17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

# Seeing Farther Than Supervision: Self-supervised Depth Completion in Challenging Environments

Seiya Ito, Naoshi Kaneko, and Kazuhiko Sumi
Aoyama Gakuin University
5-10-1 Fuchinobe, Chuo-ku, Sagamihara-shi, Kanagawa, Japan
ito.seiya@vss.it.aoyama.ac.jp, {kaneko, sumi}@it.aoyama.ac.jp

## Abstract

*This paper tackles the problem of learning a depth completion network from a series of RGB images and short-range depth measurements as a new setting for depth completion. Commodity RGB-D sensors used in indoor environments can provide dense depth measurements; however, their acquisition distance is limited. Recent depth completion methods train CNNs to estimate dense depth maps in a supervised/self-supervised manner while utilizing sparse depth measurements. For self-supervised learning, indoor environments are challenging due to many non-textured regions, leading to the problem of inconsistency. To overcome this problem, we propose a self-supervised depth completion method that utilizes optical flow from two RGB-D images. Because optical flow provides accurate and robust correspondences, the ego-motion can be estimated stably, which can reduce the difficulty of depth completion learning in indoor environments. Experimental results show that the proposed method outperforms the previous self-supervised method in the new depth completion setting and produces qualitatively adequate estimates.*

## 1 Introduction

Depth completion is the problem of estimating the dense depth map from RGB images and sparse depth measurements provided from LiDAR or SLAM. A variety of depth completion methods using Convolutional Neural Networks (CNNs) have been proposed [1, 2, 3, 4, 5, 6]. The typical design of CNNs is an encoder-decoder network that propagates sparse depth measurements to surrounding pixels. These methods train CNNs in a supervised manner and show promising results. However, sensors such as LiDAR generally do not provide ground truth dense depth. To overcome this issue, the self-supervised depth completion method has been proposed [7]. Nevertheless, self-supervised learning is challenging in indoor environments that contain many non-textured regions.

In recent years, unsupervised monocular depth estimation has been extensively studied. Most studies formulated this problem as the joint learning of depth and camera pose from monocular videos. Zhou et al. presented the photometric loss between the reference image and the image synthesized by warping the source image using estimated depth and camera pose to train depth and ego-motion networks [8]. Following [8], various unsupervised methods have been proposed [9, 10, 11]. Some studies introduced optical flow for estimating ego-motion [12, 13] to handle indoor scenes.

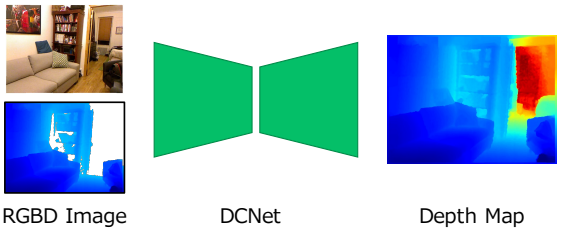



Figure 1: Concept of this work. Given an RGB image and a short-range depth image, the proposed depth completion network (DCNet) estimates the depth including the depth outside the measurement range of the sensor (filled in white in the input depth image). DCNet is trained in a self-supervised manner.

In indoor environments, relatively dense depths can be measured with sensors such as Microsoft Kinect. However, the measurable distance of such a sensor may be short, e.g., up to 10 m. Extending the measurable distance extends the possibilities of applications using the same sensor. In this paper, we propose a method to estimate a complete depth map, including *farther distances*, from an RGB image and a short-range depth map (Fig. 1). In the conventional depth completion setting, the input depth is sparse, but it contains the full range of distances to be estimated, which can be used as supervisory signals. In contrast, the problem addressed in this study has limited access to the signals, i.e., the distances to be estimated are more extensive than the supervisory signals. The proposed method combines the best aspects of depth completion and unsupervised monocular depth estimation. We extend an unsupervised depth learning framework for indoor environments to utilize depth measurements, which enables us to predict absolute depth.

The contributions of this work are as follows:

- We pose the problem of estimating long distances from RGB images and short-range depth measurements as a new setting for depth completion.

- We propose a self-supervised indoor depth completion framework consisting of depth and optical flow networks.
- We analyze relationships between the input depth measurements used during training and inference, and show effective combinations to achieve high performance.

## 2 Proposed Method

### 2.1 Overall Pipeline

Fig. 2 shows an overall pipeline of the proposed method consisting of two networks: depth completion network (DCNet) and optical flow estimation network (FlowNet). DCNet takes RGB and short-range depth images as input and estimates a dense depth map. FlowNet takes two RGB-D images as input and predicts the optical flow from the two images. These networks are trained in a self-supervised manner using input RGB-D images and estimated outputs.

During training, multiple calibrated RGB-D images are used as input. The relative camera pose of two RGB-D images is estimated via correspondences obtained by the estimated optical flow. For each of the RGB-D images, DCNet estimates the depth map. Once the depth map and camera pose are obtained, the RGB images of the different views can be warped by computing the viewpoints' transformation. The photometric error between the original and warped images is used as self-supervisory signals for DCNet. The warped images can also be generated from optical flow, and they are used to train FlowNet. We will provide a more detailed explanation in the subsequent sections.

### 2.2 Ego-Motion Estimation via Optical Flow

Ego-motion estimation is an essential component of self-supervised learning. To cope with indoor scenes, some studies introduce an optical flow estimation network and estimate ego-motion based on the optical flow [12, 13]. Since the optical flow network is trained in an unsupervised manner, the overall pipeline can be regarded as unsupervised learning. Zhou et al. [13] presented a CNN that uses optical flow images as input and ego-motion as output to improve the depth estimation performance in indoor environments. Zhao et al. [12] showed that it is more effective to estimate the ego-motion by computing the fundamental matrix from the correspondences obtained from the optical flow rather than inputting the optical flow images into the CNN. In this work, we adopt the ego-motion estimation method using correspondences.

The core idea is to seek reliable correspondences from optical flow. We estimate forward and backward optical flow from input images by FlowNet and compute occlusion mask [14] and flow consistency scores [10]. Subsequently, we randomly sample $k$ correspondences with the top 20% consistency scores in the non-occluded region. Once the correspondences are found, the fundamental matrix is computed using the 8-point algorithm [15] and RANSAC [16]. The appropriate camera pose is estimated by geometric verification.

The loss function for FlowNet is defined as follows:

$$\mathcal{L}_{flow} = w_1^f \mathcal{L}_{fp} + w_2^f \mathcal{L}_{fs} + w_3^f \mathcal{L}_{fc} \tag{1}$$

where $\mathcal{L}_{fp}$ is the photometric loss [12], $\mathcal{L}_{fs}$ is the flow smoothness loss [10], and $\mathcal{L}_{fc}$ is the forward-backward flow consistency loss [10]. We use $(w_1^f, w_2^f, w_3^f) = (1.0, 10.0, 0.01)$ during training.

### 2.3 Self-supervised Depth Completion Learning

We build the proposed self-supervised depth completion network upon the previous unsupervised depth learning framework [12]. The main difference is the use of two different depth-related supervisory signals.

First, we use the short-range depth as a supervisory signal. Specifically, we compute the difference between the input depth and the output depth from DCNet for pixels where the depth measurement exists. Let $i$ be the pixel and $D$ and $\hat{D}$ be the estimated depth map and input depth map, respectively. This loss is defined as follows:

$$\mathcal{L}_{dd} = \frac{1}{|\mathcal{M}_d|} \sum_{i \in \mathcal{M}_d} |D(i) - \hat{D}(i)| \tag{2}$$

where $\mathcal{M}_d$ is a set of pixels for which a depth measurement exists. This loss encourages DCNet to learn the absolute scale of the depth.

Second, we use the 3D points recovered using the correspondences obtained from the optical flow as supervisory signals. Similar to the camera pose estimation, we use the $k$ pixel correspondences and apply two-view triangulation [17] to obtain the depths of these pixels. Since the camera poses are relative in scale, the depth obtained by triangulation is also relative. To align the depth scale between the depth from triangulation $D_t$ and depth estimated by DCNet $D$, we introduce a single scale factor $s$ and consider the depth to be $\hat{D}_t = sD_t$. Using $s$ that minimizes the error, we calculate the loss defined as follows:

$$\mathcal{L}_{dt} = \frac{1}{|\mathcal{M}_t|} \sum_{i \in \mathcal{M}_t} |D(i) - \hat{D}_t(i)| \tag{3}$$

where $\mathcal{M}_t$ is a set of pixels for which a triangulated depth exists.

The loss function for DCNet is defined as follows:

$$\mathcal{L}_{depth} = w_1^d \mathcal{L}_{dp} + w_2^d \mathcal{L}_{ds} + w_3^d \mathcal{L}_{dt} + w_4^d \mathcal{L}_{dd} + w_5^d \mathcal{L}_{dr} \tag{4}$$

where $\mathcal{L}_{dp}$ is the photometric loss [12], $\mathcal{L}_{ds}$ is the depth smoothness loss [10], $\mathcal{L}_{dr}$ is the dense reprojection loss [12]. We use $(w_1^d, w_2^d, w_3^d, w_4^d, w_5^d) = (1.0, 10^{-4}, 1.0, 1.0, 0.1)$ during training.

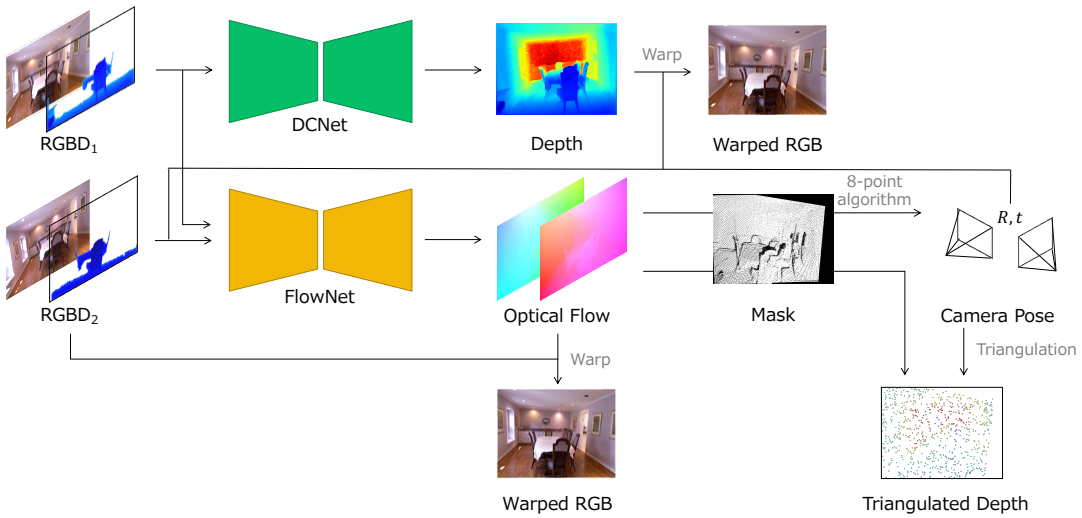



Figure 2: Overall pipeline of the proposed method. DCNet takes an RGB-D image and estimates a dense depth map. FlowNet takes two RGB-D images and predicts their optical flow. The dense depth map and optical flow are used to synthesize RGB images, which is used as supervisory signals. The optical flow is also used to estimate the ego-motion, and triangulation is applied to obtain a depth map that is used as an auxiliary supervisory signal.

Table 1: Quantitative results on the NYU Depth v2 dataset. Depth measurements of up to 3 meters are used for training and inference. The method of [12] was trained only on RGB sequences and the scale is relative.

| Method | Input | Error ↓ | | | Accuracy ↑ | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Rel | RMSE | log10 | $\delta < 1.05$ | $\delta < 1.15$ | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ |
| Zhao et al. [12] | RGB | 0.192 | 0.665 | 0.080 | 0.209 | 0.519 | 0.705 | 0.916 | 0.975 |
| Ma et al. [7] | RGB-D | 0.076 | 0.558 | 0.038 | 0.699 | 0.847 | 0.883 | 0.951 | 0.983 |
| Ours | RGB-D | **0.057** | **0.521** | **0.029** | **0.757** | **0.850** | **0.892** | **0.956** | **0.986** |

## 3 Experiments

### 3.1 Implementation Details

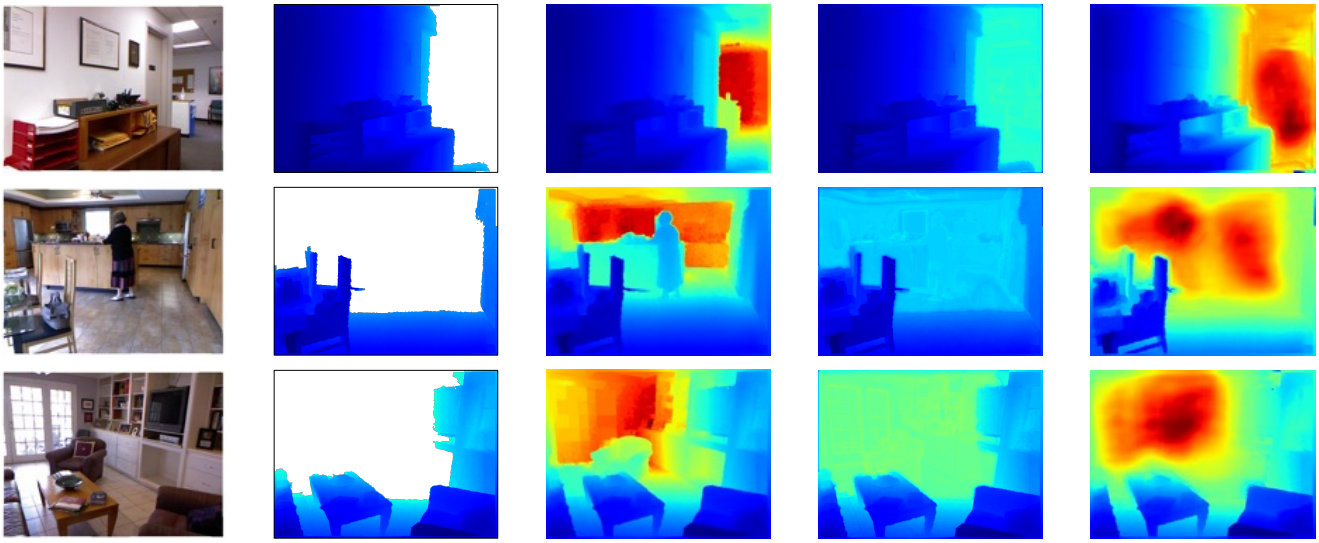
**Dataset:** We evaluate the proposed method on the NYU Depth v2 dataset [18], which contains both RGB and depth images of 464 indoor scenes taken with the Microsoft Kinect sensor. We use the official split, i.e., 249 scenes for training and 215 scenes for testing. For training, each raw video sequence in the training set is sampled spatially uniformly to produce approximately 48k synchronized RGB-D image pairs. Since large non-textured regions complicate the consistency checking between frames, the original RGB and depth images of size $640 \times 480$ are resized to $256 \times 192$. We used 654 labeled images of the original size to evaluate the proposed depth completion network. The maximum depth of the test set is 10 m.

**Training:** Our training procedure consists of two stages. We first train the optical flow network for 20 epochs. Afterward, the parameters of the optical flow network are fixed, and the depth completion network is trained in another 20 epochs. Although the proposed method can also be trained by end-to-end learning, it requires more iterations than two-stage training. Similar to prior work [12], we can add a stage where the optical flow and depth completion networks are trained jointly. However, it hardly improved the accuracy.

In each stage, we augment the input data with flips with 0.5 probability and color jitter. We also add noise of $[-0.2, 0.2]$ m to the depth and randomly remove the depth value with 0.1 probability. We train the networks using the Adam optimizer [19] with mini-batches of size 16. An initial learning rate is set to $10^{-4}$ and reduced to 10% for every 5 epochs.

### 3.2 Comparison with Existing Methods

We compare our method with the existing methods including unsupervised depth estimation [12] and self-supervised depth completion [7]. Both the proposed method and the method of [7] are trained in the same short-range setting, where the short-range depth is defined as the depth up to 3 m. We report evaluation metrics commonly used for depth estimation [10, 12] and depth completion [5, 7]: mean absolute relative dif-



(a) RGB (b) Short-range depth (c) Ground truth (d) Ma et al. [7] (e) Ours

Figure 3: Qualitative comparison between recent self-supervised depth completion and our method. (a) Input RGB image. (b) Input short-range depth. (c) Ground truth. (d) Predictions by the method of [7]. (e) Predictions by our method. Depth measurements up to 3 m are used during training and inference.

Table 2: The performance of models trained in different depth ranges.

| Depth Range | | Error | Accuracy |
|---|---|---|---|
| Training | Inference | RMSE | $\delta < 1.25$ |
| up to 3 m | up to 3 m | 0.521 | 0.892 |
| | up to 5 m | 0.616 | 0.933 |
| | up to 8 m | 0.301 | 0.989 |
| up to 5 m | up to 3 m | 0.826 | 0.778 |
| | up to 5 m | 0.186 | 0.983 |
| | up to 8 m | 0.323 | 0.974 |
| up to 8 m | up to 3 m | 1.973 | 0.695 |
| | up to 5 m | 0.337 | 0.959 |
| | up to 8 m | 0.064 | 0.999 |

ference (Rel), root mean squared error (RMSE), mean $\log 10$ (log10), and accuracy ($\delta < thr$), which is the ratio of pixels whose maximum relative error $\delta$ is below the threshold $thr$.

Table 1 shows the quantitative results on the NYU dataset. The proposed method outperforms the recent self-supervised method [7] in all metrics. The proposed method can estimate the depth much more accurately than the method which our method built upon [12]. Fig. 3 shows a qualitative comparison between our method and the method of [7]. The estimation results of [7] focus only on distances around the input depth measurements and fail to estimate the global depth scale of a scene. Meanwhile, the proposed method can estimate the depth more accurately even at a long distance.

### 3.3 Analysis of Input Depth Range

We investigate relationships of the input depth ranges used during training and inference. Since sensor specifications may vary from application to application, the generality of the model is an important aspect. Table 2 shows the performance of the proposed model trained in different depth ranges, i.e., up to 3 m, 5 m, and 8 m. In most cases, the accuracy is improved when the depth range during inference is more extensive than during training. It is worth noting that using a narrower depth range during inference than during training results in performance degradation. Comparing in terms of the depth range used during inference, the highest accuracy is achieved when the depth range during inference is the same as during training.

## 4 Conclusion

In this paper, we tackled a new depth completion problem of completing the depth of long-distance points from an RGB image and a short-range depth map. We presented a self-supervised depth completion method for indoor environments, which uses optical flow for ego-motion estimation. The experimental results demonstrated that our method outperformed the recent self-supervised depth completion method. We also investigated relationships between the input depth measurements used during training and inference. We showed that it is effective to align the depth ranges during training and inference.

## Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP20J13300.

## References

[1] F. Ma and S. Karaman: "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," in *ICRA*, pp.1–8, 2018.

[2] W. V. Gansbeke et al.: "Sparse and noisy lidar completion with RGB guidance and uncertainty," in *MVA*, pp.1–6, 2019.

[3] J. Qiu et al.: "Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image," in *CVPR*, pp.3313–3322, 2019.

[4] X. Xiong et al.: "Sparse-to-dense depth completion revisited: Sampling strategy and graph construction," in *ECCV*, pp.682–699, 2020.

[5] X. Cheng, P. Wang, and R. Yang: "Learning depth with convolutional spatial propagation network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.42, no.10, pp.2361–2379, 2020.

[6] J. Park et al.: "Non-local spatial propagation network for depth completion," in *ECCV*, pp.120–136, 2020.

[7] F. Ma et al.: "Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera," in *ICRA*, pp.3288–3295, 2019.

[8] T. Zhou et al.: "Unsupervised learning of depth and ego-motion from video," in *CVPR*, pp.6612–6619, 2017.

[9] C. Godard et al.: "Digging into self-supervised monocular depth estimation," in *ICCV*, pp.3827–3837, 2019.

[10] Z. Yin and J. Shi: "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *CVPR*, pp.1983–1992, 2018.

[11] J. Bian et al.: "Unsupervised scale-consistent depth and ego-motion learning from monocular video," in *NeurIPS*, pp.35–45, 2019.

[12] W. Zhao et al.: "Towards better generalization: Joint depth-pose learning without posenet," in *CVPR*, pp.9148–9158, 2020.

[13] J. Zhou et al.: "Moving indoor: Unsupervised video depth learning in challenging environments," in *ICCV*, pp.8617–8626, 2019.

[14] Y. Wang et al.: "Occlusion aware unsupervised learning of optical flow," in *CVPR*, pp.4884–4893, 2018.

[15] R. I. Hartley: "In defense of the eight-point algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.19, no.6, pp.580–593, 1997.

[16] M. A. Fischler and R. C. Bolles: "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol.24, no.6, pp.381–395, 1981.

[17] R. I. Hartley and P. F. Sturm: "Triangulation," *Comput. Vis. Image Underst.*, vol.68, no.2, pp.146–157, 1997.

[18] N. Silberman et al.: "Indoor segmentation and support inference from RGBD images," in *ECCV*, pp.746–760, 2012.

[19] D. P. Kingma and J. Ba: "Adam: A method for stochastic optimization," in *ICLR*, 2015.