (b) CFormer [79] (c) LRRU [56] (d) ImprovingDC [57] (e) BP-Net [50] (f) DMD^{3}C (Ours)

Figure 5. Qualitative comparison of depth completion methods. This figure demonstrates the performance of various depth completion models, including CFormer, LRRU, ImprovingDC, BP-Net, and our proposed DMD^{3}C. For each method, we show the input RGB images with sparse LiDAR points (left), along with the resulting completed depth maps and corresponding 3D point cloud reconstructions.

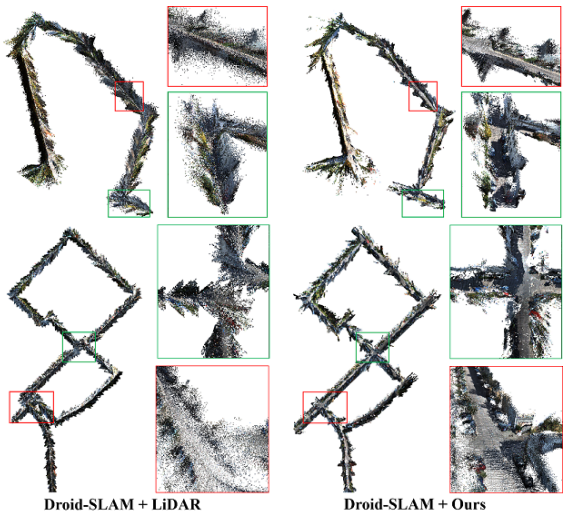


Figure 6. Qualitative comparison of Droid-SLAM with sparse LiDAR points and our method with dense depth completion.

pletion (right). Droid-SLAM suffers from structural distortions and noise, especially in occluded or low-texture areas (red boxes), due to limited depth information. In contrast, our method generates more consistent reconstructions, preserving structural details and improving alignment (green boxes). These results highlight the advantage of dense depth completion in enhancing SLAM quality.

5. Conclusion

In this work, we present DMD^{3}C, a novel two-stage distillation framework that distills knowledge from monocular depth estimation foundation models into the depth completion task. In the first stage, we introduce a data generation strategy that leverages monocular depth estimation and mesh reconstruction to simulate training data, allowing the model to learn geometric features from diverse natural images. In the second stage, we propose a scale- and shift-invariant loss (SSI Loss) combined with a supervised L1 loss with sparse ground truth, which addresses the scale ambiguity in monocular depth estimation. This ensures consistent depth completions across varying scales and focuses on learning real-world scale information. Extensive experiments on the depth completion benchmarks demonstrate that DMD^{3}C achieves state-of-the-art performance, ranking first on the KITTI leaderboard, and significantly outperforming existing methods. Our results highlight the framework's ability to produce high-quality depth maps with improved detail and structural consistency, making it a promising solution for depth completion tasks.

Acknowledgements

This work was supported by the National Key R&D Program of China (2022YFC3300704), the National Natural Science Foundation of China (62331006, 62171038, and 62088101), and the Fundamental Research Funds for the Central Universities.

References

[1] Luca Bartolomei, Matteo Poggi, Andrea Conti, Fabio Tosi, and Stefano Mattoccia. Revisiting depth completion from a stereo matching perspective for cross-domain generalization. In *International Conference on 3D Vision (3DV)*, pages 1360–1370, 2024. 2

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 2

[3] 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 *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631, 2020. 5

[4] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *International Conference on 3D Vision (3DV)*, pages 667–676, 2017. 2

[5] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *International Journal of Computer Vision*, 131(8):2198–2218, 2023. 2

[6] Linwei Chen, Ying Fu, Lin Gu, Chenggang Yan, Tatsuya Harada, and Gao Huang. Frequency-aware feature fusion for dense image prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10763–10780, 2024. 2

[7] Yun Chen, Bin Yang, Ming Liang, and Raquel Urtasun. Learning joint 2d-3d representations for depth completion. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10023–10032, 2019. 1

[8] Xinjing Cheng, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(10):2361–2379, 2019. 2, 6

[9] Xinjing Cheng, Peng Wang, Chenye Guan, and Ruigang Yang. Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In *the AAAI Conference on Artificial Intelligence (AAAI)*, pages 10615–10622, 2020. 2, 6

[10] Keunhoon Choi, Somi Jeong, Youngjung Kim, and Kwanghoon Sohn. Stereo-augmented depth completion from a single rgb-lidar image. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 13641–13647, 2021. 2

[11] Andrea Conti, Matteo Poggi, and Stefano Mattoccia. Sparsity agnostic depth completion. In *the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 5871–5880, 2023. 2

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 5

[13] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 7

[14] Zheng Fang, Ying Fu, and Lixiong Liu. A dual of transformer features-related map-intelligent generation method. *Journal of Image and Graphics*, 28(10):3281–3294, 2023. 2

[15] Ying Fu, Zichun Wang, Tao Zhang, and Jun Zhang. Low-light raw video denoising with a high-quality realistic motion dataset. *IEEE Transactions on Multimedia*, 25:8119–8131, 2022. 2

[16] Qi Guan, Zihao Sheng, and Shibei Xue. Hrpose: Real-time high-resolution 6d pose estimation network using knowledge distillation. *Chinese Journal of Electronics*, 32(1):189–198, 2023. 2

[17] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2485–2494, 2020. 7

[18] Wenxuan Guo, Yingping Liang, Zhiyu Pan, Ziheng Xi, Jianjiang Feng, and Jie Zhou. Camera-lidar cross-modality gait recognition. *arXiv preprint arXiv:2407.02038*, 2024. 2

[19] Wenxuan Guo, Zhiyu Pan, Yingping Liang, Ziheng Xi, Zhicheng Zhong, Jianjiang Feng, and Jie Zhou. Lidar-based person re-identification. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17437–17447, 2024. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[21] Junjie Hu, Chenyu Bao, Mete Ozay, Chenyou Fan, Qing Gao, Honghai Liu, and Tin Lun Lam. Deep depth completion from extremely sparse data: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(7): 8244–8264, 2022. 1, 2

[22] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1–18, 2024. 2

[23] Jinwoo Jeon, Hyunjun Lim, Dong-Uk Seo, and Hyun Myung. Struct-mdc: Mesh-refined unsupervised depth completion leveraging structural regularities from visual slam. *IEEE Robotics and Automation Letters*, 7(3):6391–6398, 2022. 1

[24] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Eurographics Symposium on Geometry Processing*, 2006. 4

[25] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9492–9502, 2024. 1, 2

[26] Muhammad Ahmed Ullah Khan, Danish Nazir, Alain Pagani, Hamam Mokayed, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. A comprehensive survey of depth completion approaches. *Sensors*, 22(18):6969, 2022. 1, 2

[27] Byeong-Uk Lee, Hae-Gon Jeon, Sunghoon Im, and In So Kweon. Depth completion with deep geometry and context guidance. In *International Conference on Robotics and Automation (ICRA)*, pages 3281–3287. IEEE, 2019. 1

[28] Byeong-Uk Lee, Kyunghyun Lee, and In So Kweon. Depth completion using plane-residual representation. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13916–13925, 2021. 1

[29] Miaoyu Li and Ying Fu. Joint self-attention transformer for multispectral and hyperspectral image fusion. *Journal of Image and Graphics*, 28(12):3922–3934, 2023. 2

[30] Miaoyu Li, Ying Fu, Tao Zhang, and Guanghui Wen. Supervise-assisted self-supervised deep-learning method for hyperspectral image restoration. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 1

[31] Yingping Liang and Ying Fu. Relation-guided adversarial learning for data-free knowledge transfer. *International Journal of Computer Vision*, pages 1–18, 2024. 3

[32] Yingping Liang, Jiaming Liu, Debing Zhang, and Ying Fu. Mpi-flow: Learning realistic optical flow with multiplane images. In *the IEEE International Conference on Computer Vision (ICCV)*, pages 13857–13868, 2023. 2, 3

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 5

[34] Yuankai Lin, Tao Cheng, Qi Zhong, Wending Zhou, and Hua Yang. Dynamic spatial propagation network for depth completion. In *the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1638–1646, 2022. 2, 6

[35] Lina Liu, Xibin Song, Xiaoyang Lyu, Junwei Diao, Mengmeng Wang, Yong Liu, and Liangjun Zhang. Fcfr-net: Feature fusion based coarse-to-fine residual learning for depth completion. In *the AAAI Conference on Artificial Intelligence (AAAI)*, pages 2136–2144, 2021. 6

[36] Qiankun Liu, Yuqi Jiang, Zhentao Tan, Dongdong Chen, Ying Fu, Qi Chu, Gang Hua, and Nenghai Yu. Transformer based pluralistic image completion with reduced information loss. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6652–6668, 2024.

[37] Qiankun Liu, Yichen Li, Yuqi Jiang, and Ying Fu. Siamese-detr for generic multi-object tracking. *IEEE Transactions on Image Processing*, 33:3935–3949, 2024.

[38] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *IEEE international conference on robotics and automation (ICRA)*, pages 4796–4803, 2018. 2, 6

[39] Fabiola Maffra, Lucas Teixeira, Zetao Chen, and Margarita Chli. Real-time wide-baseline place recognition using depth completion. *IEEE Robotics and Automation Letters*, 4(2): 1525–1532, 2019. 1

[40] Hyoungseob Park, Anjali Gupta, and Alex Wong. Test-time adaptation for depth completion. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20519–20529, 2024. 2

[41] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *the European Conference on Computer Vision (ECCV)*, pages 120–136, 2020. 6

[42] Jin-Hwi Park and Hae-Gon Jeon. A simple yet universal framework for depth completion. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[43] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 724–732, 2016. 5

[44] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10106–10116, 2024. 2

[45] Jiaxiong Qiu, Zhaopeng Cui, Yinda Zhang, Xingdi Zhang, Shuaicheng Liu, Bing Zeng, and Marc Pollefeys. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3313–3322, 2019. 2, 6

[46] 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 (TPAMI)*, 44(3), 2022. 1, 2, 3, 5

[47] Ali Osman Serhatoglu, Oguzhan Guclu, and Ahmet Burak Can. Rgb-d slam with deep depth completion. In *International Conference on Artificial Intelligence and Soft Computing (ICAISC)*, pages 59–67, 2022. 1

[48] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *the European Conference on Computer Vision (ECCV)*, pages 120–136, 2012. 2, 5, 6

[49] Jie Tang, Fei-Peng Tian, Wei Feng, Jian Li, and Ping Tan. Learning guided convolutional network for depth completion. *IEEE Transactions on Image Processing (TIP)*, 30: 1116–1129, 2020. 2, 6

[50] Jie Tang, Fei-Peng Tian, Boshi An, Jian Li, and Ping Tan. Bilateral propagation network for depth completion. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9763–9772, 2024. 1, 2, 5, 6, 7, 8

[51] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:16558–16569, 2021. 7

[52] Ye Tian, Ying Fu, and Jun Zhang. Transformer-based undersampled single-pixel imaging. *Chinese Journal of Electronics*, 32(5):1151–1159, 2023. 2

[53] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV)*, 2017. 1, 2, 3, 5, 6

[54] Haotian Wang, Meng Yang, and Nanning Zheng. G2-monodepth: A general framework of generalized depth inference from monocular rgb+ x data. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023. 2, 3

[55] Haotian Wang, Meng Yang, Xinhu Zheng, and Gang Hua. Scale propagation network for generalizable depth completion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024. 2

[56] Yufei Wang, Bo Li, Ge Zhang, Qi Liu, Tao Gao, and Yuchao Dai. Lrru: Long-short range recurrent updating networks for depth completion. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9422–9432, 2023. 1, 6, 7, 8

[57] Yufei Wang, Ge Zhang, Shaoqian Wang, Bo Li, Qi Liu, Le Hui, and Yuchao Dai. Improving depth completion via depth feature upsampling. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21104–21113, 2024. 1, 2, 5, 6, 7, 8

[58] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8927–8948, 2021. 1

[59] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2575–2584, 2020. 5

[60] Alex Wong, Xiaohan Fei, Stephanie Tsuei, and Stefano Soatto. Unsupervised depth completion from visual inertial odometry. *IEEE Robotics and Automation Letters*, 5(2): 1899–1906, 2020. 7

[61] Xiaopei Wu, Liang Peng, Honghui Yang, Liang Xie, Chenxi Huang, Chengqi Deng, Haifeng Liu, and Deng Cai. Sparse fuse dense: Towards high quality 3d detection with depth completion. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5418–5427, 2022. 1

[62] Yangchao Wu, Tian Yu Liu, Hyoungseob Park, Stefano Soatto, Dong Lao, and Alex Wong. Augundo: Scaling up augmentations for monocular depth completion and estimation. In *European Conference on Computer Vision (ECCV)*, pages 274–293, 2025. 2

[63] Weijian Xie, Guanyi Chu, Quanhao Qian, Yihao Yu, Hai Li, Danpeng Chen, Shangjin Zhai, Nan Wang, Hujun Bao, and Guofeng Zhang. Depth completion with multiple balanced bases and confidence for dense monocular slam. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2024. 1

[64] Zexiao Xie, Xiaoxuan Yu, Xiang Gao, Kunqian Li, and Shuhan Shen. Recent advances in conventional and deep learning-based depth completion: A survey. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 35(3):3395–3415, 2024. 2

[65] Yan Xu, Xinge Zhu, Jianping Shi, Guofeng Zhang, Hujun Bao, and Hongsheng Li. Depth completion from sparse lidar data with depth-normal constraints. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2811–2820, 2019. 1

[66] Zhiqiang Yan, Xiang Li, Kun Wang, Zhenyu Zhang, Jun Li, and Jian Yang. Multi-modal masked pre-training for monocular panoramic depth completion. In *European Conference on Computer Vision*, pages 378–395. Springer, 2022. 2

[67] Zhiqiang Yan, Kun Wang, Xiang Li, Zhenyu Zhang, Jun Li, and Jian Yang. Rignet: Repetitive image guided network for depth completion. In *the European Conference on Computer Vision (ECCV)*, pages 214–230, 2022. 1, 6

[68] Zhiqiang Yan, Xiang Li, Kun Wang, Shuo Chen, Jun Li, and Jian Yang. Distortion and uncertainty aware loss for panoramic depth completion. In *International Conference on Machine Learning*, pages 39099–39109. PMLR, 2023.

[69] Zhiqiang Yan, Kun Wang, Xiang Li, Zhenyu Zhang, Jun Li, and Jian Yang. Desnet: Decomposed scale-consistent network for unsupervised depth completion. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3109–3117, 2023.

[70] Zhiqiang Yan, Yuankai Lin, Kun Wang, Yupeng Zheng, Yufei Wang, Zhenyu Zhang, Jun Li, and Jian Yang. Tri-perspective view decomposition for geometry-aware depth completion. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4874–4884, 2024. 1, 2, 6, 7

[71] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10371–10381, 2024. 1, 2, 3

[72] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 1, 2, 3, 5

[73] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9043–9053, 2023. 2

[74] Zhu Yu, Zehua Sheng, Zili Zhou, Lun Luo, Si-Yuan Cao, Hong Gu, Huaqi Zhang, and Hui-Liang Shen. Aggregating feature point cloud for depth completion. In *the IEEE International Conference on Computer Vision (ICCV)*, pages 8732–8743, 2023. 2, 6

[75] Fan Zhang, Shaodi You, Yu Li, and Ying Fu. Atlantis: Enabling underwater depth estimation with stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11852–11861, 2024. 1

[76] Tao Zhang, Ying Fu, and Jun Zhang. Deep guided attention network for joint denoising and demosaicing in real image. *Chinese Journal of Electronics*, 33(1):303–312, 2024. 2

[77] Xuanmeng Zhang, Zhedong Zheng, Minyue Jiang, and Xiaoqing Ye. Self-ensembling depth completion via density-aware consistency. *Pattern Recognition*, 154:110618, 2024. 2

[78] Yinda Zhang and Thomas Funkhouser. Deep depth completion of a single rgb-d image. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 175–185, 2018. 1

[79] Youmin Zhang, Xianda Guo, Matteo Poggi, Zheng Zhu, Guan Huang, and Stefano Mattoccia. Completionformer: Depth completion with convolutions and vision transformers. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18527–18536, 2023. 1, 2, 5, 6, 7, 8

[80] Shanshan Zhao, Mingming Gong, Huan Fu, and Dacheng Tao. Adaptive context-aware multi-modal network for depth completion. *IEEE Transactions on Image Processing (TIP)*, 30:5264–5276, 2021. 2, 6

[81] Jin Zheng, Botao Jiang, Wei Peng, and Qiaohui Zhang. Multi-scale binocular stereo matching based on semantic association. *Chinese Journal of Electronics*, 33(4):1010–1022, 2024. 2

[82] Wending Zhou, Xu Yan, Yinghong Liao, Yuankai Lin, Jin Huang, Gangming Zhao, Shuguang Cui, and Zhen Li. Bev@ dc: Bird's-eye view assisted training for depth completion. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9233–9242, 2023. 2, 6

[83] Yunhao Zou, Ying Fu, Tsuyoshi Takatani, and Yinqiang Zheng. Eventhdr: From event to high-speed hdr videos and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):32–50, 2024. 2

[84] Yiming Zuo and Jia Deng. Ogni-dc: Robust depth completion with optimization-guided neural iterations. In *the European Conference on Computer Vision (ECCV)*, pages 78–95, 2025. 2