

SelfDeco: Self-Supervised Monocular Depth Completion in Challenging Indoor Environments

Jaehoon Choi^{1,2}, Dongki Jung¹, Yonghan Lee¹, Deokhwa Kim¹, Dinesh Manocha², and Donghwan Lee¹
¹NAVER LABS ²University of Maryland

Abstract—We present a novel algorithm for self-supervised monocular depth completion. Our approach is based on training a neural network that requires only sparse depth measurements and corresponding monocular video sequences without dense depth labels. Our self-supervised algorithm is designed for challenging indoor environments with textureless regions, glossy and transparent surfaces, moving people, longer and diverse depth ranges and scenes captured by complex ego-motions. Our novel architecture leverages both deep stacks of sparse convolution blocks to extract sparse depth features and pixel-adaptive convolutions to fuse image and depth features. We compare with existing approaches in NYUv2, KITTI and NAVERLABS indoor datasets, and observe 5 - 34 % improvements in root-means-square error (RMSE) reduction.

I. INTRODUCTION

Depth completion has been widely studied in the field of robot navigation, computer vision, and autonomous driving. Its goal is to convert a sparse depth map from active depth sensors such as LiDAR or RGB-D cameras to a dense depth map. In practice, robots with accurate sensing and dense depth maps have a lower chance of collision [29] and are better able to compute safe navigation routes through challenging environments [30], [31]. Currently, most recent methods for depth completion rely on supervised learning [22], [25], [24], requiring large amounts of high-quality and dense depth maps as ground truth. Unfortunately, generating dense depth maps is expensive and challenging due to the sparse and noisy depth values from other active sensors. Even existing expensive 3D LiDAR sensors only have a limited number of scan lines and provide sparse depth measurements.

Self-supervised methods with monocular videos that are able to train depth completion neural networks without the need for dense depth maps. These methods [2], [1], [3], [4] are based on an unsupervised learning framework that use depth and camera ego-motion from monocular videos and use photometric loss as supervision. This loss measures the difference between a reference image and a synthesized image obtained by the depth-guided reprojection of other views into the reference view. However, the previous self-supervised training process for depth completion [17], [16] may not work well on public indoor datasets such as NYUv2 [39]. This is due to large areas of textureless regions and scenes captured by complex ego-motions of a handheld camera [5], [8]. Large indoor environments are more challenging not only because of the above issues but also because of large regions of glossy and transparent surface, non-Lambertian surfaces, longer and diverse ranges, and moving people. To overcome these issues, our proposed training process requires

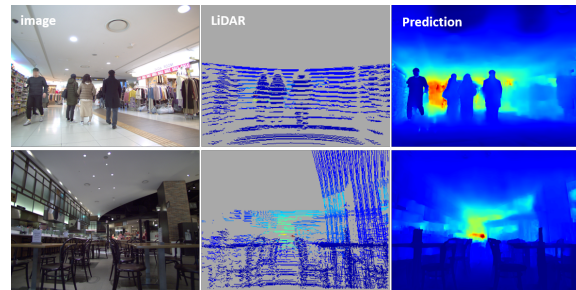



Fig. 1. Depth completion results in challenging scenes: Metro Station (top) and the Department Store (bottom) from NAVERLABS indoor dataset. Our method takes an RGB image and a sparse depth map (projection of LiDAR pointcloud) as input, and predicts a dense depth map. (low  high; grey means empty depth values.)

effective ways to compute complex camera ego-motions attached to a robot, mask out regions where photometric loss is unreliable, and output depth maps robust to diverse depth range. In this paper, we address the problem of developing self-supervised depth completion methods that work robustly on general indoor environments.

Main Results: We improve the training framework by leveraging learning-based local features to obtain accurate relative poses, making robust matching with auto-masking and minimum reprojection loss [4], and predicting inverse depth using a sigmoid function. However, we observe that merely masking out all regions still leads to depth artifacts including blurred boundaries and irregular patterns copied from sparse depth input. Therefore, we also propose a novel architecture with two unique components: sparsity-aware convolutions that extract features properly from sparse input and pixel-adaptive convolutions which propagate this information to the whole depth prediction with the guidance of image features. We highlight the effectiveness of our method by evaluating our approach on challenging datasets such as the Metro Station or the Department Store (Fig. 1). We also demonstrate the applicability of our method on datasets for which prior self-supervised depth estimation and completion fail to provide reliable depth maps. **Our novel contributions** can be summarized as follows:

- We develop a robust and stable training framework for self-supervised depth completion in challenging indoor environments.
- We propose a novel architecture that can extract sparse depth features guided by high-level image features.
- Our method outperforms prior self- and fully-supervised methods on challenging indoor datasets like NAVERLABS. We also achieve the state-of-the-art per-

formance among self-supervised methods on public benchmark datasets, including KITTI and the NYUv2 dataset based on various evaluation metrics.

II. RELATED WORK

Recently, depth completion methods using a deep neural network have shown promising results. Uhrig et al. [9] propose a sparsity-invariance convolution layer to handle sparse depth or features by using mask normalizations. Eldesokey et al. [11], [15] propose the normalized convolution to produce both a confidence map and dense depth output. In [25], they further extend their work into uncertainty estimation with depth completion methods. Teixeira et al. [23] use this confidence propagation from [11] to compute dense depth from LiDAR. Cheng et al. [14], [22] utilize the neural network to learn the affinity among neighboring pixels in order to perform the propagation process. In addition to using RGB images, some methods take advantage of information from other modalities, e.g., surface normals [13], [19], [20] or segmentation [12]. Yang et al. [16] predict the posterior over depth maps with a conditional prior network. Some works [27], [28] only focus on an indoor depth completion task. However, all aforementioned works on indoor scenarios are limited to a small-scale indoor environment and handle non-LiDAR indoor data (e.g., on the NYU-Depth-v2 [39] and Matterport3D [13] datasets). In this paper, we introduce challenging indoor scenarios such as the department store or metro station containing LiDAR sensor data.

The lack of dense depth maps is a significant limitation of supervised depth completion. As an alternative, several current works study self-supervised depth completion methods. Ma et al. [17] present self-supervised methods based on a photometric loss with pose estimation using the PnP method. Yang et al. [16] present a conditional prior network (CPN) that predicts a probability of depths at each pixel. Zhang et al. [18] propose an end-to-end framework to jointly estimate pose and depth without the PnP method. Yoon et al. [26] combine monocular depth estimation and Gaussian process-based depth regression for depth completion.

III. PROPOSED METHOD

A. Training Framework

In Fig. 2, our proposed depth completion model takes a target view image I_t and a sparse depth S_t as input, and outputs a dense inverse-depth \hat{d}_t . We invert \hat{d}_t to compute a dense depth map \hat{D}_t . In our method, the source view images contains its two adjacent temporal frames, i.e., $I_s \in \{I_{t-1}, I_{t+1}\}$, although including a larger temporal context is possible. Also, we estimate the relative pose $T_{t \rightarrow s}$ between the target image I_t and the set of source images I_s by using a Perspective-n-Point (PnP) [46] with Random Sample Consensus (RANSAC) [47].

1) *Relative Pose Estimation*: Although most prior works [4], [18] adopt learning-based pose estimation, we observe that learning-based pose estimation reveals degradation in performance when applied to challenging environments such as indoor scenes [5], [7], [8]. Additionally, according to

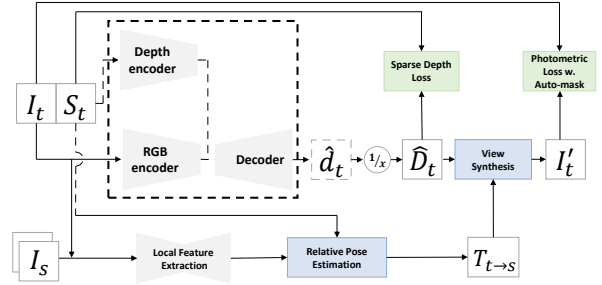


Fig. 2. Overview of our proposed framework. The black block with a dash-dotted line shows the depth completion network (details in Fig. 4).

recent findings [6], learning-based pose estimation methods suffer from scale ambiguity problems and make it difficult to provide absolute-scale depth supervision. Unlike self-supervised depth estimation, pose estimation via the PnP method is aligned with the depth completion task because our model needs to predict absolute scale depth.

Our method first detects learning-based features and computes their corresponding descriptors via learning-based algorithms. We select R2D2 [36], which shows better performance than traditional hand-crafted features on the visual localization benchmarks. Given the predicted feature in source and target images, we can compute high-quality correspondences by using a FLANN-based search algorithm [48] and then output the relative camera pose by solving PnP with corresponding sparse depths. Compared to hand-crafted features used by [17], dense correspondences obtained by R2D2 enable us to fully utilize sparse depth measurements with corresponding feature points for PnP pose estimation.

2) *Photometric Loss*: Given the camera intrinsic matrix K and the relative pose $T_{t \rightarrow s}$, we compute the warped pixel coordinates and synthesize target image I'_t from source images I_s via a bilinear sampling function. Following the self-supervised depth estimation [2], [1], we use the combination of the L1 and SSIM [33] as the photometric loss L_{ph} :

$$L_{ph}(I_t, I'_t) = \alpha \frac{1 - \text{SSIM}(I_t, I'_t)}{2} + (1 - \alpha) \|I_t - I'_t\|, \\ I'_t(x) = I_s \langle \pi(K T_{t \rightarrow s} \hat{D}_t(x) K^{-1} x_h) \rangle. \quad (1)$$

where $x_h = [x^T, 1]^T$ are the homogeneous coordinates from the target image, π gives pixel coordinates, and $\langle \cdot \rangle$ denotes the bilinear sampling function. We adopt the minimum reprojection loss and auto-masking from [4]. Our methods often face the occlusion issue, meaning that some pixels from frame I_t do not appear in either I_{t-1} or I_{t+1} frame due to moving people. To mitigate this occlusion issue, our per-pixel minimum photometric loss is defined as follows,

$$L_{photo} = \min_{I_s} L_{ph}(I_t, I'_t). \quad (2)$$

Additionally, auto-masking encourages our method to mask out static pixels for which the appearance does not change between the current frame and temporally adjacent frames. This auto-masking results in removing pixels belonging to

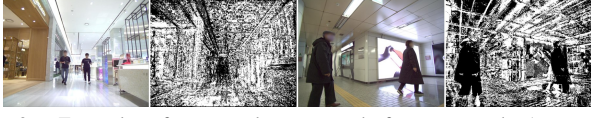


Fig. 3. Examples of auto-masks computed after one epoch. Auto masks (black pixels) usually contain low-texture scenes or non-Lambertain surfaces where photometric loss is unreliable.

low-texture scenes (e.g. ceiling or fluorescent light) that have undesirable effects on the minimum photometric loss. Furthermore, these masks often remove non-Lambertian surfaces (e.g. mirrors or digital bilboard screens), which are common in the department store or the metro station.

3) *Sparse Depth Loss*: Our proposed model can learn to encode absolute scale from sparse depth loss by aligning the depth prediction \hat{D}_t with sparse depth S_t . We use the L1-norm to penalize the difference between \hat{D}_t and S_t over Ω , where sparse depth is available,

$$L_{depth} = \sum_{x \in \Omega} |\hat{D}_t(x) - S_t(x)|. \quad (3)$$

4) *Total Loss Function*: Following [2], [17], [4], we also incorporate an edge-aware smoothness loss function L_{smooth} over the predicted depth map to regularize the depth in texture-less low-image gradient regions.

$$L_{smooth} = |\partial_x d^*| e^{-|\partial_x I_t|} + |\partial_y d^*| e^{-|\partial_y I_t|}, \quad (4)$$

where ∂_x, ∂_y denotes the gradients along the either the x or y direction. $d^* = \hat{d}_t / \hat{d}_t$ is the normalized inverse depth [3]. Our overall loss function is a weighted sum over the aforementioned losses,

$$L_{tot} = L_{photo} + \lambda_d L_{depth} + \lambda_s L_{smooth} \quad (5)$$

where λ_d and λ_s denote weighting terms selected through a grid search.

B. Network Architecture

In our training framework, the objective function defines a trade-off between a sparse depth loss, which enables our network to predict absolute scale depth on sparse regions, and a photometric loss, which encourages the network to predict dense depth where sparse information does not exist. In challenging indoor environments, self-supervised methods with photometric loss have shown that finding the optimal depth value is very difficult because the loss only comes from the appearance difference between target and synthesized images; measuring these differences is difficult due to the large non-texture regions, repeating structures, and occlusions.

The sparse depth loss might easily overwhelm the whole self-supervised training, and the depth predictions often result in overfitting on sparse depth input. Both feature extraction from the sparse depth and feature fusion schemes of multi-modal data are critical to avoid such negative effects. Our network adopts a late fusion scheme that uses each encoder for different sources because image and sparse depth features are heterogeneous. Throughout several experiments in Sec. IV-A.1, a late fusion scheme is shown to be particularly suitable for self-supervised depth completion. We focus

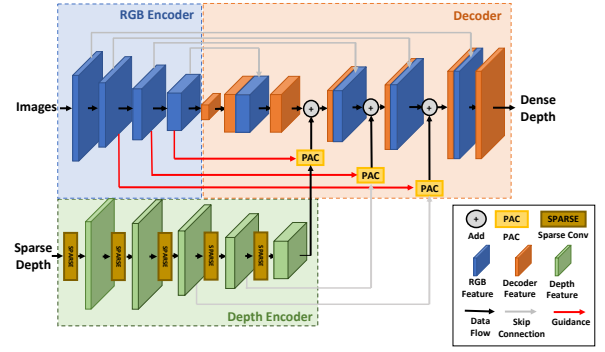


Fig. 4. Our model consists of three main components: (1) an RGB encoder that extracts RGB features, reducing the spatial resolution of the input images to consider the global context; (2) a depth encoder that extracts sparse depth features from sparse input via sparse convolutional blocks; (3) a decoder that takes both RGB and depth features, and transforms them into dense depth maps. Data flow contains convolution operations.

on designing an effective architecture that leverages sparse depth features to fill the empty regions rather than getting stuck producing sparse depth prediction.

1) *Sparse Convolution*: In order to extract features from sparse depth, recent depth completion methods fall into two groups. One group [9], [11] designs shallow networks with sparse convolution layers due to the poor performance of standard convolution layers on sparse input. In contrast, the other group [14], [17], [22] demonstrates that employing standard convolution layers yields better performance with their training schemes; these methods also perform well on different levels of sparsity. Comparing both approaches, we conclude that deep stacks of standard convolution layers are key to handling sparse depth input. Therefore, we choose to combine these two approaches by stacking sparse convolution layers [9] deeper. Sparse convolution blocks take a sparse depth feature and a binary observation mask as inputs, which both have the same spatial size. Since an observation mask indicates the reliability of each location of the feature maps, the convolution operations only attend on reliable input features which are involved with sparse depth input. As a result, the depth encoder is able to propagate depth information from reliable sparse depth input to its surrounding depth features. In contrast to sparsity invariant convolutions [9], we do not apply mask normalization, which causes a degradation problem due to the small values extracted from the depth encoder when stacking deep layers.

2) *Pixel-Adaptive Convolution (PAC)*: Recent works utilize operations like concatenation or element-wise addition to fuse the information from both the RGB and the depth encoder. However, this simple method is not the optimal way to fuse the information from two modalities because extracted depth features are not guided by RGB images in a late-fusion manner. To achieve this, we require spatially-variant convolution weights that different convolution kernels depending on image features are applied to different spatial positions of the depth features. Thus, we adopt pixel-adaptive convolutions (PACs) introduced by Su et al. [35]. Assume that the PACs take a sparse depth feature map \mathbf{v} and an image feature \mathbf{f} as input, and produce the output sparse depth feature

\mathbf{v}' . We denote the convolution weight as \mathbf{W} and the bias term as \mathbf{b} . PACs augment the spatially invariant convolution weight \mathbf{W} with a local-adaptive kernel function \mathbf{K} . Following the notation in [35], the PAC convolution is defined as

$$\mathbf{v}'_i = \sum_{j \in \Omega(i)} K(\mathbf{f}_i, \mathbf{f}_j) \mathbf{W}[\mathbf{p}_i - \mathbf{p}_j] \mathbf{v}_j + \mathbf{b} \quad (6)$$

where \mathbf{f} are image features from the RGB encoder and act as guidance for the PACs in the depth encoder. High-level image features are preferable because they contain image context information. The kernel function \mathbf{K} has a standard Gaussian form: $K(\mathbf{f}_i, \mathbf{f}_j) = \exp(-\frac{1}{2}(\mathbf{f}_i - \mathbf{f}_j)^T(\mathbf{f}_i - \mathbf{f}_j))$. This kernel computes the correlation between image features and guides the standard convolutional weights to generate depth features depending on the content learned from the RGB encoder. This spatially-variant kernel is helpful because, for instance, depth features on a person should not be applied to generate depth maps on the wall. Thus, PACs enable our proposed architecture to generate intermediate depth features consistent with image content.

3) *Decoder*: The decoder is similar to Ma et al [17] with one major difference. Unlike their method, our proposed network predicts inverse depth instead of depth. Thus, we replace the ReLU nonlinearity function with sigmoid activation. We observe that applying a sigmoid function is critical for training stability and is more robust to diverse depth ranges on challenging indoor scenes. Without inverse depth prediction, we fail to train the network proposed by [17] on our challenging indoor datasets.

IV. IMPLEMENTATION AND PERFORMANCE

A. Datasets

To demonstrate that our proposed method performs robustly in the challenging environment, we experiment with NAVERLABS indoor dataset. Additionally, we evaluate our method on two benchmarks: the NYUv2 dataset and KITTI. For indoor scenes, we report five evaluation metrics: root mean squared error (RMSE), mean absolute relative error (ABS Rel), and δ_i , which means the percentage of predicted pixels for which the relative error is less than a threshold i . The RMSE is represented in *mm* scale. For the KITTI benchmark, we use four official metrics for evaluation.

1) *NAVERLABS indoor dataset*: NAVERLABS dataset [37] comprises data collected from three different places: Department Store B1 (Dept. B1), Department Store 1F (Dept. 1F), and Metro Station (MS). All data were collected using a mapping robot equipped with two 16-channel LiDAR sensors and six RGB cameras with 2592×2048 resolution. LiDAR SLAM is used to estimate sensor poses for our dataset, and this pose information enables us to generate depth maps. We captured 17K images for Dept. 1F, 32K images for Dept. B1, and 61K images for MS. The MS dataset was captured at a highly crowded metro station.

We accumulate consecutive LiDAR sweeps and project the point clouds onto each of six cameras in order to generate depth maps. These depth maps can be used as sparse input and ground truth for evaluation. In order to create sparse

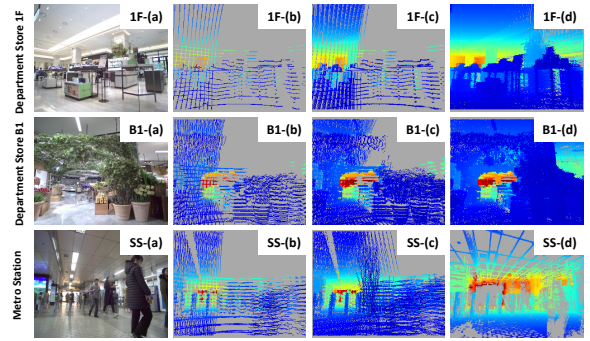


Fig. 5. Samples from the NAVERLABS indoor dataset. We include more samples and demonstrate many challenging features from our datasets in supplementary video.

input, we accumulate 0.3 seconds of LiDAR scans, which correspond to one LiDAR scan since the LiDAR is sampled at 10 Hz (see the second column in Fig. 5). In the Dept. 1F and Dept. B1 datasets, we collect 1 second of LiDAR scans to generate ground truth depth for training supervised depth completion algorithms (see the third column in Fig. 5). However, the depth maps generated from the MS dataset are too noisy to serve as ground truth due to noise caused by moving objects (see SS-(c) in Fig. 5). For evaluation, we select the high-quality depth maps generated from projecting a LiDAR 3D model onto the camera frames in the Dept. 1F and Dept. B1 datasets (see 1F-(d) and B1-(d) in Fig. 5). In contrast, we use the SfM method [42] to reconstruct a sparse 3D model of the MS dataset and MVS system [43] to produce dense depth maps instead of noisy depth maps from LiDAR point clouds. Given the pose from LiDAR SLAM, we can obtain a 3D model aligned with the LiDAR data. SS-(d) in Fig. 5 generated by the SfM is the example of depth maps for evaluation.

2) *NYU-Depth-v2 dataset*: The NYUv2 dataset [39] comprises RGB images and depth maps collected from Microsoft Kinect in 464 indoor scenes. The raw training dataset contains 268K images. We downsampled 10 times to remove redundant frames, resulting in 47k frames sampled for the training set. Our method is tested on the 654 official labeled test set for evaluation.

3) *KITTI depth completion dataset*: The KITTI depth completion dataset [9] provides RGB images, sparse lidar points, and dense ground truth depth. Following the official split, it contains 86k frames for the training set, 1k frames for the validation set, and 1k frames for the evaluation.

B. Implementation Details

We adopt ResNet34 [34] as the RGB encoder, and it was pretrained on ImageNet [41]. The proposed deep networks were implemented with PyTorch [40] and trained on 4 Tesla V100 GPUs. We trained our network from scratch for 30 epochs, with a batch size of 16 using Adam [44], where $\beta_1 = 0.9$, $\beta_2 = 0.999$; we used an input resolution of 1024×760 for NAVERLABS, 640×480 for NYUv2, and 1024×320 for KITTI. We used an initial learning rate of 10^{-4} for the first 10 epochs and halved it every 10 epochs. In loss functions, we set the α to 0.85, λ_d to 0.001 and λ_s to 0.1.

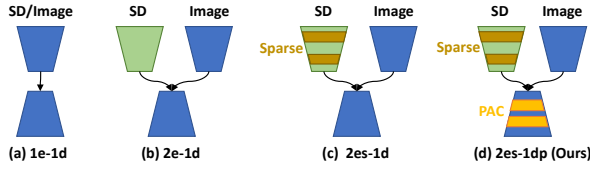


Fig. 6. SD represents a sparse depth. We design 4 architectures: (a) 1e-1d: an encoder and a decoder; (b) 2e-1d: an RGB encoder, a depth encoder, and a decoder; (c) 2es-1d: an RGB encoder, a depth encoder with sparse convolutions, and a decoder; (d) 2es-1dp: the same architecture as Fig. 4.

TABLE I

DEPTH COMPLETION RESULTS ON NAVERLABS INDOOR DATASET.

(a) Dept. 1F Dataset						
Method	T	RMSE ↓	ABS Rel ↓	$\delta_{1.25} \uparrow$	$\delta_{1.25^2} \uparrow$	$\delta_{1.25^3} \uparrow$
UNET [17]	S	5043.5	0.571	0.273	0.463	0.553
nUNET [15]	S	4993.8	0.171	0.898	0.933	0.951
CSPN [14]	S	2857.1	0.266	0.808	0.886	0.924
MD2 [4]	SS	3862.4	0.493	0.488	0.716	0.818
1e-1d	SS	2899.6	0.256	0.784	0.869	0.911
2e-1d	SS	2801.4	0.265	0.783	0.868	0.909
2es-1d	SS	2700.3	0.242	0.790	0.875	0.915
Ours	SS	2692.9	0.234	0.815	0.887	0.926

(b) Dept. B1 Dataset						
Method	T	RMSE ↓	ABS Rel ↓	$\delta_{1.25} \uparrow$	$\delta_{1.25^2} \uparrow$	$\delta_{1.25^3} \uparrow$
UNET [17]	S	3048.5	0.346	0.599	0.711	0.764
nUNET [15]	S	2963.1	0.324	0.635	0.732	0.779
CSPN [14]	S	2162.2	0.199	0.774	0.890	0.938
MD2 [4]	SS	3592.7	0.338	0.47	0.729	0.862
1e-1d	SS	2210.5	0.186	0.823	0.914	0.947
2e-1d	SS	2157.5	0.170	0.832	0.916	0.949
2es-1d	SS	2116.8	0.183	0.829	0.916	0.949
Ours	SS	2024.1	0.178	0.841	0.922	0.953

(c) Metro Station Dataset						
Method	T	RMSE ↓	ABS Rel ↓	$\delta_{1.25} \uparrow$	$\delta_{1.25^2} \uparrow$	$\delta_{1.25^3} \uparrow$
UNET [17]	S	4029.7	0.749	0.201	0.219	0.23
nUNET [15]	S	2705.1	0.472	0.444	0.528	0.582
MD2 [4]	SS	1321.2	0.166	0.8	0.916	0.939
Ours	SS	1248.4	0.138	0.866	0.923	0.942

“S” represents *supervised* training methods and “SS” denotes *self-supervised* training methods. Different from other depth completion methods, MD2 is based on self-supervised depth estimation with monocular video sequences. Compared to prior self- and fully- supervised methods, our proposed method demonstrates improvement in RMSE metric by 5.8% in Dept. 1F, by 6.4% in Dept. B1 and by 5.5% in Metro Station.

C. Experiments on NAVERLABS indoor datasets

We compare our proposed method with 4 baseline approaches with public codes: 1) UNET is a U-shaped ResNet [17]; 2) nUNET is an early-fusion encoder-decoder network combined with multiple normalization convolutions [15]; 3) CSPN learns the affinity to perform sparse depth propagation [14]; 4) Monodepth2 (MD2) is based on ResNet34 [4]. The same relative pose generated from our method is applied to training MD2 for a fair comparison. MD2 benefits from the use of ground truth median scaling at inference, while other depth completion methods utilize unmodified network prediction. Except for 4), the aforementioned methods are based on supervised training with ground truth depth. As mentioned in Section IV-A, it is difficult to create ground truth depth for the Metro Station dataset (MS). In the case of the MS dataset, we only use the sparse depth maps as weak supervisions for supervised depth completion. Quantitative results are shown in Table I, and the associated qualitative

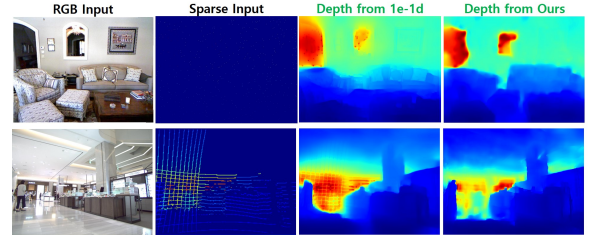


Fig. 7. Examples of depth artifacts generated by 1e-1d model. The top (NYUv2) and bottom (Dept. 1F) figures in third column shows irregular sparse patterns copied from sparse inputs and blurred depth boundaries

comparisons are shown in Fig. 8. We train the model from Ma et al. [17] with its network architecture and the relative poses from our method, which are more accurate. Note that Ma et al. failed to produce meaningful results. We observe that the network just copies the sparse inputs, meaning that the sparse depth loss only serves as supervision for training while the photometric loss is unreliable.

In contrast, our method successfully performs depth completion in the NAVLERLABS dataset. This indicates that our training framework is more robust to variant supervision compared to other self-supervised depth completion methods. The performance of our method is superior to existing supervised completion methods. From columns 4 through 6 of Fig. 8, we can observe that supervised depth completion methods rely heavily on the density of ground truth depth. We notice that these algorithms fail to fill the holes where LiDAR information does not exist (e.g., the upper part of the image) due to the sensor configuration. In contrast, thanks to the effectiveness of photometric loss, the proposed method succeed in filling empty regions such as the ceiling, which LiDAR sensors are not able to measure, with accurate depth values. In addition, our method even predicts the correct depth map on shiny or transparent surface which cannot be scanned accurately by sensors (see First row in Fig. 8). The experimental results of CSPN bolster this observation because our method achieves performance comparable to the supervised setting in Table I-(a), (b), while its training with weak supervision collapsed on the MS dataset. If we collect the high-quality dense depth maps (see the last column in Fig. 8), the supervised depth completion methods can be superior to our method. However, collecting dense depth maps requires a high cost and can be impractical in certain environments such as the MS dataset.

To investigate the impact of our network architecture, we conduct ablation studies for different network configurations on the NAVERLABS dataset. Figure 6 illustrates the network variants, and their quantitative results are shown in Table I. Comparing the results of 1e-1d with those of 2e-1d, we observe that fusing the RGB and depth feature at the earliest stage worsens the results. With the early fusion, we often observe depth with blurrier depth discontinuities, as seen in Fig. 7. With PAC, depth maps can eliminate either irregular lattice patterns from LiDAR or sparse points from sampled depth maps. We conclude that the late fusion is more effective for self-supervised learning because image and depth features are heterogeneous data, and early fusion

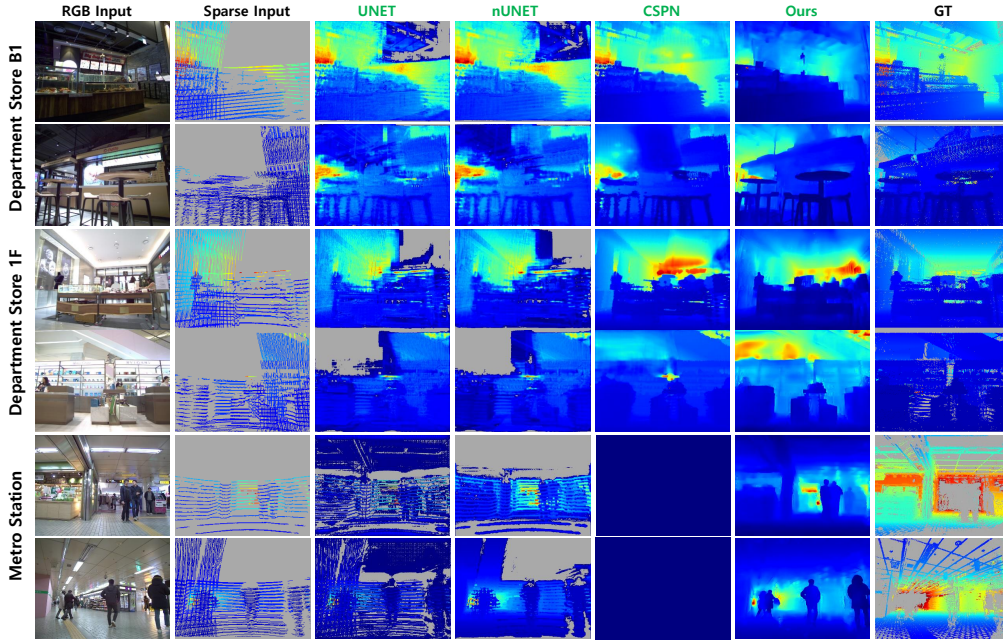


Fig. 8. Qualitative comparison with “UNET” [17], “nUNET” [15], and “CSPN” [14] on the NAVERLABS indoor dataset. As mentioned in Sec. IV-A.1, the groundtruth of the Metro Station dataset is the results of MVS. (low high; grey means empty depth values.)

TABLE II

DEPTH COMPLETION RESULTS ON THE NYUV2 DATASET

Method	T	RMSE ↓	ABS Rel ↓	$\delta_{1.25} \uparrow$	$\delta_{1.25^2} \uparrow$	$\delta_{1.25^3} \uparrow$
(a) 500 Samples						
Ma et al. [10]	S	0.204	0.043	97.8	99.6	99.9
CSPN [14]	S	0.117	0.016	99.2	99.9	100
Ma et al. [17]	SS	0.271	0.068	-	-	-
Ours	SS	0.178	0.033	98.1	99.7	100
(b) 200 Samples						
Ma et al. [10]	S	0.230	0.044	97.1	99.4	99.8
Yang et al. [16]	SS	0.569	0.171	-	-	-
Yoon et al. [26]	SS	0.309	-	-	-	-
Ours	SS	0.240	0.048	96.6	99.4	99.9

TABLE III

DEPTH COMPLETION RESULTS ON THE KITTI VALIDATION SET

Method	T	RMSE [mm]	MAE [mm]	iRMSE [1/km]	iMAE [1/km]
Uhrig et al. [9]	S	1601.33	481.27	4.94	1.78
Ma et al. [17]	S	814.73	249.95	2.80	1.21
Yoon et al. [26]	SS	1593.37	547.00	27.98	2.36
Ma et al. [17]	SS	1343.33	358.66	4.28	1.64
Yang et al. [16]	SS	1310.03	347.17	-	-
Ours	SS	1212.89	346.12	3.54	1.29

“S” represents *supervised* training methods and “SS” denotes *self-supervised* training methods. Compared to prior self-supervised method, our method reduces the RMSE by 34.3% in NYUv2 and by 7.4% in KITTI.

is more likely to produce noisy initial depth values.

D. Experiments on Public Benchmark Datasets

To verify the performance of our method on public datasets, we train and evaluate our method on the NYUv2 dataset. Following previous works, we uniformly sample 200 and 500 sparse depth points separately for sparse input. Except for supervised depth completion methods, our proposed method achieves the best performance among self-supervised depth completion methods in Table II.

In order to prove the generalization capability of our method, we train and evaluate our method on an outdoor dataset, the KITTI depth completion dataset. The comparison

results are shown in Table III. We observe that our method outperforms other self-supervised depth completion methods on every metric. Although Yang [16] pretrains their network using a virtual dataset such as the Virtual KITTI dataset [45], which is similar to the benchmark dataset, our method performs better than their method. We include qualitative results in the supplementary video.

V. CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

In this paper, we investigated the self-supervised depth completion framework for a challenging indoor environment. We provide two main research contributions. First, we introduced a training framework that leads to stable self-supervised training. Second, we proposed a novel architecture to combine image features and sparse depth features effectively. In indoor environments, our method outperforms other self-supervised methods and is competitive with supervised methods. Furthermore, it generalizes better to all public benchmark datasets. As part of future work, we would like to apply this method to complement depth sensors for mobile robots in indoor environments and evaluate its benefits for robot navigation. A limitation of the current method is the high computational cost for embedded platforms. For robotic tasks, we would like to explore network compression algorithms to reduce computational complexity and latency.

ACKNOWLEDGMENT

This work was supported by the Institute of Information communications Technology Planning Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2019-0-01309, Development of AI Technology for Guidance of a Mobile Robot to its Goal with Uncertain Maps in Indoor/Outdoor Environments). This work was supported in part by ARO Grants W911NF1910069, W911NF2110026 and W911NF1910315 and NSF grant 2031901.

REFERENCES

- [1] Zhou, Tinghui, Matthew Brown, Noah Snavely, and David G. Lowe. "Unsupervised learning of depth and ego-motion from video." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1851-1858. 2017.
- [2] Godard, Clément, Oisín Mac Aodha, and Gabriel J. Brostow. "Unsupervised monocular depth estimation with left-right consistency." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 270-279. 2017.
- [3] Wang, Chaoyang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. "Learning depth from monocular videos using direct methods." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2022-2030. 2018.
- [4] Godard, Clément, Oisín Mac Aodha, Michael Firman, and Gabriel J. Brostow. "Digging into self-supervised monocular depth estimation." In *Proceedings of the IEEE international conference on computer vision*, pp. 3828-3838. 2019.
- [5] Zhou, Junsheng, Yuwang Wang, Kaihuai Qin, and Wenjun Zeng. "Moving indoor: Unsupervised video depth learning in challenging environments." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8618-8627. 2019.
- [6] Sattler, Torsten, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe. "Understanding the limitations of cnn-based absolute camera pose regression." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3302-3312. 2019.
- [7] Zhao, Wang, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. "Towards Better Generalization: Joint Depth-Pose Learning without PoseNet." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9151-9161. 2020.
- [8] Bian, Jia-Wang, Huangying Zhan, Naiyan Wang, Tat-Jun Chin, Chunhua Shen, and Ian Reid. "Unsupervised Depth Learning in Challenging Indoor Video: Weak Rectification to Rescue." *arXiv preprint arXiv:2006.02708* (2020).
- [9] Uhrig, Jonas, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. "Sparsity invariant cnns." In *2017 international conference on 3D Vision (3DV)*, pp. 11-20. IEEE, 2017.
- [10] Mal, Fangchang, and Sertac Karaman. "Sparse-to-dense: Depth prediction from sparse depth samples and a single image." In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1-8. IEEE, 2018.
- [11] Eldesokey, Abdelrahman, Michael Felsberg, and Fahad Shahbaz Khan. "Propagating confidences through cnns for sparse data regression." *arXiv preprint arXiv:1805.11913* (2018).
- [12] Jaritz, Maximilian, Raoul De Charette, Emilie Wirbel, Xavier Perrotton, and Fawzi Nashashibi. "Sparse and dense data with cnns: Depth completion and semantic segmentation." In *2018 International Conference on 3D Vision (3DV)*, pp. 52-60. IEEE, 2018.
- [13] Zhang, Yinda, and Thomas Funkhouser. "Deep depth completion of a single rgb-d image." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 175-185. 2018.
- [14] Cheng, Xinjing, Peng Wang, and Ruigang Yang. "Depth estimation via affinity learned with convolutional spatial propagation network." In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 103-119. 2018.
- [15] Eldesokey, Abdelrahman, Michael Felsberg, and Fahad Shahbaz Khan. "Confidence propagation through cnns for guided sparse depth regression." *IEEE transactions on pattern analysis and machine intelligence* (2019).
- [16] Yang, Yanchao, Alex Wong, and Stefano Soatto. "Dense depth posterior (ddp) from single image and sparse range." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3353-3362. 2019.
- [17] Ma, Fangchang, Guilherme Venturini Cavalheiro, and Sertac Karaman. "Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera." In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 3288-3295. IEEE, 2019.
- [18] Zhang, Yilun, Ty Nguyen, Ian D. Miller, Shreyas S. Shivakumar, Steven Chen, Camillo J. Taylor, and Vijay Kumar. "Dfnet: Ego-motion estimation and depth refinement from sparse, noisy depth input with rgb guidance." *arXiv preprint arXiv:1903.06397* (2019).
- [19] Qiu, Jiaxiong, 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3313-3322. 2019.
- [20] Xu, Yan, Xinge Zhu, Jianping Shi, Guofeng Zhang, Hujun Bao, and Hongsheng Li. "Depth completion from sparse lidar data with depth-normal constraints." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2811-2820. 2019.
- [21] Zhong, Yiqi, Cho-Ying Wu, Suya You, and Ulrich Neumann. "Deep rgb-d canonical correlation analysis for sparse depth completion." In *Advances in Neural Information Processing Systems*, pp. 5331-5341. 2019.
- [22] Cheng, Xinjing, Peng Wang, Chenye Guan, and Ruigang Yang. "CSPN++: Learning Context and Resource Aware Convolutional Spatial Propagation Networks for Depth Completion." In *AAAI*, pp. 10615-10622. 2020.
- [23] Teixeira, Lucas, Martin R. Oswald, Marc Pollefeys, and Margarita Chli. "Aerial Single-View Depth Completion With Image-Guided Uncertainty Estimation." *IEEE Robotics and Automation Letters* 5, no. 2 (2020): 1055-1062.
- [24] Tang, Jie, Fei-Peng Tian, Wei Feng, Jian Li, and Ping Tan. "Learning guided convolutional network for depth completion." *arXiv preprint arXiv:1908.01238* (2019).
- [25] Eldesokey, Abdelrahman, Michael Felsberg, Karl Holmquist, and Michael Persson. "Uncertainty-Aware CNNs for Depth Completion: Uncertainty from Beginning to End." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12014-12023. 2020.
- [26] Yoon, Sungho, and Ayoung Kim. "Balanced Depth Completion between Dense Depth Inference and Sparse Range Measurements via KISS-GP." *arXiv preprint arXiv:2008.05158* (2020).
- [27] Huang, Yu-Kai, Tsung-Han Wu, Yueh-Cheng Liu, and Winston H. Hsu. "Indoor Depth Completion with Boundary Consistency and Self-Attention." In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0-0. 2019.
- [28] Senushkin, Dmitry, Ilia Belikov, and Anton Konushin. "Decoder Modulation for Indoor Depth Completion." *arXiv preprint arXiv:2005.08607* (2020).
- [29] Park, Jae Sung, and Dinesh Manocha. "Efficient probabilistic collision detection for non-Gaussian noise distributions." *IEEE Robotics and Automation Letters* 5, no. 2 (2020): 1024-1031.
- [30] Sathyaamoorthy, Adarsh Jagan, Utsav Patel, Tianrui Guan, and Dinesh Manocha. "Frozone: Freezing-Free, Pedestrian-Friendly Navigation in Human Crowds." *IEEE Robotics and Automation Letters* (2020).
- [31] Sathyaamoorthy, Adarsh Jagan, Jing Liang, Utsav Patel, Tianrui Guan, Rohan Chandra, and Dinesh Manocha. "Densecavoid: Real-time navigation in dense crowds using anticipatory behaviors." *arXiv preprint arXiv:2002.03038* (2020).
- [32] Hirschmuller, Heiko. "Stereo processing by semiglobal matching and mutual information." *IEEE Transactions on pattern analysis and machine intelligence* 30, no. 2 (2007): 328-341.
- [33] Wang, Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. "Image quality assessment: from error visibility to structural similarity." *IEEE transactions on image processing* 13, no. 4 (2004): 600-612.
- [34] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016.
- [35] Su, Hang, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. "Pixel-adaptive convolutional neural networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11166-11175. 2019.
- [36] Revaud, Jerome, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. "R2d2: Reliable and repeatable detector and descriptor." In *Advances in Neural Information Processing Systems*, pp. 12405-12415. 2019.
- [37] Min Young Chang, Suyong Yeon, Soohyun Ryu and Donghwan Lee. "SpoxelNet: Spherical Voxel-based Deep Place Recognition for 3D Point Clouds of Crowded Indoor Spaces" In *Intelligent Robots and Systems (IROS)*, 2020 IEEE/RSJ International Conference on. IEEE, 2020, pp. 8564-8570.
- [38] Geiger, Andreas, Philip Lenz, and Raquel Urtasun. "Are we ready for autonomous driving? the kitti vision benchmark suite." In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354-3361. IEEE, 2012.
- [39] Silberman, Nathan, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. "Indoor segmentation and support inference from rgb-d images." In

- European conference on computer vision, pp. 746-760. Springer, Berlin, Heidelberg, 2012.
- [40] Paszke, Adam, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. "Automatic differentiation in pytorch." (2017).
 - [41] Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang et al. "Imagenet large scale visual recognition challenge." *International journal of computer vision* 115, no. 3 (2015): 211-252.
 - [42] Schonberger, Johannes L., and Jan-Michael Frahm. "Structure-from-motion revisited." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4104-4113. 2016.
 - [43] Schönberger, Johannes L., Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. "Pixelwise view selection for unstructured multi-view stereo." In *European Conference on Computer Vision*, pp. 501-518. Springer, Cham, 2016.
 - [44] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).
 - [45] Gaidon, Adrien, Qiao Wang, Yohann Cabon, and Eleonora Vig. "Virtual worlds as proxy for multi-object tracking analysis." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4340-4349. 2016.
 - [46] Lepetit, Vincent, Francesc Moreno-Noguer, and Pascal Fua. "Epnnp: An accurate $O(n)$ solution to the pnp problem." *International journal of computer vision* 81, no. 2 (2009): 155.
 - [47] Fischler, Martin A., and Robert C. Bolles. "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography." *Communications of the ACM* 24, no. 6 (1981): 381-395.
 - [48] Muja, Marius, and David G. Lowe. "Fast approximate nearest neighbors with automatic algorithm configuration." *VISAPP* (1) 2, no. 331-340 (2009): 2.