2022 IEEE International Conference on Robotics and Automation (ICRA)
May 23-27, 2022. Philadelphia, PA, USA

Unsupervised Depth Completion and Denoising for RGB-D Sensors

Lei Fan, Yunxuan Li, Chen Jiang and Ying Wu

***Abstract*— Depth information is considered valuable as it describes geometric structures, which benefits various robotic tasks. However, the depth acquired by RGB-D sensors still suffers from two deficiencies, i.e., *incompletion* and *noises*. Previous methods complete depth by exploring hand-tuned models or raising surface assumptions, while nowadays, deep approaches intend to solve this problem with rendered image pairs. For depth denoising, as a consequence of different sensor mechanisms, most methods can only work under specific devices. With existing methods, three challenges emerge: the onerous training set collecting process, the mismatch between existing models and present RGB-D sensors, and the non-real-time computation. In this paper, we first state depth completion and denoising are inherently different and without the need to collect or render complete and noiseless ground truths. We address all mentioned challenges with two separate unsupervised learning procedures. The completion network takes color and incomplete depth as input and predicts values to the unobserved area, which combines prior knowledge and color-depth correlations. The denoising step exploits image sequences to construct noise models in a self-supervised manner with the ability to cater to different sensors. Experimental comparisons and ablation studies demonstrate that even without human-labeled ground truths, the proposed method could produce better completion results and also reduce noises in real-time.**

I. Introduction

Depth sensing provides an important information dimension for various visual and robotic tasks. Compared to RGB cameras providing texture and color knowledge, depth sensors capture the underlying structure of the environment. Recent works, including but not limited to semantic segmentation [6], object detection [4], action recognition [33] and visual localization [28] already achieve improvements with additional depth information. However, recent leading commodity-level RGB-D cameras like Intel RealSense and Microsoft Kinect still have deficiencies like incompletion and sensor noises. As demonstrated in Fig. 1, these drawbacks bring ambiguities and impede further depth-related applications.

The blank and noise in depth are closely related to the latent operations of sensors. In general, depth sensors can be categorized into active and passive based on whether they interact with the real world and can be categorized into stereo, Time-of-Flight (ToF) and structure light according to their techniques. Therefore, most depth completion and denoising methods address these problems under a specific camera setting [29]. Recently, researches [35], [37] are focusing on achieving this goal using deep data-driven approaches. To train their models, intact and less-noise training image pairs need to be rendered first by reconstructing the whole environment. Besides building such a dataset is time-consuming and expensive work, they can only cover limited RGB-D sensor types and indoor scene categories.

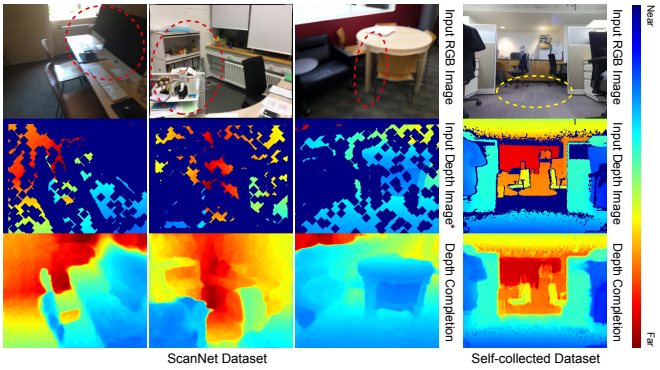


Fig. 1: Depth completion results on the public ScanNet dataset [7] and also the self-collected dataset captured with Microsoft Azure Kinect. We circle two kinds of completion problems with red and yellow dot lines, respectively.

Recently, Ma *et al.* utilize the photometric loss to addresses LiDAR depth completion [25]. Different from the LiDAR depth completion process, blanks in depth are not regularly distributed for RGB-D sensors. Nevertheless, a blessing for RGB-D depth completion is the percentage of valid depth data, which could provide valuable correlation information between colors and depths. In this way, we separate depth absence into two kinds. The first kind that cannot be inferred from current valid color and depth requires prior knowledge that should be learned like the monocular depth prediction. The second kind of missing can be propagated from existing depth for their color and texture similarities. These two types of missings are marked with red and yellow dot circles in Fig. 1, respectively.

In this paper, the contributions go with the following two observations and insights. (1) The prior knowledge for depth completion is not influenced by the sensors being adopted, which enables us to learn such information from existing RGB-D datasets. (2) For robotic and other practical applications, the depth noise model could change thoroughly with different sensor types. Hence, the two problems are better addressed with separate learnable models. For depth completion, we first produce further degrade depth images from already incomplete sensor inputs, which share similar ideas with the random masking on images. One of the learning objectives of our designed end-to-end trainable network is to fill these newly generated missing areas. Affinity matrices are additionally predicted in our decoder for the

Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, 60208

2022 IEEE International Conference on Robotics and Automation (ICRA) | 978-1-7281-9681-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICRA46639.2022.9812392

final spatial diffusion process, which especially recovers the second kind of missings in Fig. 1. Moreover, low-level partial convolution layers are applied to extract depth features to prevent the negative effect of invalid depth values. For the depth denoising step, we treat it as the following stage of depth completion, which is trained with the photometric loss under the self-supervision fashion. In reality, we can adapt the depth denoising models to different sensor types by train on collected new sequences. In sum, the proposed method could achieve both depth completion and denoising without any human-labeling process.

II. Related Work

As a fundamental problem of computer vision and robotics, there is an abundance of researches focusing on depth prediction, depth completion and depth denoising. Here we mainly summarize works closely related to our method.

Depth prediction from a monocular camera. Estimating depth from a single color image is an ill-posed problem according to its inherent ambiguities [18]. The depth prediction problem is quite similar to the focus of this paper, i.e., color-guided depth completion, especially when inferring large unobserved areas. Compared to classic methods [36], recent approaches adopt deep networks [3], [18], [20], [38]. Laina *et al.* performed residual learning to predict dense depth [18]. Godard *et al.* proposed a novel training loss for unsupervised monocular depth estimation, which relies on the left-right consistency [12]. Fu *et al.* treated the depth prediction as an ordinal regression problem that could achieve faster convergence compared to previous methods [11]. All of these methods demonstrate their ability to transfer learned prior knowledge from the training set into depth prediction of new color images. They are not suitable for the depth completion task since they could not mine the existing correlation between valid depth and color values.

Depth completion of RGB-D sensors. Missing values in the depth image from RGB-D sensors are often irregularly scattered compared to LiDAR depth images [32]. Various inpainting methods have been applied to the specific RGB-D depth completion and can be categorized into *color-guided* and *depth-only* approaches. The *color-guided* method includes ones utilizing fast marching [13], anisotropic diffusion [10], [22], low-rank matrix completion [24] and bilateral filters [1], [26]. The *depth-only* methods tend to use the information of the surrounding area of holes in depth images [2], [34]. These methods are mostly not designed to handle large loss areas where blur effects could be heavily introduced. Recently, Zhang *et al.* first predicted dense normals and occlusion boundaries from the color image and then performed the completion as an optimization problem [37]. However, it requires different kinds of ground truth labels, and the optimization step requires non-real-time computations. Cheng *et al.* proposed a more efficient convolutional spatial propagation network to reconstruct dense depth from sparse samples [5], [23].

Depth denoising. As depth sensors working on different principles, the noise model varies. For ToF cameras like Kinect, Herrera *et al.* [16] proposed an accurate calibration technique by modeling noise into a scale and a distortion portion. Shen *et al.* utilized a probabilistic model to catch uncertainties of structured-light cameras [29]. Additionally, by deriving the method proposed by Shen *et al.*, small incompleted depth regions, especially those caused by occlusions, could be completed. Jeon *et al.* rendered raw-clean depth pairs by reconstructing the ScanNet dataset [8] and then use these pairs as the supervision to learn the potential noise model [17]. In [35], Yan *et al.* proposed a dense surface reconstruction method to provide pairwise data and designed a multi-scale network to reduce noises from coarse inputs. The main drawback of these data-driven denoising approaches is that generating enough pairs for just one specific sensor type is time-consuming and expensive. Sterzentsenko *et al.* proposed a self-supervised depth denoising method in a multi-view setting with the photometric loss [31]. With inputs from four different cameras, corresponding denoised depth maps are predicted by an autoencoder network. Our method also utilizes the photometric loss as the supervision while adapting to the denoising of a single RGB-D camera, which is more widely used for indoor robots.

III. Depth Completion

In this paper, we mainly focus on the depth completion of the depth channel from RGB-D visual sensors. Depth denoising is then implemented to our depth completion results. During the depth completion step, there are mainly three different kinds of missings in depth caused by the underlying depth acquiring technique, occlusions and the range limit. For the first type, considering an active stereo-based depth sensor working in an indoor environment, the depth map could be disrupted in the fluorescent area, like a ceiling light source or reflective wooden floor. ToF sensors are more likely to fail when dealing with Multiple Path Interference (MPI), mirror-like and low-reflection regions. The second type of noise, i.e., occlusions, arises when aligning depth with one color image due to the position difference between sensors. The third type is the most challenging part, which is large blank areas caused by range limits. Current commodity RGB-D sensors typically have a most perception range between 2 to 10 meters, leaving absolutely no reliable depth data for distant objects. To complete these various missing depth values, the relation between color and depth information in one image, and the prior knowledge learned from training data should be both employed to generate satisfactory results.

We solve depth completion through an end-to-end trainable generative deep network. Given the sensor depth d^*, we first create the degraded depth representation d^-. The degraded depth d^- and its corresponding color image I are fed to the proposed network. The loss function is calculated between predicted depth $\hat{d}$ and d^*. The rationality and benefits are introduced in the following section.

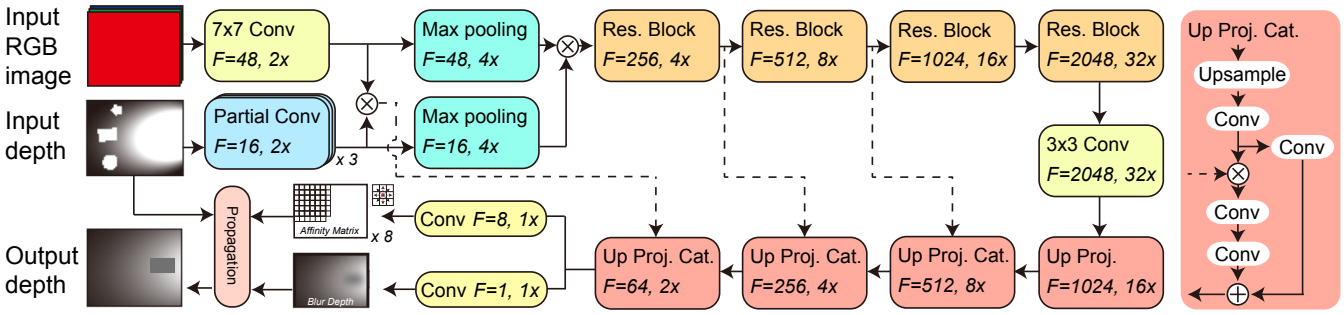


Fig. 2: The network of proposed depth completion method. With an input RGB image and an incomplete depth image, the network predicts a blurred depth and affinity matrices by two branches in the decoder. A spatial propagation step is then adopted, combined with the input depth, to give the final complete result.

A. Depth Completion with Learned Affinities

The local context plays an important role in the depth completion task, which makes learning affinity matrices with spatial propagation [5], [23] advantageous. As demonstrated in Fig. 2, the output contains two branches that separately produce affinity matrices and the blur depth d^{blur}. The affinity matrix describes the neighboring similarity information of eight directions in the depth image. First, we combine affinity matrices into each transformation kernel $\kappa_{i,j} \in \mathbb{R}^{k\times k}$ for the pixel p with the coordinate (i, j), and k is the kernel size. Let d_t^{blur} stand for the blurred depth at the t^{th} updating propagation step. The evolution equation during each spatial propagation step is then formulated as:

$$d_{i,j,t+1}^{blur} = \sum_{a,b=\frac{-(k-1)}{2}}^{\frac{k-1}{2}} \kappa_{i,j} \cdot d_{i-a,j-b,t}^{blur}. \tag{1}$$

To preserve the input depth, The binary mask M is constructed for each pixel p in the input d by letting $M(\mathrm{p}) = 1$ with $d(\mathrm{p}) > 0$ and $M(\mathrm{p}) = 0$ otherwise. Note the input depth d could be d^- or d^*. The updated depth d_{t+1}^{blur} can be then written as:

$$d_{t+1}^{blur} = (1 - M) \odot d_{t+1}^{blur} + M \odot d, \tag{2}$$

where $\odot$ is the element-wise product. This final propagation step of depth completion is implemented parrallel for all pixels which could produce completed depth efficiently.

B. Network Architecture and Loss Function

We present our generative depth completion network and then introduce our training data in the next subsection. The network architecture is illustrated in Fig. 2.

Since the input depth contains invalid values and sparse depth regions, deriving classic convolution layers could bring negative effects on learning meaningful depth features. We implement low-level feature learning to RGB and depth inputs independently and adopt the recently presented partial convolutions [21] to the depth channel. The valid depth location is first calculated from the input depth then updated after each partial convolution.

Learning to predict depth values is strongly dependent on the low-level spatial details of the input image. During the forward propagation with downsampling operations, useful spatial information is weakened or lost. We adopt similar skip connections as in the U-net [27] by directly concatenating features from the encoder to up-projection layers [5], [18].

The proposed network is trained with two terms:

$$\mathcal{L}_{completion} = \mathcal{L}_{depth} + \lambda_1 \mathcal{L}_{smoothness}, \tag{3}$$

where $\lambda_1 \in (0, 1)$ is one hyperparameter.

Depth Supervision. Instead of training only on missing areas, we find that training on all valid depth pixels between the prediction $\hat{d}$ and the ground truth, i.e., the sensor depth d^*, yields better performance. The depth loss term is defined as:

$$\mathcal{L}_{depth} = ||\mathbb{1}_{d^*>0.01m} \cdot (\hat{d} - d^*)||_1, \tag{4}$$

where $||.||_1$ indicates the $\mathcal{L}_1$ loss function, and the threshold $0.01m$ for valid depth in the ground truth is chosen for robustness.

Smoothness Loss. The depth loss measures the sum of individual errors without constraints on neighboring values. During the prediction of depth values, minimizing only the depth loss neglects that local depth areas usually share a common surface, which follows the piecewise planar assumption. The smoothness loss is then adopted to encourage this property, which is penalized on the second-order derivatives of depth predictions. The smoothness loss is formulated as:

$$\mathcal{L}_{smoothness} = ||\nabla^2 \hat{d}||_1. \tag{5}$$

In summary, the loss function of the entire model promotes both feature extraction and two branches in the decoder for a better depth completion result.

C. Training Data Generating

Previous deep RGB-D depth completion methods require rendered complete depth, which includes elaborate data collection, 3D scene reconstruction with camera poses and complete depth map projection. Despite noises introduced during reconstruction, the rendered scene still cannot achieve 100% completeness. One of our depth completion insights is that the unobserved depth contains two types of missings. Considering these two conditions, we first randomly select $\beta\%$ of valid depths in the sensor depth d^* and set them to 0. Image erosion is then adopted with s iterative steps to

produce the further incomplete depth image d^-. By doing this, the original holes in depth are enlarged, and the depth d^- also contains random novel blank areas (examples are presented in the left part of Fig. 1). Such training sets enable the network to learn the correlation between color and depth, especially around the originally unobserved area, which is essential to infer initial blanks. More importantly, we can train our model by using a bunch of existing RGB-D datasets and just simply self-collected frames.

IV. Depth Denoising

The depth denoising is implemented with another deep network. As depth sensors working on different approaches, the systematic noise generated by them is entirely camera-specific. The depth noise for sensors like stereo cameras lies in the disparity space, while sensors that can directly measure depth are different. We regard depth denoising as a further step of depth completion, which is briefly introduced here.

Three assumptions are made to achieve self-supervised depth denoising. The first is that the noise from the RGB camera is much less prominent than noises from the depth sensor. The second assumption is the scene is mostly composed of Lambertian surfaces. The third is that the scene is static. The first two assumptions hold in most cases, and moving objects can be pre-excluded, which enables us to derive the photometric loss to learn the noise model guided by the color information. An important truth here is that we can never achieve completely noise-free depth images since measuring the absolute depth in the real world is impossible. In other words, we intend to do depth denoising by shrinking the gap between depth completion results and the collected or rendered so-called ground truth.

The denoising network is also based on an encoder-decoder architecture with residual blocks of ResNet-18 [15] to extract features. The decoder part is composed of transposed layers to recover the final output. Skip connections are adopted to pass information from each encoding layer to its corresponding decoding layers. The output of the proposed denoising network is added to the original input as the final depth denoising result. During training, we adopt both the photometric loss for consistency, which is defined in detail in [25], [31] and the Huber penalty [19] for regularization.

V. Experimental Results

In this section, we demonstrate the effectiveness of the proposed method quantitatively and qualitatively. We mainly focus on the proposed depth completion part, whose task is more challenging and also show the improvement after our depth denoising part. Despite comparison on public datasets, we also collect our own data to show that our method does not rely on the labeling process. Ablation studies on different input densities and also the improvement of adopting the affinity matrix branch are explored. In all experiments, the proposed method is not trained with any manually labeled or rendered ground truths.

A. Implementation Details

In the experiments, the weights of ResNet in the proposed network are pre-trained on the ImageNet dataset [9]. The depth completion model is trained with the SGD optimizer, while the Adam optimizer is applied for our depth denoising model. In both training and evaluation stages, the image is resized to a lower resolution of $320 * 256$.

B. Datasets and Metrics

The experiments are evaluated on two public datasets, including the ScanNet [7] and the Matterport [8], and a self-collected dataset of real-world indoor scenes.

ScanNet and Matterport3D datasets. The ScanNet v1 dataset [7] contains 1513 indoor scans with 2.5 million views, captured by a Structure sensor sharing a similar design with Microsoft Kinect v1. The Matterport3D dataset [8] adopts the same device to capture RGB-D frames, which contains 90 scenes with 194K frames. Since the training of the proposed method does not need additional rendered ground truth, we randomly sample 215K images and 155K images from the ScanNet and Matterport datasets for training. Two small parts with each 2K images from these two datasets are used to evaluate the final depth completion and denoising performances. The completion network is trained on both training sets, while the depth denoising network is trained only on the ScanNet v1 dataset [7], considering two aspects. (1) The data capturing device is different from dataset to dataset, which could influence the noise model. (2) The ScanNet dataset is suitable for our self-supervised approach based on the photometric loss, which requires consecutive images sharing dominant scene overlaps. Therefore, we randomly choose 100 scans with 170K images to train our depth denoising model.

Self-collected datasets. We collect our small indoor datasets with Microsoft Azure Kinect, a ToF RGB-D sensor. During the data collecting process, we set Azure Kinect to the mode with operating range of 0.25 - 2.88 m.

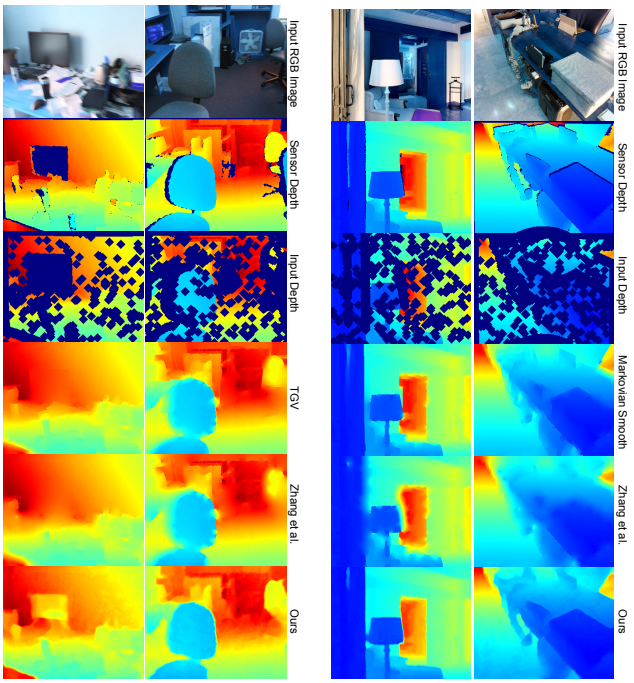
Metrics. For both depth completion and denoising, we adopt the same metrics as in [37]. Given ground truth depth $D^* = \{d^*\}$ and results $\hat{D} = \{\hat{d}\}$, the metrics include: (1) RMSE: $\sqrt{\frac{1}{|\hat{D}|}\sum_{\hat{d}\in\hat{D}}||\hat{d}-d^*||}$. (2) Abs Rel: $\frac{1}{|\hat{D}|}\sum_{\hat{d}\in\hat{D}}|\hat{d}-d^*|/d^*$. (3) δ_t: the percentage of $\hat{d}\in\hat{D}$, s.t. $\max(\frac{d^*}{\hat{d}},\frac{\hat{d}}{d^*}) < t$ where $t \in \{1.05, 1.10, 1.25, 1.25^2, 1.25^3\}$.

C. Comparison

Tab. I shows the result of comparing our method with both classic and deep learning methods. Several well-known non-data-driven methods are demonstrated, which includes the interpolation with the average of nearest values in four directions (Basic Smooth), the guided anisotropic diffusion [22] (Anisotropic Diffusion), the guided edge-aware energy optimization [10] (TGV) and the smooth with second-order Markov Random Field [14] (Markovian Smooth). We also compare our method with the most relevant deep depth completion method [37] proposed by Zhang *et al.*. They

TABLE I: Comparison to both baseline classic and deep depth completion methods.

Method	Lower the better		Higher the better						Time (ms)
	RMSE (mm)	**Abs Rel** (mm)	$\delta_{1.02}$	$\delta_{1.05}$	$\delta_{1.10}$	$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$	
Classic Depth Completion Methods									
Basic Smooth	667.3	160.4	68.46	81.49	88.03	92.24	93.98	95.59	-
Anisotropic Diffusion [22]	659.4	159.9	69.88	82.50	88.58	93.34	95.60	96.33	1075 (CPU)
TGV [10]	796.15	180.15	33.74	43.79	50.82	63.84	79.29	88.41	1560 (CPU)
Markovian Smooth [14]	212.4	44.1	**70.70**	83.30	89.92	95.41	97.74	98.47	620 (CPU)
Deep Depth Completion Methods									
Zhang *et al.* [37]	219.3	48.1	66.22	79.37	87.34	94.75	97.82	98.65	1056 (GPU&CPU)
Ours	**186.7**	**43.0**	60.12	**85.34**	**92.97**	**97.22**	**98.23**	**99.51**	19 (GPU)
Deep Denoising on the ScanNet [7] dataset									
Ours (ScanNet Only)	**143.9**	33.1	62.64	88.05	93.55	97.45	99.21	99.73	19 (GPU)
Ours+Denoising (ScanNet Only)	147.8	**32.8**	**63.88**	**88.90**	**93.97**	**98.34**	**99.50**	**99.82**	19+16 (GPU)



(a) ScanNet dataset [7] (b) Matterport dataset [8]

Fig. 3: Comparison to classic and deep learning depth completion methods. These images are chosen for containing small objects and large depth variation. Our method preserves detail structures and predicts accurate depth for large absence.

trained his method on both the normal and occlusion prediction on the SUNCG dataset [30], the ScanNet dataset [7] and the Matterport dataset [8] with complete ground truths. The final output is generated by optimizing a combination of the sparse input, the normal image and the occlusion boundary image. Therefore, we directly derive their published trained model on predicting normals and occlusions and using our sparse depth for optimization. The result is conducted on the sampled ScanNet dataset [7] and Matterport dataset [8]. The input depth is degraded from the sensor depth with $\beta = 2.4\%$ and $s = 10$ resulting in a loss of 63% valid depth values.

Our results in Tab. I outperform other methods, including the data-driven method [37]. The RMSE of the proposed method on two datasets is 186.7 mm, while other methods lie in the range of 212.4 - 667.3 mm. Furthermore, our method is advantageous on most of the metrics measuring predicted pixels fall in a given range. The accuracy after depth denoising is presented at the bottom two rows of Tab. I. During depth denoising, the predicted noise is reducted to the input completed depth to give the final results. The depth denoising part is evaluated on the ScanNet dataset [7] with only predicted depth values. It is shown that the proposed depth denoising method improves depth results to some extent. Additionally, we show the average time required for each method with a resolution of $320 * 256$. The running time and devices are also identified in Tab. I for reference. Note methods that achieve results with a gradual completion process might not be suitable for parallel implementations on the GPU. The proposed method could achieve real-time performances.

Qualitative comparisons between different methods [10], [37] are demonstrated in Fig. 3. The sensor depth is regarded as the ground truth while only input RGB image and input degraded depth are utilized to generate complete depth. As shown in the first column of Fig. 3 (a), the depth of the display is not observed in the original sensor depth while predicted in our results, which reflect that the prior knowledge could be learned in a data-driven manner without providing rendered depth. The variation of scenes from the Matterport dataset [8] as in Fig. 3 (b) is much larger than the ScanNet dataset [7], which could make the depth completion more difficult. The second column of Fig. 3 (b) contains several relatively small objects whose boundaries are mostly degraded. The proposed method predicts their shape precisely with the affinity information of local color and depth context. A scene of drastic depth changes is demonstrated in the second column of Fig. 3 (b). Compared to the other two methods, the proposed method provides relatively more reasonable outputs.

D. Results on Self-collected Dataset

This part demonstrates the result of the self-collected dataset. Since the proposed method can be trained without rendered ground truth depth, we fine-tune our model with degraded depth images and sensor outputs directly. The result is demonstrated in Fig. 6. The lost area in the second

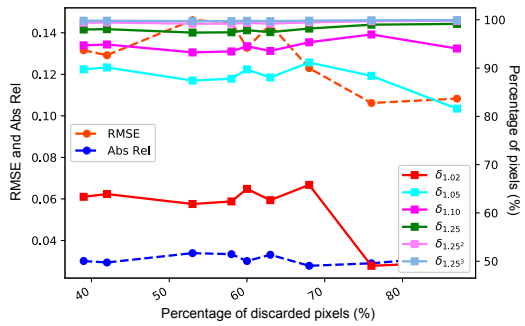


Fig. 4: Effects of depth density on the accuracy of depth completion, which is measured on the ScanNet dataset [7].

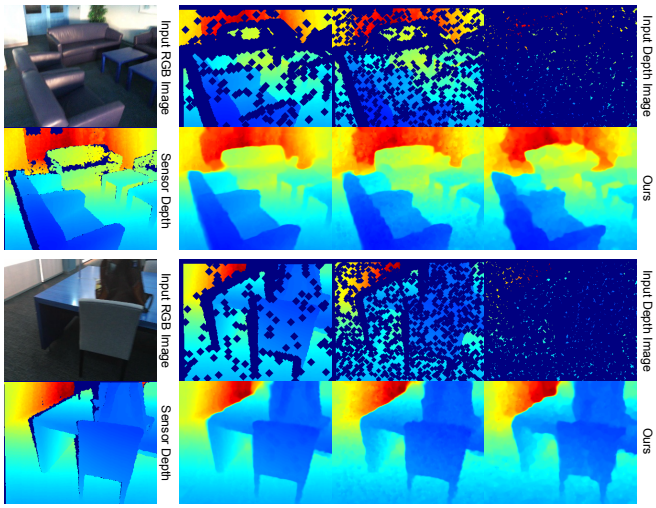


Fig. 5: The result of the proposed method with different depth densities. The input depth density gradually drops, which demonstrates the ability of our method dealing with dense to sparse depths.

column containing large depth variations can be effectively completed by our proposed method.

E. Ablation Study

We conduct two sets of experiments to investigate how different inputs, the affinity branch, influence the results.

Results with different input densities. This part is measured on generated degraded depths from the ScanNet dataset [7]. We justify the percent of random samples $\beta\%$ and also erosion steps s to get different inputs. These incomplete depths images are fed to the same version of our network for depth completion. The original sensor depth is regarded as the ground truth for evaluation.

Fig. 4 shows the result with different quality of input depth, i.e., the density. The horizontal axis denotes the percentage of discarded pixels. The vertical axis contains three metrics, including RMSE, Abs Rel and percentages of pixels. From this quantitative result, we see that the input density does not affect the proposed depth completion method obviously. The accuracy does not directly drop with sparser inputs because of two aspects. The first reason is that the proposed method only requires a small number of valid

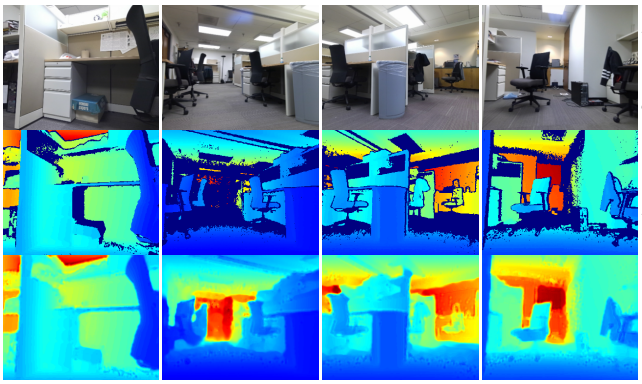


Fig. 6: The result of the proposed method on self-collected with the Azure Kinect.

TABLE II: Comparison results on the ScanNet dataset [7] between different variants of proposed method.

Method	Lower the better		Higher the better			
	RMSE (mm)	Abs Rel (mm)	$\delta_{1.02}$	$\delta_{1.05}$	$\delta_{1.10}$	$\delta_{1.25}$
Ours *omit affinity*	167.7	39.8	20.30	84.36	91.90	96.23
Ours	**143.9**	**33.1**	**62.64**	**88.05**	**93.55**	**97.45**

pixels to predict their surrounding depth values. The second is the metric we adopt is related to the number of evaluated pixels. Further qualitative results are demonstrated in Fig. 5. As the input depth deteriorated, most object boundaries still maintain clear, and not many blur effects are introduced.

Results without affinity learning. In this part, we aim to show the significance of learning the affinity matrix. We prune the branch of generating affinity matrices and then train the simplified network with the same objective function and training settings. The results in Tab. II demonstrate that, without affinity learning, the percentage dramatically drops, especially for more delicate metrics. The reason is when omitting affinity learning, minor depth fluctuations could happen severely, which are reflected on the $\delta_{1.02}$ metric.

In the experiments, different kinds of missing could be effectively addressed with the proposed depth completion network. The prior knowledge of depth could be effectively learned during the training stage. Another branch of affinity matrices assists in propagating depth for more precise results.

VI. Conclusion

In this paper, we describe a novel depth completion and denoising method of RGB-D sensors that can be performed in a fully unsupervised manner. The proposed depth completion method first predicts blurred depth with the prior knowledge and then propagates iteratively with learned affinity matrices. Various comparisons are conducted in our experimental results, both qualitatively and quantitatively. Further depth denoising is employed with a separate network under the self-supervision with the self-supervised photometric loss, which could address different noise models. The proposed method also achieves real-time performance and is suitable to be adopted on indoor robotic platforms.

REFERENCES

[1] Jonathan T Barron and Ben Poole. The fast bilateral solver. In *European Conference on Computer Vision*, pages 617–632. Springer, 2016.

[2] Marcelo Bertalmio, Andrea L Bertozzi, and Guillermo Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[3] Ayan Chakrabarti, Jingyu Shao, and Greg Shakhnarovich. Depth from a single image by harmonizing overcomplete local network predictions. In *Advances in Neural Information Processing Systems*, pages 2658–2666, 2016.

[4] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017.

[5] Xinjing Cheng, 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)*, pages 103–119, 2018.

[6] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*, 2013.

[7] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[8] Angela Dai, Matthias Nießner, Michael Zollöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface re-integration. *ACM Transactions on Graphics 2017 (TOG)*, 2017.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] David Ferstl, Christian Reinbacher, Rene Ranftl, Matthias Rüther, and Horst Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 993–1000, 2013.

[11] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018.

[12] Clément Godard, Oisin 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*, pages 270–279, 2017.

[13] Xiaojin Gong, Junyi Liu, Wenhui Zhou, and Jilin Liu. Guided depth enhancement via a fast marching method. *Image and Vision Computing*, 31(10):695–703, 2013.

[14] Alastair Harrison and Paul Newman. Image and sparse laser fusion for dense scene reconstruction. In *Field and Service Robotics*, pages 219–228. Springer, 2010.

[15] Kaiming He, 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*, pages 770–778, 2016.

[16] Daniel Herrera, Juho Kannala, and Janne Heikkilä. Joint depth and color camera calibration with distortion correction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):2058–2064, 2012.

[17] Junho Jeon and Seungyong Lee. Reconstruction-based pairwise depth dataset for depth image enhancement using cnn. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 422–438, 2018.

[18] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.

[19] Sophie Lambert-Lacroix and Laurent Zwald. The adaptive berhu penalty in robust regression. *Journal of Nonparametric Statistics*, 28(3):487–514, 2016.

[20] Chao Liu, Jinwei Gu, Kihwan Kim, Srinivasa G Narasimhan, and Jan Kautz. Neural rgb (r) d sensing: Depth and uncertainty from a video camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10986–10995, 2019.

[21] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[22] Junyi Liu and Xiaojin Gong. Guided depth enhancement via anisotropic diffusion. In *Pacific-Rim conference on multimedia*, pages 408–417. Springer, 2013.

[23] Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, and Jan Kautz. Learning affinity via spatial propagation networks. In *Advances in Neural Information Processing Systems*, pages 1520–1530, 2017.

[24] Si Lu, Xiaofeng Ren, and Feng Liu. Depth enhancement via low-rank matrix completion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3390–3397, 2014.

[25] Fangchang Ma, Guilherme Venturelli 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)*, pages 3288–3295. IEEE, 2019.

[26] Georg Petschnigg, Richard Szeliski, Maneesh Agrawala, Michael Cohen, Hugues Hoppe, and Kentaro Toyama. Digital photography with flash and no-flash image pairs. *ACM transactions on graphics (TOG)*, 23(3):664–672, 2004.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[28] Johannes L Schönberger, Marc Pollefeys, Andreas Geiger, and Torsten Sattler. Semantic visual localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6896–6906, 2018.

[29] Ju Shen and Sen-Ching S Cheung. Layer depth denoising and completion for structured-light rgb-d cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1187–1194, 2013.

[30] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1754, 2017.

[31] Vladimiros Sterzentsenko, Leonidas Saroglou, Anargyros Chatzitofis, Spyridon Thermos, Nikolaos Zioulis, Alexandros Doumanoglou, Dimitrios Zarpalas, and Petros Daras. Self-supervised deep depth denoising. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1242–1251, 2019.

[32] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 International Conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017.

[33] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1290–1297. IEEE, 2012.

[34] Hongyang Xue, Shengming Zhang, and Deng Cai. Depth image inpainting: Improving low rank matrix completion with low gradient regularization. *IEEE Transactions on Image Processing*, 26(9):4311–4320, 2017.

[35] Shi Yan, Chenglei Wu, Lizhen Wang, Feng Xu, Liang An, Kaiwen Guo, and Yebin Liu. Ddrnet: Depth map denoising and refinement for consumer depth cameras using cascaded cnns. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 151–167, 2018.

[36] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 21(8):690–706, 1999.

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

[38] Tinghui Zhou, 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*, pages 1851–1858, 2017.