

# Decoder Modulation for Indoor Depth Completion

Dmitry Senushkin<sup>1,2</sup>, Mikhail Romanov<sup>1</sup>, Iliia Belikov<sup>1</sup>, Nikolay Patakin<sup>1</sup> and Anton Konushin<sup>1,2</sup>

**Abstract**—Depth completion recovers a dense depth map from sensor measurements. Current methods are mostly tailored for very sparse depth measurements from LiDARs in outdoor settings, while for indoor scenes Time-of-Flight (ToF) or structured light sensors are mostly used. These sensors provide semi-dense maps, with dense measurements in some regions and almost empty in others. We propose a new model that takes into account the statistical difference between such regions. Our main contribution is a new decoder modulation branch added to the encoder-decoder architecture. The encoder extracts features from the concatenated RGB image and raw depth. Given the mask of missing values as input, the proposed modulation branch controls the decoding of a dense depth map from these features differently for different regions. This is implemented by modifying the spatial distribution of output signals inside the decoder via Spatially-Adaptive Denormalization (SPADE) blocks. Our second contribution is a novel on-the-fly sensor simulation strategy that allows us to train on a semi-dense sensor data when the ground truth depth map is not available. Our model achieves the state of the art results on indoor *Matterport3D* dataset [1]. Being designed for semi-dense input depth, our model is still competitive with LiDAR-oriented approaches on the *KITTI* dataset [2]. Our sensor simulation strategy significantly improves prediction quality with no dense ground truth available, as validated on the *NYUv2* dataset [3].

## I. INTRODUCTION

In recent years, depth sensors have become an essential component of many devices, from self-driving cars to smartphones. However, the quality of modern depth sensors is still far from perfect. LiDAR systems provide accurate but spatially sparse measurements while being quite expensive and large. Commodity-grade depth sensors based on the active stereo with structured light (e.g., Microsoft Kinect) or Time-of-Flight (e.g., Microsoft Kinect Azure or depth sensors in many smartphones) provide estimations that are relatively dense but less accurate and within a limited distance range. LiDAR-based sensors are widely used in outdoor environments, especially for self-driving cars, while the other sensors are mainly applicable in an indoor setting. Due to the rapid growth of the self-driving car industry, the majority of recent depth completion methods are mostly focused on outdoor depth completion for LiDAR data [2], [4], [5], often overlooking other types of sensors and scenarios. Nevertheless, these sensors are an essential part of many modern devices (such as mobile phones, AR glasses, and others).

LiDAR-oriented methods mainly deal with sparse measurements. Applying these methods to depth data captured

with semi-dense sensors as-is may be a suboptimal strategy. This kind of transfer requires additional heuristics such as sparse random sampling. The most popular approach [4], [6], [7], [8] of training LiDAR-oriented methods on a semi-dense depth map proceeds as follows. First, the gaps in semi-dense depth maps are filled using simple interpolation methods such as bilateral filtering or the approach of [9]. Then, some depth points are uniformly sampled from the resulting map. This heuristic approach is used due to the lack of LiDAR data for indoor environments, but such kind of preprocessing suggests that it may be better to use a model originally designed to operate with semi-dense data. Such an approach would take into account the features of semi-dense sensor data and would not require separate heuristics for transfer.

Inspired by these observations, we present a novel solution for the indoor depth completion from semi-dense depth maps guided by color images. Since sensor data may be present for 60% of pixels and more, we propose to use a single encoder for the joint RGBD signal. Taking into account the statistical differences between regions with and without depth measurements, we design a decoder modulation branch that takes a mask as input and modifies the distributions of activation maps in the decoder. This modulation mechanism is based on Spatially-Adaptive Denormalization (SPADE) blocks [10]. Since there are few publicly available datasets with both sensor and dense ground truth depth, we additionally propose a special sensor simulation strategy for depth completion models that emulates semi-dense sensors on-the-fly and does not require dense depth reconstruction.

As a result, we offer the following **contributions**:

- a novel network architecture for indoor depth completion with a decoder modulation branch;
- a novel sensor simulation strategy that emulates semi-dense sensors on-the-fly and does not require dense depth reconstruction;
- large-scale experimental validation on real datasets including *Matterport3D*, *ScanNet*, *NYUv2*, and *KITTI*.

The paper is organized as follows. In Section II, we review related work on depth estimation and dense image labeling. Section III presents our approach, including the new architecture and simulation strategies. Section IV describes the experimental setup, Section V presents the results of our experiments, and Section VII concludes the paper.

## II. RELATED WORK

In this section, we review works on several topics related to depth processing for images or works that have served as the original inspiration for our work. Namely, we cover depth

<sup>1</sup>Samsung AI Research Center, Moscow, Russia; d.senushkin@partner.samsung.com, {m.romanov, ilia.belikov, n.patakin, a.konushin}@samsung.com

<sup>2</sup>Lomonosov Moscow State University, Moscow, Russia

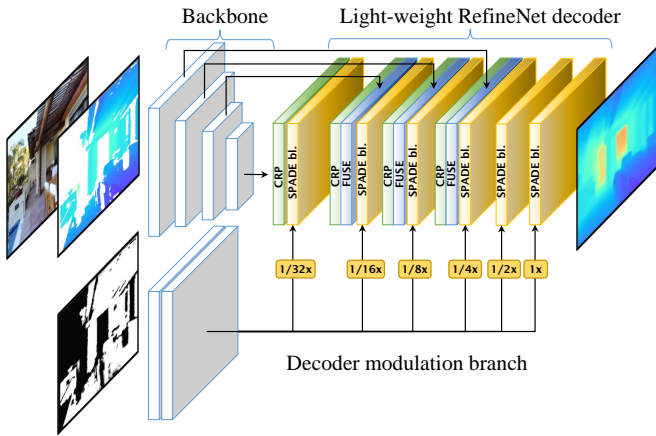


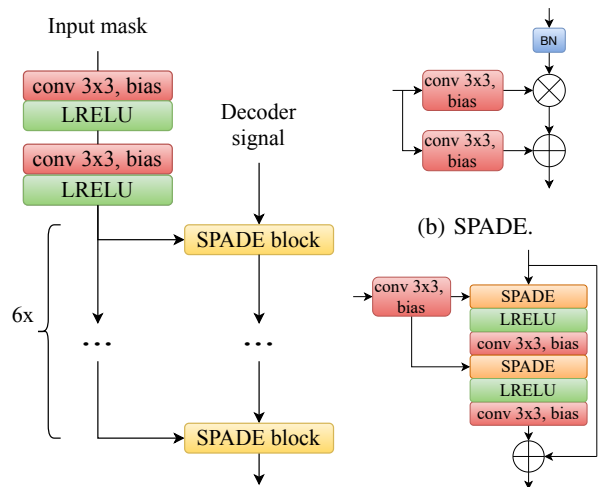
Fig. 1: High-level architecture of the proposed DM-LRN network. Pretrained EfficientNet [22] backbone encodes the input RGBD signal. Extracted features are fed into the lightweight RefineNet decoder [23]. The decoder modulation branch modifies the spatial distribution of output signals inside the decoder via SPADE blocks [10].

estimation, depth completion, and semantic segmentation as a well-studied case of dense image labeling.

*a) Depth Estimation.*: Methods for single view depth estimation based on deep neural networks have significantly evolved in recent years, by now rapidly approaching the accuracy of depth sensors [11], [12], [13], [14], [15]; some of these methods are able to run in real-time [16] or even on embedded platforms [17]. However, the acquisition of accurate ground truth depth maps is complicated due to certain limitations of existing depth sensors. To overcome these difficulties, various approaches focusing on data acquisition, data refinement, and the use of additional alternative data sources have been proposed [18], [19]. We also note several recently developed weakly supervised and unsupervised approaches [20], [21].

*b) Depth Completion.*: Pioneering works on depth completion adopted complicated heuristic algorithms for processing raw sensor data. These algorithms were based on compressed sensing [24] or used a combined wavelet-contourlet dictionary [25]. Uhrig *et al.* [2] were the first to present a successful learnable depth completion method based on convolutional neural networks, developing special sparsity-invariant convolutions to handle sparse inputs. Learnable methods were further improved by image guidance [5], [26]. Tang *et al.* [4] proposed an approach to train content-dependent and spatially-variant kernels for sparse depth features processing. Li *et al.* [27] suggested a multi-scale guided cascade hourglass architecture for depth completion. Chen *et al.* [28] presented a 2D-3D fusion pipeline based on continuous convolutions. Apart from utilizing images, some recently proposed methods make use of surface normals [6], [29], [30], [31] and object boundaries [29], [31].

Most of the above-mentioned works focus on LiDAR-based sparse depth completion in outdoor scenarios and report results on the well-known KITTI benchmark [2].



(a) Decoder modulation branch. (c) SPADE block.

Fig. 2: Architecture of the decoder modulation branch (2a). It contains a simple encoder composed of two biased convolutions with activations and a series of SPADE blocks (2c). These blocks include the SPADE layer (2b) that performs modulation. We use LeakyReLU activations, as we predict logarithmic depth directly.

There are only a few works that consider processing non-LiDAR semi-dense depth data obtained with Kinect sensors. Recently, Zhang *et al.* [31] introduced *Matterport3D*, a large-scale RGBD dataset for indoor depth completion, and used it to showcase a custom depth completion method. This method implicitly exploits extra data by using pretrained networks for normal estimation and boundary detection, and the resulting normals and boundaries are used in global optimization. Overall, the complexity of this method strictly limits its practical usage. Huang *et al.* [29] was the first to outperform the original results on this dataset. Similar to Zhang *et al.* [31], their results were achieved via a complicated multi-stage method that involved resource-intensive preprocessing. Although it does not rely on pretrained backbones, it uses a normal estimation network explicitly trained on external data. In this work, we propose a novel depth completion method that presents strong baseline results while being scalable and straightforward.

Depth completion and depth estimation can be formulated as a dense labeling problem. Hence, techniques and architectures that are designed for other dense labeling tasks might be useful for depth completion as well. Encoder-decoder architectures with skip connections originally developed for semantic segmentation [32] have shown themselves to be capable of solving a wide range of tasks. Chen *et al.* [33] proposed a powerful architecture based on atrous spatial pyramid pooling for semantic segmentation and improved it in further work [34]. Other important approaches include the refinement network [35] and the pyramid scene parsing network [36]. At the same time, lightweight networks such as [37] capable of running in a resource-constrained device

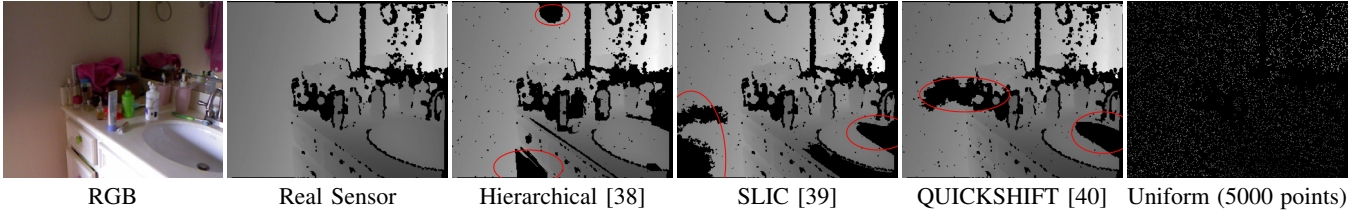


Fig. 3: Qualitative comparison of different sampling strategies based on classical image segmentation methods applied to a raw NYUv2 [3] instance. All methods perform image partitioning, then we replace depth data with zeros in segments with area below a predefined threshold value. Threshold was chosen empirically, this parameter may vary.

in real-time.

### III. PROPOSED APPROACH

*a) Architecture overview.:* The general structure of the proposed architecture is shown in Fig. 1. Our architecture design follows the standard encoder-decoder paradigm with a pretrained backbone network modified for RGBD input. In our experiments, we use the EfficientNet family [22] as a backbone. The decoder part is based on a lightweight RefineNet decoder [23] combined with a custom modulation branch described below. The network takes an image, sensor depth, and a mask as input and outputs a completed depth map. No additional data is required.

*b) Decoder Modulation Branch.:* To introduce the decoder modulation branch, let us take a closer look at the forward propagation path of the network. The backbone network generates feature maps from the input RGBD signal. The input signal initially has an inhomogeneous spatial distribution, since a part of the depth data is missing. The signal compression inside a backbone smoothes this inhomogeneity, which works well for small depth gaps. If the depth gaps are too large, the convolutions generate incorrect activations due to the domain shift between RGB and RGBD signals. Aiming to reduce this domain gap, we propose to learn spatially-dependent scale and bias for normalized feature maps inside the decoder part of the architecture. This procedure is called spatially-adaptive denormalization (SPADE) and was first introduced by Park *et al.* [10].

Let  $f_{n,c,y,x}^i$  denote the activation maps of the  $i$ th layer of the decoder for a batch of  $N$  samples with shape  $C_i \times H_i \times W_i$ , and let  $\mathbf{m}$  denote a modulation signal. The output value from SPADE  $g_{n,c,y,x}^i$  at location  $(n \in N, c \in C_i, y \in H_i, x \in W_i)$  is

$$g_{n,c,y,x}^i = \gamma_{n,c,y,x}^i(\mathbf{m}) \frac{f_{n,c,y,x}^i - \mu_c^i}{\sigma_c^i} + \beta_{n,c,y,x}^i(\mathbf{m}), \quad (1)$$

where  $\mu_c^i = \frac{1}{N_i W_i H_i} \sum_{n,x,y} f_{n,c,y,x}^i$  is the sample mean and  $\sigma_c^i = \sqrt{\frac{1}{N_i W_i H_i} \sum_{n,x,y} (f_{n,c,y,x}^i - \mu_c^i)^2}$  is the sample (biased) standard deviation, and  $\gamma_{n,c,y,x}^i$  and  $\beta_{n,c,y,x}^i$  are the spatially dependent scale and bias for batch normalization respectively. In our case, the modulation signal  $\mathbf{m}$  is the input mask of missing depth values.

Fig. 2 illustrates the decoder modulation branch in detail. This subnetwork consists of a simple mask encoder composed of convolutions with bias terms and activations and

SPADE blocks that perform modulation. A bias term in the convolutions is necessary to avoid zero signals that can cover a significant part of the input mask.

*c) Sensor simulation strategy.:* Existing highly annotated large-scale indoor datasets do not always include both sensor depth data and ground truth depth data [41], [3], which might be an issue for the development of depth completion models. If a dataset provides a reconstruction, different physics based approaches [42], [43] can be used for simulation. But it requires a time and computational resources. If the real sensor data (or reconstructed depth too, but rendering techniques are more suitable in this case) is only available, we propose to use specially developed and fast corruption techniques in order to obtain synthetic semi-dense sensor data.

Let  $t \in T$  be a target sample that we want to degrade. Our goal is to construct a function  $h: T \rightarrow S$  that transforms a depth map from the target domain  $T$  to pseudo-sensor domain  $S$ . We assume that this procedure is sample-specific and can be factorized:  $h(\cdot) = z_g(\cdot|q) \circ z_n(\cdot) = z_n(z_g(\cdot|q))$ , where  $q$  is the input RGB image. The term  $z_g$  emulates a zero masking process guided by the image and  $z_n$  is the zero masking caused by noise. The noise term  $z_n$  represents a random spattering procedure uniformly distributed over the entire image. The specific form of  $z_g$  may vary. Fig. 3 presents some possible approaches results. We performed ablation experiments for the best method among them below.

*d) Loss function.:* Recent works underline two primary families of losses that are conceptually different: pixel-wise and pair-wise. Pixel-wise loss functions measure the mean per-pixel distance between prediction and target, while their pair-wise counterparts express the error by comparing the relationships between pairs of pixels  $i, j$  in the output. The latter loss functions force the relationship between each pair of pixels in the prediction to be similar to that of the corresponding pair in the ground truth. In this work, we have experimented with several different single-term loss functions, including pair-wise and pixel-wise approaches in a logarithmic and actual domain. The logarithmic  $L_1$  pair-wise loss function [44] appears to be the most suitable for our network. It can be expressed as

$$\mathcal{L}(y_i, y_j^*) = \frac{1}{|\mathcal{O}|^2} \sum_{i,j \in \mathcal{O}} \left| \log \frac{y_i}{y_j} - \log \frac{y_i^*}{y_j^*} \right|, \quad (2)$$

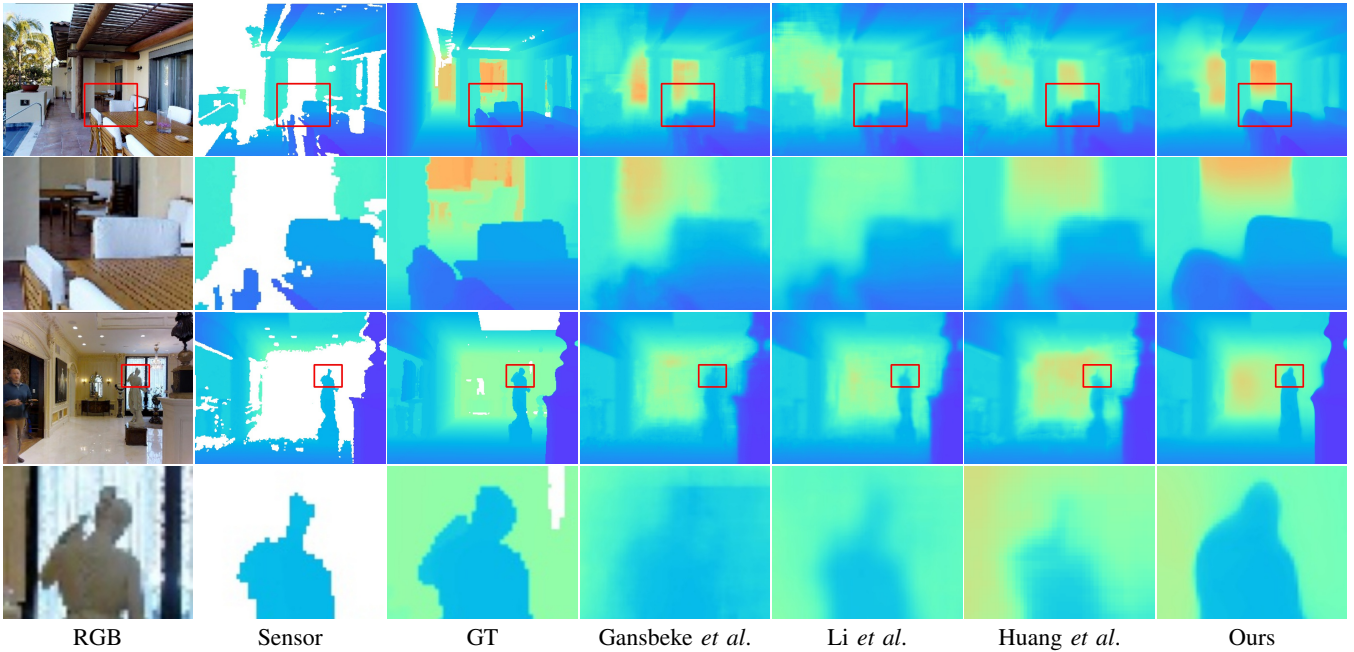


Fig. 4: Qualitative comparison with Gansbeke *et al.* [46], Li *et al.* [47], Huang *et al.* [29] on Matterport3D test set. We train [46] and [47] on Matterport3D using the official code of the corresponding approaches. Results for [29] are based on the official pretrained model. Rows 2 and 4 represent zoomed-in fragments from rows 1 and 3, respectively. All images are created using color maps with the same value limits. Our model generates the completed depth map with very sharp boundaries.

	RMSE ↓	MAE ↓	$\delta_{1.05} \uparrow$	$\delta_{1.10} \uparrow$	$\delta_{1.25} \uparrow$	$\delta_{1.25^2} \uparrow$	$\delta_{1.25^3} \uparrow$	SSIM ↑
Huang <i>et al.</i> [29]	1.092	0.342	0.661	0.750	0.850	0.911	0.936	0.799
Zhang <i>et al.</i> [31]	1.316	0.461	0.657	0.708	0.781	0.851	0.888	0.762
Gansbeke <i>et al.</i> [46]	1.161	0.395	0.542	0.657	0.799	0.887	0.927	0.700
Li <i>et al.</i> [47]	1.054	0.397	0.508	0.631	0.775	0.874	0.920	0.700
Gansbeke <i>et al.</i> [46] (ours)	1.264	0.484	0.675	0.741	0.826	0.888	0.920	0.780
Li <i>et al.</i> [47] (ours)	1.134	0.426	0.649	0.729	0.834	0.899	0.928	0.774
<b>DM-LRN (ours)</b>	<b>0.961</b>	<b>0.285</b>	0.726	<b>0.813</b>	<b>0.890</b>	<b>0.933</b>	0.949	<b>0.844</b>
<b>LRN (ours)</b>	1.028	0.299	0.719	0.805	<b>0.890</b>	0.932	<b>0.950</b>	0.843
<b>LRN + mask (ours)</b>	1.054	0.298	<b>0.737</b>	<b>0.815</b>	0.889	<b>0.933</b>	<b>0.950</b>	<b>0.844</b>

TABLE I: *Matterport3D TEST*. We use the results for Huang *et al.* [29] and Zhang *et al.* [31] reported in [29]. Gansbeke *et al.* [46] and Li *et al.* [47] are trained on Matterport3D using their official implementations. Models labeled as “ours” are trained using our proposed pipeline. The two bottom rows represent models without the decoder modulation branch, with and without the mask on the input. RMSE and MAE are measured in meters.

where  $\mathcal{O}$  is the set of pixels where the ground truth depth exists,  $i, j$  are pixel indices,  $y_i, y_i^*$  are the predicted and target depth respectively. Following Eigen *et al.* [45], our model predicts  $\log y_i$  directly.

#### IV. EXPERIMENTAL SETUP

*a) Datasets.:* We perform comparative experiments on the following datasets: Matterport3D [1], ScanNet [48], NYUv2 [3] and KITTI[2]. Matterport3D includes real sensor data and ground truth depth data obtained from official reconstructed meshes. We use it as the primary target dataset. In order to investigate the generalization capabilities of the model, we perform validation of the models trained on the Matterport3D dataset directly on ScanNet. NYUv2 does not provide dense depth reconstruction for the entire dataset, so we evaluate our sensor simulation strategy on this dataset.

Although our approach is not intended to be applied to sparse depth sensors, we compare it with the best performing models on the KITTI dataset.

*b) Evaluation metrics:* Following the standard evaluation protocol for indoor depth estimation and completion, we use root mean squared error (RMSE), mean absolute error (MAE),  $\delta_i$ , and SSIM. The  $\delta_i$  metric denotes the percentage of predicted pixels where the relative error is less than a threshold  $i$ . Specifically, we evaluate  $\delta_i$  for  $i$  equal to 1.05, 1.10, 1.25, 1.25<sup>2</sup>, and 1.25<sup>3</sup>; smaller values of  $i$  correspond to making the  $\delta_i$  metric more sensitive, while larger values reflect a more accurate prediction. RMSE and MAE directly measure absolute depth accuracy. RMSE is more sensitive to outliers than MAE and is usually chosen as the main metric for ranking models. In general, our testing pipeline for indoor



	RMSE ↓	MAE ↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑	$\delta_{1.25^2}$ ↑	$\delta_{1.25^3}$ ↑	SSIM ↑
Huang <i>et al.</i> [29]	0.244	0.097	0.736	0.850	0.945	0.982	0.992	0.812
Zhang <i>et al.</i> [31]	0.214	0.080	0.769	0.881	0.958	0.985	0.993	0.850
Gansbeke <i>et al.</i> [46]	0.223	0.074	0.829	0.899	0.954	0.980	0.990	0.850
Li <i>et al.</i> [47]	<b>0.190</b>	0.067	0.828	0.903	0.961	<b>0.986</b>	<b>0.995</b>	0.875
DM-LRN (ours)	0.198	<b>0.054</b>	<b>0.900</b>	<b>0.933</b>	<b>0.962</b>	0.982	0.992	<b>0.918</b>

TABLE II: *ScanNet TEST*. Cross-dataset testing demonstrates the strong generalization capability of our method. All models are trained on Matterport3D. RMSE and MAE are measured in meters.

depth completion is similar to Huang *et al.* [29].<sup>1</sup> Following the KITTI leaderboard, we evaluate RMSE, MAE, iRMSE and iMAE metrics on the KITTI dataset.

*c) Implementation details:* In our experiments, we use the Adam [49] optimizer with initial learning rate set to  $10^{-4}$ ,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  and without weight decay. The pretrained EfficientNet-b4 [22] backbone is used unless otherwise stated. Batch normalization is controlled by the modulation process, so we fine-tune its parameters during the first epoch only, and afterwards these parameters are fixed. The training process is performed end-to-end for 100 epochs on a single Nvidia Tesla P40 GPU. We implement all models in Python 3.7 using the PyTorch library [50].

## V. RESULTS

*a) Matterport3D.:* We begin by inferencing our indoor pipeline on the *Matterport3D* dataset. Since very few previous approaches have been tested and achieved good results on this dataset, we train some of the best performing open-source KITTI models [47], [46] for a fair comparison. Assuming that the original training pipeline of these models might be designed specifically for LiDAR data, we also perform a complementary training procedure in our training setup.

The results of this quantitative comparison are presented in Table I. Our training pipeline applied to KITTI models improves the results in terms of  $\delta_i$ , especially with smaller values of  $i$ , but leads to artifacts captured by RMSE values. The original training setup of these methods also does not show state of the art performance on *Matterport3D* (see Table I). We use the original training procedure for further experiments. These methods do not produce sharp edges (see Fig. 4) that are crucial for indoor applications. Zhang *et al.* [31] and Huang *et al.* [29] managed to address this problem and received less blurry results. Our model produces improved completed depth while being more accurate in terms of both RMSE and MAE. In Table I, we also present ablation experiments including different masking strategies.

A visual comparison is shown in Figure 4. Our model keeps the sensor data almost unchanged and sharp. Moreover, the geometric shapes of the interior layout and objects in the scene remain distinct.

<sup>1</sup>The evaluation code is available on the official page <https://github.com/patrickwu2/Depth-Completion>. To keep a fair comparison, we opt for an evaluation procedure based on the official code.

*b) ScanNet.:* In order to evaluate the generalization capability of our method, we conduct a cross-dataset evaluation. Since the test split was not provided for depth completion on *ScanNet*, we use 20% of the original scenes for testing. For the sake of data diversity, we split all frames into intervals of consecutive 20 frames and take one out of each interval. We take the image with the largest variance of Laplacian [51] and the image with the largest file size (which indicates the level of details for a frame). We test the models trained on Matterport3D [1] on this subset that was not seen by the models during the training process. Quantitative results are presented in Table II. Our method significantly improves  $\delta_{1.05}$ ,  $\delta_{1.10}$ , SSIM, and MAE metrics.

*c) NYUv2.:* Since this dataset provides both sensor and reconstruction depth data only for the test subset, we use it to verify our sensor simulation strategy that does not require ground truth. We first cut off black borders (45, 15, 45, 40 pixels from the top, bottom, left, and right side, respectively) from the original  $640 \times 480$  RGBD images. Then the images and depths are interpolated to  $320 \times 256$  resolution using bilinear and nearest-neighbors sampling respectively. These preprocessed RGBD images are used for pseudo sensor data sampling. At test time, the original sensor and ground truth depth data are used. We compare our sampling strategy with the widely used random uniform sampling approach [8], [4]. Qualitative and quantitative results are presented in Fig. 5 and Table III. Since the original semi-dense depth maps contain more accurate information, our training approach demonstrates significant improvements in all target metrics. The compared performance of models originally designed for sparse inputs is shown in Table III. Our model demonstrates strong results in this setup as well.

*d) KITTI.:* In general, this dataset is out of our scope, since it consists of sparse LiDAR depth measurements. It is a hard case for our model, because the architecture includes a unified encoder for the joint RGBD signal, expecting *segments* filled with correct depth values. Previous work [47] has demonstrated that it is a suboptimal design for a sparse depth completion model.

Since LiDAR-based outdoor depth completion differs from our use-case scenario, we perform an additional search for the most suitable loss function. As a result, we have chosen the  $L_2$  loss in the logarithmic domain. As the LiDAR points at the top of an image are rare, input images were cropped to  $256 \times 1216$  for both training and testing, following [4]. A horizontal flip was used as data augmentation.

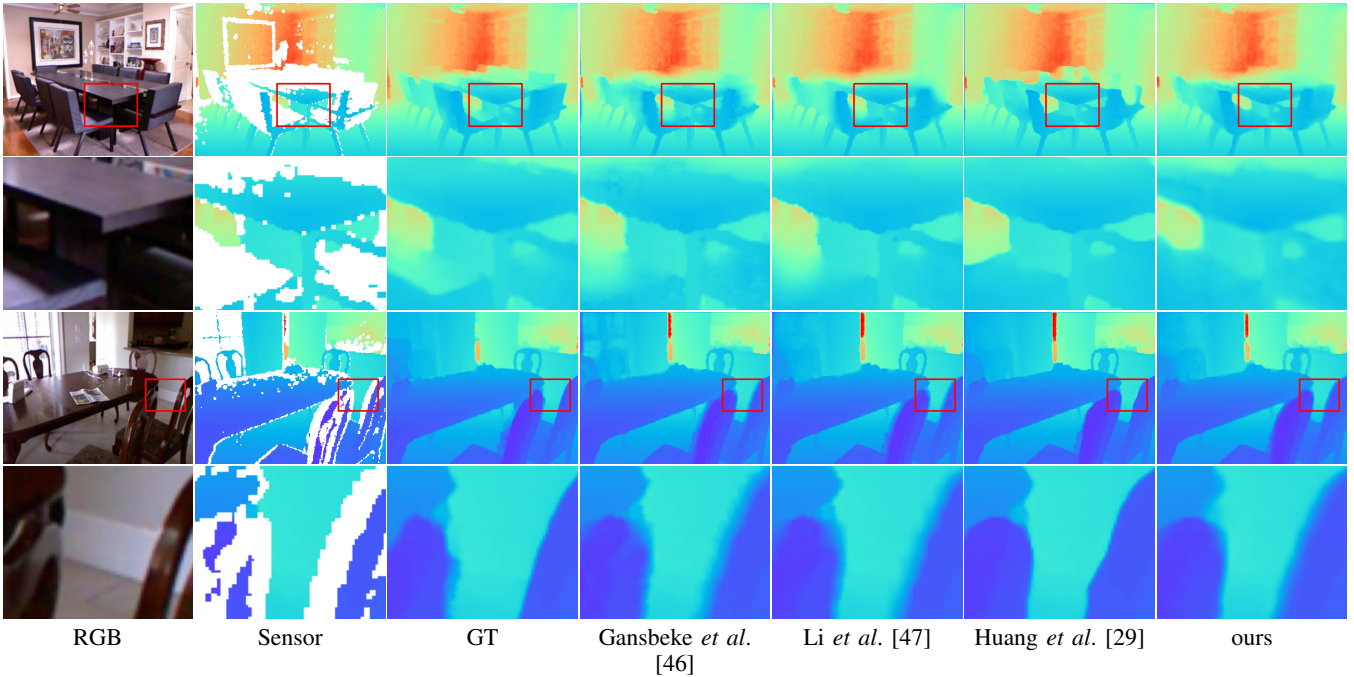


Fig. 5: Qualitative comparison of methods on NYUv2 [3] test set trained using our semi-dense sampling strategy.

	semi-dense (SLIC)					sparse (uniform, 500 points)				
	RMSE ↓	rel ↓	$\delta_{1.25} \uparrow$	$\delta_{1.25^2} \uparrow$	$\delta_{1.25^3} \uparrow$	RMSE ↓	rel ↓	$\delta_{1.25} \uparrow$	$\delta_{1.25^2} \uparrow$	$\delta_{1.25^3} \uparrow$
Huang <i>et al.</i> [29]	0.271	0.016	0.981	0.991	0.994	–	–	–	–	–
Gansbeke <i>et al.</i> [46]	0.240	0.018	0.979	0.994	0.998	0.344	0.042	0.961	0.985	0.995
Li <i>et al.</i> [47]	0.192	<b>0.013</b>	0.988	<b>0.997</b>	<b>0.999</b>	0.272	<b>0.034</b>	0.973	0.992	0.997
DM-LRN (ours)	<b>0.188</b>	0.016	<b>0.989</b>	<b>0.997</b>	<b>0.999</b>	<b>0.263</b>	0.035	<b>0.975</b>	<b>0.993</b>	<b>0.998</b>

TABLE III: *NYUv2 TEST*. Quantitative comparison of training setups for different models. Semi-dense sampling preserves more accurate information that leads to better results. We chose SLIC segmentation for semi-dense sampling. See Section VI for details. Although our DM-LRN model is not intended to be applied to sparse depth sensors, it demonstrates strong results in the sparse training setting in indoor environments. We do not use any densification scheme for target depth reconstruction. Pseudo-sensor data is directly sampled from real sensor data.

A quantitative comparison<sup>2</sup> is shown in Table IV. Being designed for semi-dense sensors, our approach demonstrates mid-level performance compared to the KITTI leaderboard. In general, our model produces accurate depth maps, even though there are some errors at the borders of the image.

	RMSE	MAE	iRMSE	iMAE
Cheng <i>et al.</i> [7]	1019	279	2.93	1.15
Gansbeke [46]	773	215	2.19	0.93
Lee <i>et al.</i> [52]	807	254	2.73	1.33
Qiu <i>et al.</i> [6]	758	226	2.56	1.15
Tang <i>et al.</i> [4]	736	218	2.25	0.99
Chen <i>et al.</i> [28]	753	221	2.34	1.14
Li <i>et al.</i> [47]	762	220	2.30	0.98
Ours	984	287	2.67	1.17

TABLE IV: *KITTI TEST*. Quantitative comparison with top ranked KITTI models. All metrics are measured in millimeters.

<sup>2</sup>Depth visualisation is available on official KITTI benchmark page. It is also available from [this link](#).

## VI. ABLATION STUDY

In these experiments we introduce some changes for our model and simulation baselines to ensure that all components do improve the final result. The architectural ablations were performed on Matterport3D. The results are presented in Table V. Our decoder modulation (DM) branch serves as an adaptive batch normalization allowing to significantly improve relative metrics being slightly better in absolute ones comparing with standard batch normalization.

	RMSE	MAE	$\delta_{1.05}$	$\delta_{1.25}$
no ImageNet pretraining	1.068	0.335	0.669	0.860
no fixed BatchNorm	0.985	0.297	0.609	0.882
no DM branch	1.028	0.299	0.719	0.889
<b>DM-LRN</b>	<b>0.961</b>	<b>0.285</b>	<b>0.726</b>	<b>0.890</b>

TABLE V: *Matterport3D TEST*. Quantitative results with selected feature excluded. RMSE and MAE are measured in meters.

We compared some pseudo-sensor sampling strategies

looking for the most suitable for our scenario. All strategies were investigated on NYUv2 using our network. The summary results are presented in Table VI and VII.

	RMSE	rel	$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
Semi-dense					
QUICKSHIFT [40]	0.201	0.015	0.988	0.997	0.999
SLIC [39]	<b>0.188</b>	0.016	<b>0.989</b>	<b>0.997</b>	<b>0.999</b>
HIERARCHICAL [38]	0.205	<b>0.014</b>	0.988	0.996	0.999
Sparse					
Uniform [8]	0.263	0.035	0.975	0.993	0.998

TABLE VI: NYUv2 TEST. Quantitative results with different segmentation approaches. The threshold value is set to 3000 pixels. RMSE is measured in meters. SLIC segmentation provides the best result.

	RMSE	rel	$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
t = 2000	0.196	0.015	0.989	0.997	0.999
t = 3000	<b>0.188</b>	0.016	<b>0.989</b>	<b>0.997</b>	<b>0.999</b>
t = 4000	0.193	<b>0.014</b>	0.989	0.997	0.999

TABLE VII: NYUv2 TEST. Different threshold values for strategy based on SLIC segmentation. RMSE is measured in meters.

Finally, we performed experiments with backbone depth. As it can be seen from Figure 6, our approach demonstrates stable improvement with respect to backbone complexity. It makes our model scalable.

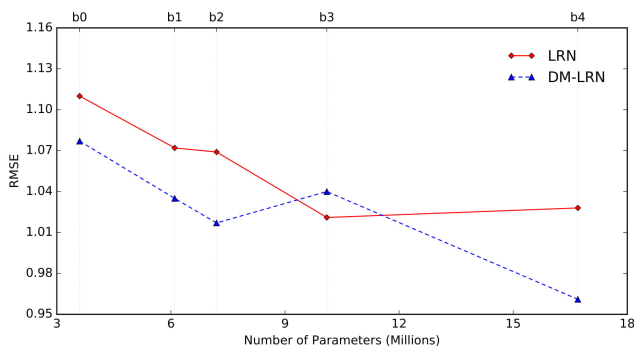


Fig. 6: Matterport3D TEST. A dependency of RMSE of the baseline model and the model with the decoder modulation concerning the size of the backbone. LRN is the baseline model with RGBD inputs. DM-LRN is the baseline with the decoder modulation branch. The mask modulation consistently gives an improvement in the target metric with the exception "B3" configuration that demonstrated an unexpected behavior, assumed to be a random outlier.

## VII. CONCLUSION

In this work, we have proposed a new depth completion method for semi-dense depth sensor maps with auxiliary color images. Our main innovation is a novel decoder architecture that exploits statistical differences between mostly

filled and mostly empty regions. It is implemented by an additional decoder modulation branch that takes a mask of missing values as input and adjusts the activation mask distribution in the decoder via SPADE blocks.

In experimental evaluation, our model has shown state-of-the-art results on the Matterport3D dataset with generalization to ScanNet, and even competitive performance on the KITTI dataset with sparse depth measurements. We have also proposed a new fast sensor simulation strategy for datasets with raw sensor data and without reconstructed ground truth depth, which allows us to achieve strong results on the NYUv2 dataset.

## REFERENCES

- [1] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *International Conference on 3D Vision (3DV)*, 2017.
- [2] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant cnns," *2017 International Conference on 3D Vision (3DV)*, pp. 11–20, 2017.
- [3] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgb-d images," in *ECCV*, 2012.
- [4] J. Tang, F. P. Tian, W. Feng, J. Li, and P. Tan, "Learning guided convolutional network for depth completion," *IEEE Transactions on Image Processing*, vol. 30, pp. 1116–1129, 2021.
- [5] X. Cheng, P. Wang, G. Chenye, and R. Yang, "Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 10615–10622, 04 2020.
- [6] J. Qiu, Z. Cui, Y. Zhang, X. Zhang, S. Liu, B. Zeng, and M. Pollefeys, "DeepLidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [7] X. Cheng, P. Wang, and R. Yang, "Depth estimation via affinity learned with convolutional spatial propagation network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [8] F. Ma and S. Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," 2018.
- [9] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," *ACM Trans. Graph.*, vol. 23, no. 3, p. 689–694, Aug. 2004. [Online]. Available: <https://doi.org/10.1145/1015706.1015780>
- [10] T. Park, M. Liu, T. Wang, and J. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2332–2341.
- [11] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 8001–8008, Jul. 2019. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/4801>
- [12] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2002–2011.
- [13] Z. Liang, Y. Feng, Y. Chen, and L. Zhang, "Learning for disparity estimation through feature constancy," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2811–2820.
- [14] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5695–5703.
- [15] N. Durasov, M. Romanov, V. Bubnova, P. Bogomolov, and A. Konushin, "Double refinement network for efficient monocular depth estimation," 11 2019, pp. 5889–5894.
- [16] Wofk, Diana and Ma, Fangchang and Yang, Tien-Ju and Karaman, Sertac and Sze, Vivienne, "FastDepth: Fast Monocular Depth Estimation on Embedded Systems," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

- [17] "Ambarella cvflow technology overview," <https://www.ambarella.com/technology/technology-overview>, accessed: 2018-10-30.
- [18] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from internet photos," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [19] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.
- [20] H. Ren, A. Raj, M. El-Khamy, and J. Lee, "Suw-learn: Joint supervised, unsupervised, weakly supervised deep learning for monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.
- [21] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6602–6611.
- [22] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, California, USA: PMLR, 09–15 Jun 2019, pp. 6105–6114. [Online]. Available: <http://proceedings.mlr.press/v97/tan19a.html>
- [23] V. Nekrasov, C. Shen, and I. D. Reid, "Light-weight refinenet for real-time semantic segmentation," in *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*. BMVA Press, 2018, p. 125. [Online]. Available: <http://bmvc2018.org/contents/papers/0494.pdf>
- [24] S. Hawe, M. Kleinstueber, and K. Diepold, "Dense disparity maps from sparse disparity measurements," in *2011 International Conference on Computer Vision*, 2011, pp. 2126–2133.
- [25] L. Liu, S. H. Chan, and T. Q. Nguyen, "Depth reconstruction from sparse samples: Representation, algorithm, and sampling," *IEEE Transactions on Image Processing*, vol. 24, no. 6, pp. 1983–1996, 2015.
- [26] Y. Yang, A. Wong, and S. Soatto, "Dense depth posterior (ddp) from single image and sparse range," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [27] A. Li, Z. Yuan, Y. Ling, W. Chi, s. zhang, and C. Zhang, "A multi-scale guided cascade hourglass network for depth completion," in *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.
- [28] Y. Chen, B. Yang, M. Liang, and R. Urtasun, "Learning joint 2d-3d representations for depth completion," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [29] Y.-K. Huang, T.-H. Wu, Y.-C. Liu, and W. H. Hsu, "Indoor depth completion with boundary consistency and self-attention," *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Oct 2019. [Online]. Available: <http://dx.doi.org/10.1109/ICCVW.2019.00137>
- [30] Y. Xu, X. Zhu, J. Shi, G. Zhang, H. Bao, and H. Li, "Depth completion from sparse lidar data with depth-normal constraints," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [31] Y. Zhang and T. Funkhouser, "Deep depth completion of a single rgb-d image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241, (available on [arXiv:1505.04597 \[cs.CV\]](https://arxiv.org/abs/1505.04597)). [Online]. Available: <http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a>
- [33] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
- [34] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 06 2017.
- [35] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," 07 2017, pp. 5168–5177.
- [36] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6230–6239.
- [37] V. Nekrasov, T. Dharmasiri, A. Spek, T. Drummond, C. Shen, and I. Reid, "Real-time joint semantic segmentation and depth estimation using asymmetric annotations," 05 2019, pp. 7101–7107.
- [38] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vision*, vol. 59, no. 2, p. 167–181, Sept. 2004. [Online]. Available: <https://doi.org/10.1023/B:VISI.0000022288.19776.77>
- [39] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.
- [40] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *ECCV (4)*, ser. Lecture Notes in Computer Science, D. A. Forsyth, P. H. S. Torr, and A. Zisserman, Eds., vol. 5305. Springer, 2008, pp. 705–718. [Online]. Available: <http://dblp.uni-trier.de/db/conf/eccv/eccv2008-4.html#VedaldiS08>
- [41] A. R. Zamir, A. Sax, W. B. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.
- [42] M. Gschwandtner, R. Kwitt, A. Uhl, and W. Pree, "Blensor: Blender sensor simulation toolbox," in *Advances in Visual Computing*, G. Bebis, R. Boyle, B. Parvin, D. Koracin, S. Wang, K. Kyungnam, B. Benes, K. Moreland, C. Borst, S. DiVerdi, C. Yi-Jen, and J. Ming, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 199–208.
- [43] B. Planche, Z. Wu, K. Ma, S. Sun, S. Kluckner, O. Lehmann, T. Chen, A. Hutter, S. Zakharev, H. Kosch, and J. Ernst, "Depthsynth: Real-time realistic synthetic data generation from cad models for 2.5d recognition," in *2017 International Conference on 3D Vision (3DV)*, 2017, pp. 1–10.
- [44] M. Romanov, N. Patatkin, A. Vorontsova, and A. Konushin, "Towards general purpose and geometry preserving single-view depth estimation," 2020.
- [45] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2366–2374.
- [46] W. Van Gansbeke, D. Neven, B. De Brabandere, and L. Van Gool, "Sparse and noisy lidar completion with rgb guidance and uncertainty," in *2019 16th International Conference on Machine Vision Applications (MVA)*. IEEE, 2019, pp. 1–6.
- [47] A. Li, Z. Yuan, Y. Ling, W. Chi, C. Zhang, *et al.*, "A multi-scale guided cascade hourglass network for depth completion," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 32–40.
- [48] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3d reconstructions of indoor scenes," in *Proc. Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017.
- [49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015. Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: <http://arxiv.org/abs/1412.6980>
- [50] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.
- [51] J. L. Pech Pacheco, G. Cristobal, J. Chamorro-Martinez, and J. Fernandez-Valdivia, "Diatom autofocus in brightfield microscopy: A comparative study," vol. 3, 02 2000, pp. 314–317 vol.3.
- [52] S. Lee, J. Lee, D. Kim, and J. Kim, "Deep architecture with cross guidance between single image and sparse lidar data for depth completion," *IEEE Access*, vol. 8, pp. 79 801–79 810, 2020.