SEED: Self-Ensembling Depth Completion via Density-Aware Consistency

Anonymous CVPR submission

Paper ID 1596

Abstract

The acquisition of ground truth annotations for depth completion is labor-intensive and non-scalable. Therefore, we resort to semi-supervised learning, where we only need a few labeled images and leverage the rest unlabeled data without ground truth labels to facilitate model learning. In this paper, we propose ***SEED****, a* ***SE****lf-****E****nsembling* ***D****epth completion framework to enhance the generalization of the model on unlabeled and unseen data. Specifically, SEED contains a pair of teacher and student models, which are given high-density and low-density depth maps as input respectively. The main idea underpinning SEED is to enforce density-aware consistency by encouraging consistent prediction across different-density input depth maps. One empirical challenge is that the output of the teacher model, i.e., pseudo-depth labels, inevitably contain wrong depth values, which would mislead the convergence of the student model. To evaluate the reliability of the pseudo-depth labels, we propose an adaptive method to model the pixel-wise uncertainty map via prediction variance automatically. By leveraging the discrepancy of prediction distributions, we explicitly incorporate the uncertainty estimation to rectify the learning from noisy labels. To our knowledge, we are among the early semi-supervised attempts on the depth completion task. Extensive experiments on both indoor and outdoor datasets demonstrate that SEED consistently improves the performance of the baseline model by a large margin and even is on par with several fully-supervised methods.*

1. Introduction

Dense and accurate depth perception is critical information for subsequential applications, such as autonomous driving [13, 14], simultaneously localization and mapping (SLAM) [44], 3D reconstruction [31], and augmented reality (AR) [1]. To obtain the depth of the surrounding environment, various depth sensors are developed such as RGB-D cameras, stereo camera systems, and LiDAR sensors. Among these devices, the RGB-D cameras are not applicable for outdoor scenarios due to the short ranging distance. Stereo algorithms [17, 26] usually fail to predict accurate depth in ill-posed areas and textureless regions. At present, the LiDAR scanners are the most accurate depth perception sensors and have been widely adopted in robots and autonomous driving vehicles. However, existing 3D LiDARs have a limited number of horizontal scan lines and thus provide only sparse measurements [34], *e.g.*, the 64-line Velodyne scan. The sparse depth is insufficient for practical applications such as navigation and planning. Therefore, increasing the density of such sparse data is desirable in real-world scenarios. Depth completion [45] is a promising solution to estimate dense depth map from the sparse depth. Various deep learning based algorithms have achieved significant performance, including depth-only approaches [9, 45] and image-guided methods [5, 6, 18, 32, 38, 42]. The widely-used image-guided approaches take a sparse depth map and aligned RGB image as input, and require densely annotated ground truth for training. Despite the remarkable success, existing methods typically rely on sufficient annotated training data. In the real world, the acquisition of ground truth labels for depth completion can be challenging and not easily scalable [34]. Therefore, one problem occurs: how to improve the generalization of the model on extensive unlabeled data. In this paper, we regard the training data **with ground truth depth annotations as labeled data**, in contrast, only the raw sparse depth captured by LiDARs **without annotations are viewed as unlabeled data.**

Inspired by the success of semi-supervised learning methods [20, 36], we propose a SElf-Ensembling Depth completion framework (SEED) to bring performance gain by leveraging both labeled and unlabeled data. We resort to semi-supervised learning, where we only need to annotate a few images and leverage the rest of unlabeled data to facilitate model training. Theoretically, a robust depth completion model should output similar predictions when given depth maps with different densities as input. However, we notice the predictions inherently fluctuate when the density of input depth changes. This observation inspires us to improve the generalization ability of the model on unlabeled and unseen data by encouraging consensus predictions with different densities depth as inputs. Specifically, we design a self-ensembling paradigm to explore the density-aware consistency on unlabeled data by reducing the prediction gap between high-density and low-density input depth. Taking the raw depth and color image pairs as input, SEED first generates pseudo-depth labels from the prediction of the teacher model on unlabeled images. The student model is then supervised by the pseudo-depth labels when fed a low-density version of the depth map. By simulating the output of the teacher model, the student model is forced to mine more geometry information from the input data and ensure prediction consistency across high density s^u and low density $\widetilde{s}^u$ input depth maps.

The generated pseudo-depth maps inevitably contain incorrect predictions, which would mislead the convergence of the student model. Existing semi-supervised methods [30, 41] filter out the low-score pseudo-labeled samples by manually setting the threshold. But these approaches can not be directly extended to depth completion because the predictions of regression tasks do not have class scores. Therefore, to evaluate the reliability of the pseudo-depth labels, we propose an adaptive method to model the uncertainty [11, 23] via prediction variance automatically. By leveraging the distribution discrepancy of outputs, we model the uncertainty as the prediction variance **without introducing extra parameters or modules.** Specifically, we alleviate the negative effect of noisy labels on the unlabeled data by incorporating the uncertainty rectification into the optimization. SEED can dynamically filter out the unreliable predictions from the teacher model and focus the student model on reliable pseudo annotations according to the uncertainty criterion. Finally, we perform iterative bootstrapping by updating parameters of the teacher model with current student model and re-train a new student. During inference, SEED **only requires the student model** to perform depth completion without the teacher model.

In general, SEED distills the reliable knowledge (high-confidence pseudo-depth labels) from the teacher model to supervise the student model and then updates the knowledge the student model learned back to teacher model. Therefore, we call the semi-supervised training algorithm as a self-ensembling process. Overall, the contributions are as follows:

1. We propose a semi-supervised depth completion framework (SEED) to explore the feasibility of one real-world setting, *i.e.*, limited annotated data and lots of unlabeled data. SEED encourages prediction consistency across high and low density input depth maps on unlabeled data.

2. To alleviate the negative effect of noisy pseudo labels, we propose a density-based method to model the pixel-wise uncertainty adaptively and incorporate the uncertainty estimation to rectify the learning from unreliable pseudo-depth explicitly.

3. We demonstrate the effectiveness of our approach through evaluating on both indoor and outdoor datasets, *i.e.*, KITTI [45] and NYUv2 [37]. Extensive experiments substantiate that SEED consistently improves the performance of the baseline model by a large margin and even is on par with several fully-supervised methods which require full annotations.

2. Related Works

Depth Completion. In recent years, convolutional neural networks have been the dominating solution for depth completion. Based on the modality of input data, previous works can be divided into depth-only methods [7, 45] and image-guided methods [35, 42, 53]. The depth-based methods [7, 45] mainly design different kernels to handle sparse inputs, while the image-guided methods [35, 42, 53] focus on modeling long-range context relationship and fusing multi-model features. Especially, a variety of approaches [5, 6, 18, 32, 38] follow a coarse-to-fine pipeline to recover the dense depth map progressively, producing a sequence of output predictions in the refine process. Unlike previous methods, we exploit both labeled and unlabeled data with the proposed semi-supervised learning strategy.

Semi-Supervised Learning. Semi-supervised learning aims to facilitate model learning with limited labeled data. Early works [20, 36] utilize discriminators to align the distributions of labeled and unlabeled data with an adversarial loss. Pseudo-based methods [2, 41] perform self-training to generate pseudo-labels and train augmented unlabeled data with the corresponding labels. Tarvainen *et al.* [4, 43] regularize the model to be invariant to augmentations injected into the input or parameters. Chen *et al.* [4, 43, 52, 54] leverage the self-training policy to bootstrap the model learning. Although semi-supervised learning has made remarkable progress in classification, there is a paucity of literature focus on depth regression tasks. Kuznietsov *et al.* [27] enforce the network to produce photo-consistent dense depth maps using an image alignment loss. Guizilini *et al.* [16] propose a supervised loss term as auxiliary supervision to complement the photometric loss.

Uncertainty Estimation. Understanding whether a model is under-confident or falsely over-confident can help us to evaluate the reliability of the prediction [11, 12, 23, 55]. Kendall *et al.* [23] present a Bayesian framework to learn a mapping from the input data to aleatoric uncertainty. Considering the homoscedastic uncertainty of each task, Kendall *et al.* [24] propose a principled approach that weighs multiple loss functions to multi-task deep learning. Recently, several approaches are proposed to estimate the

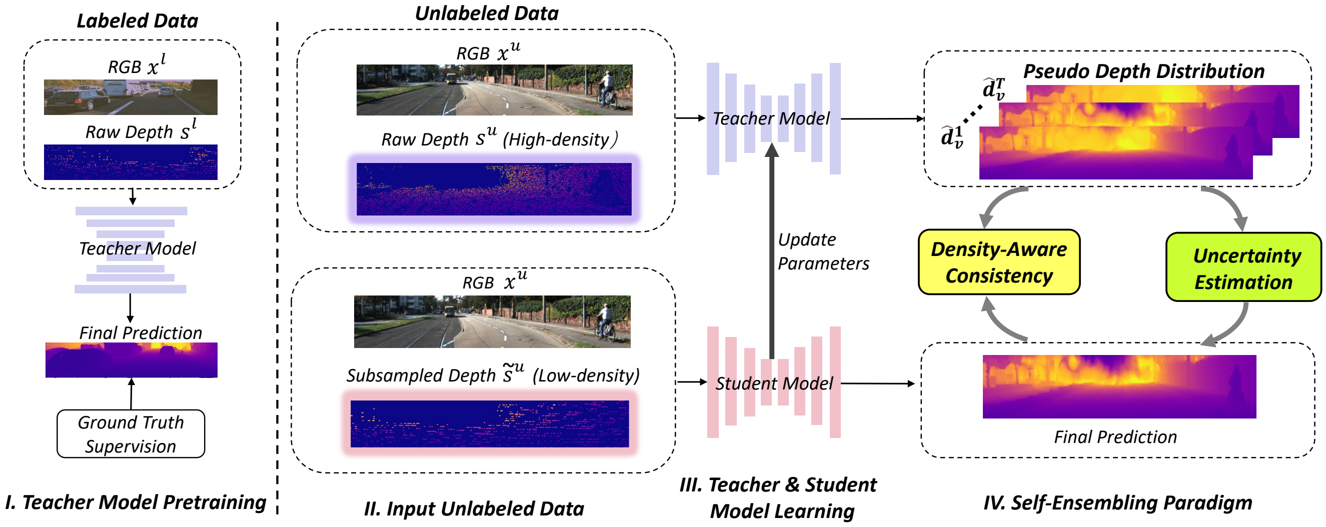

Figure 1. **Overview of SEED**. **I**. The teacher is first trained on limited labeled data with ground-truth annotations. **II**. For the unlabeled data, we cannot obtain the ground-truth label. Given the high-density s^u (raw) depth map and RGB image pair as input, the teacher model generates T predictions in the refine process (from $\hat{d}_v^1$ to $\hat{d}_v^T$) progressively. **III**. The uncertainty is modeled as the prediction variance, which can be estimated from the distribution of T predicted depth maps. Given the low-density $\widetilde{s}^u$ (subsampled) depth map and the corresponding RGB image, the student model is trained to maintain the density-aware consistency with the guidance of uncertainty rectification. **IV**. Finally, we perform iterative bootstrapping by updating parameters of the teacher model with current student model and re-train a new student.

uncertainty on depth regression tasks. Poggi *et al.* [39] make a comprehensive evaluation of uncertainty estimation approaches. Yang *et al.* [50] model the photometric uncertainties of pixels on the input images to improve the accuracy of depth estimation. A probabilistic normalized convolutional neural network is introduced in [8] to produce a statistically meaningful uncertainty measure for the prediction. However, the investigation on how to employ uncertainty to resist the noisy pseudo depth on depth completion is still unexplored. In this work, we attempt to fill this gap by incorporate the uncertainty estimation to rectify the training from unreliable pseudo-depth labels.

3. Our Approach

3.1. Problem Definition

Given both labeled data and unlabeled data, we intend to address depth completion in the semi-supervised setting. Here we use $\mathcal{D}_l = \{(x_i^l, s_i^l, y_i)\}_{i=1}^{N_l}$ and $\mathcal{D}_u = \{(x_i^u, s_i^u)\}_{i=1}^{N_u}$ to represent the labeled and unlabeled training set respectively. N_l and N_u denote the number of images in labeled and unlabeled training set seperately. Concretely, x, s, and y represent the RGB image, the input sparse depth map, and the ground-truth annotation respectively. It is worth noting that the sparse input depth s is sparse scanned data captured by LiDAR sensors, x is the RGB image aligned with s, and y is the ground truth depth map.

3.2. Overview

The brief pipeline of SEED is showed in Fig. 1. Following the common practice in previous approaches [5, 6, 18, 32, 38, 49], SEED adopts the coarse-to-fine structure as the baseline model to be compatible with most state-of-the-art models [5, 6, 18, 32, 38, 49]. We first train the initial teacher model with the limited labeled data. On the unlabeled data, we aim to evolve both the teacher and student models via a density-consistent mechanism, where the teacher and student model are given input depth maps with high-density and low-density respectively. Taking the RGB image and raw depth map (high-density) as input, the teacher model generates the pseudo labels by refining the predictions progressively. However, there exist some pixels do not converge to a stable depth value during the refine process, which indicates the teacher model is not confident in the predictions and thus predictions are not reliable. Therefore, we formulate the uncertainty as the prediction variance of T predicted depth maps (generate from the refine process) to measure the reliability of the pseudo-depth labels. Besides, we explicitly involve the estimated uncertainty in the density-aware consistency to rectify the student model training. The student model fed with low-density depth is trained to be consistent with the high-confidence pseudo-depth labels distilled from the teacher model. Finally, as shown in Fig. 1 (**IV**), we update the parameters of the teacher model with current student model and iterate the process to re-train a new student. During inference,

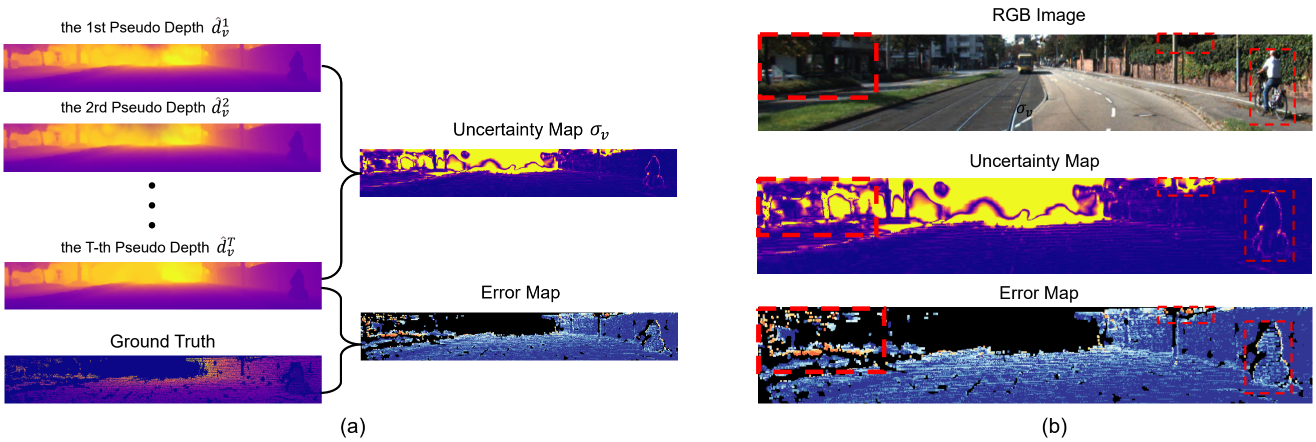
the 1st Pseudo Depth $\hat{d}_v^1$
the 2rd Pseudo Depth $\hat{d}_v^2$
the T-th Pseudo Depth $\hat{d}_v^T$
Ground Truth
Uncertainty Map σ_v
Error Map
(a)
RGB Image
σ_v
Uncertainty Map
Error Map
(b)

Figure 2. **Visualization of the uncertainty and the error map.** (a) We calculate the uncertainty and the error map of the final prediction produced by the teacher model. (b) We make a comparison between the uncertainty and the error map. The red regions of the error map denote large errors while the yellow areas denote high variance in the uncertainty map. As shown in the red dotted box, the high uncertainty area has remarkable overlaps with the large-error region.

only the student model is required to conduct depth completion without the need of the teacher model. In general, we perform the self-ensembling paradigm by distilling the reliable knowledge from the teacher model to the student model and then updating the knowledge the student model learned back to teacher model.

3.3. Density-Aware Consistency

We propose SEED, a semi-supervised learning algorithm to enhance the generalization of the model on unlabeled and new unseen data. In theory, a robust depth completion model should output similar predictions when given depth maps with different densities as input. However, we notice the predictions inherently fluctuate when the density of input depth changes. As observed in [42, 45], the performance drops significantly with the density of the input depth map decreasing. This observation inspires us to promote the generalization of the model on unlabeled data by encouraging consensus predictions with different densities depth maps as inputs. Specifically, we design a teacher-student paradigm to explore the density-aware consistency on unlabeled data by reducing the prediction gap between high-density and low-density. We first use the labeled data to train a teacher model with the conventional supervised learning methods. Following the practice in [5,6,18,32,38], we adopt the widely-used coarse-to-fine pipeline to refine predictions progressively. As shown in the Fig. 1(b), SEED generates pseudo-depth labels of the unlabeled data with the raw depth and the corresponding color image as input. Under the supervision of the teacher model, the student model learns to recover the dense depth map by minimizing the combined reconstruction loss as follows:

$$\mathcal{L}_{reconstruct} = \frac{1}{|\mathcal{V}|} \sum_{\rho \in \{1,2\}} \sum_{v \in \mathcal{V}} |\hat{d_v^T} - d_v^T|^{\rho}, \qquad (1)$$

where $\hat{d_v^T}$ is the T-th pseudo-depth label (final prediction of the teacher model) of point v, d_v^T is the T-th prediction (final prediction) of the student model at point v, and $\mathcal{V}$ represents the point set of the full depth map. The combined reconstruction loss is the sum of $\mathcal{L}_1$ loss ($\rho = 1$) and $\mathcal{L}_2$ loss ($\rho = 2$). By reducing the density of the raw input depth deliberately, the student model is trained to be consistent with the output of the teacher model that takes raw depth (high-density) as input. Considering the teacher are provided with more depth information, there exist certain regions where the prediction of the student is not accurate enough while the teacher performs better. Through mimicking the output of the teacher model, the student is forced to learn harder from pseudo-depth labels and explore the structural information from the RGB guidance with limited depth information.

3.4. Noise-robust Uncertainty Rectification

The pseudo-depth labels generated from the teacher model inevitably contain incorrect predictions and can mislead the convergence of the student model. To alleviate the negative effect of noisy labels, we resort to the uncertainty [11, 23] to evaluate the reliability of pseudo-depth maps. Taking the widely-used coarse-to-fine model [5, 6, 18, 32, 38] as the baseline, the teacher model generates the pseudo labels through a refinement process, where a sequence of predictions are produced progressively. Under ideal circumstances, every pixel of the predicted pseudo-depth maps will converge to a stable depth value in the refinement process (from $\hat{d}_v^1$ to $\hat{d}_v^T$). However, there are some hard pixels and regions do not converge, which indicates the model is not confident in the predictions and thus the predictions are unreliable. Therefore, we propose to model the uncertainty by the prediction variance of teacher model's output (see Fig. 1 (c)), because the variance between the intermediate and final outputs can measure the predictive stability of the teacher model. Specifically, the uncertainty σ_v of the point v is calculated from the distribution of T predictions:

$$\sigma_v = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{d}_v^t - \mu_v)^2}, \quad \mu_v = \frac{1}{T}\sum_{t=1}^{T}\hat{d}_v^t, \qquad (2)$$

where $\hat{d}_v^t$ denotes the t-th prediction of the teacher model and the μ_v represents the prediction mean of the point v. To verify the effectiveness of our uncertainty estimation, we make a comparison between the error map and uncertainty map. As illustrated in the Fig. 2, the red regions of the error map denote large errors while the bright yellow areas denote high variance in the uncertainty map. We observe that the high-variance area has remarkable overlaps with the large-error region (the red dotted box), showing the strong correlation between the error map and uncertainty map. It is worth noting that plenty of high-uncertain regions are concentrated in the middle part of the picture, which confirmed the accuracy of the uncertainty map because these remote areas are beyond the perception range of the depth sensor (corresponding to the black area in the error map).

To resist the noise of pseudo-depth annotations, we explicitly incorporate uncertainty rectification into the optimization. We reshape the loss function to down-weight noisy points and thus focus training on reliable pseudo-depth points. The new loss function is formulated as:

$$\mathcal{L}_{reconstruct} = \frac{1}{|\mathcal{V}|}\sum_{\rho\in\{1,2\}}\sum_{v\in\mathcal{V}}\alpha_v|\hat{d_v^T} - d_v^T|^{\rho}, \qquad (3)$$

where α_v is the new added weighting factor compared to Eq. 1 and depends on the uncertainty σ_v. Our goal is to reduce the contribution of the noisy points by assigning a lower weighting factor to the point with higher uncertainty. Here we adopt a mapping function $\alpha_v = e^{-\sigma_v}$ to compute the weights according to the uncertainty criteria. Other inversely correlated function such as $\alpha_v = \frac{1}{\sigma_v+1}$ also works in our experiments. When the uncertainty σ_v equals to zero, the optimization loss degrades to conventional supervised learning with ground truth labels ($\alpha_v = 1$). In contrast, for the points with ambiguous predictions ($\sigma_v \rightarrow +\infty$), the proposed uncertainty rectification guides the model to neglect noisy pseudo-depth labels ($\alpha_v \rightarrow 0$). Finally, following the self-training manner [47], we iterate the process by updating the parameters of the teacher model with the student model and re-train a new student. In the inference stage, SEED **only requires the student model** to make predictions without the teacher model. In general, our method takes advantage of the multiple outputs of the model itself to estimate the uncertainty **without introducing the extra modules or Gaussian noise**. Besides, we also provide an alternative choice to exploit the uncertainty estimated by the image flipping, which can be applied seamlessly to any depth completion frameworks. The details can be found in the ablation study of the experiment section.

Dataset	KITTI				NYUv2			
	Full	1/2	1/4	1/8	Full	1/8	1/16	1/32
Frames^l	85,898	41,042	22,048	13,792	47,584	6,443	3,016	1,598
Sequences^l	138	69	34	17	280	35	17	8

Table 1. The partition protocols of two datasets. The Frames^l and Sequences^l represent the number of labeled frames and sequences under different partition protocols respectively. For example, 1/8 denotes there are 1/8 labeled data and the rest data are divided into unlabeled set.

4. Experiment

4.1. Datasets and Evaluation Metrics

KITTI. The KITTI depth completion dataset [45] is a large outdoor dataset for autonomous driving. The standard training, validation, and test sets consist of 85,898, 1,000, and 1,000 frames respectively. For the training data, there are 138 recording image sequences in total. Each sequence consists of consecutive image frames captured by the sensors. Following the partition protocols of previous semi-supervised works [3, 33, 41], we divide the whole training set into two groups via randomly **sub-sampling 1/2, 1/4, and 1/8 sequences of the whole set as the labeled set** and regard **the remaining sequences as the unlabeled set**. The number of labeled frames and sequences under different partition protocols can be found in Tab. 1.

NYUv2. The NYUv2 [37] dataset consists of the RGBD sequences from 464 indoor scenes captured by the Microsoft Kinect. The standard training set consists of 47,584 RGBD images. To make a fair comparison with previous approaches [6, 35, 38, 42], we first down-size the input frames to 320 × 240, and center-crop to the resolution of 304 × 228. Following early methods [5, 38, 53], we use 500 randomly sampled points as the sparse input. Similar to the semi-supervised partition protocols of KITTI [45], we divide the standard training sequences into the labeled and unlabeled sets. Considering NYUv2 [37] is simpler than KITTI DC [45], we randomly subsample with smaller ratios: **1/16, 1/32, and 1/64 of total sequences from the training set to construct the labeled set.** The number of labeled frames and sequences in different partition protocols is shown in Tab. 1.

Baseline	DC	UR	IS	RMSE$_{KITTI}$ (mm) ↓ 1/8	 1/4	 1/2
✓				884.2	858.8	830.7
✓	✓			859.9	850.2	820.2
✓	✓	✓		853.5	836.1	810.8
✓	✓	✓	✓	851.6	834.2	808.1

Table 2. Ablation study of designed components. ”DC” means depth-aware consistency. ”UR” indicates the proposed uncertainty rectification. ”IS” represents the iterative self-training [47].

Evaluation Metrics. To make a fair comparison with existing works [5, 6, 35, 38] on the KITTI benchmark [45], we adopt four commonly-used metrics for quantitative evaluation: root mean squared error (RMSE), mean absolute error (MAE), root mean squared error of the inverse depth (iRMSE), and mean absolute error of the inverse depth (iMAE). For the indoor dataset NYUv2 [37], the evaluation metrics are selected as root mean squared error (RMSE) and mean absolute relative error (REL) and δ_τ which means the percentage of predicted pixels where the relative error is less the threshold τ [6, 35, 40]. It is worth noting that **RMSE is chosen as the primary metric** for all the experiment evaluations.

4.2. Implementation Details

Following the practice in [5, 6, 18, 32, 38], we adopt the widely-used coarse-to-fine structure to conduct experiments. If not specified, our algorithm takes NLSPN [38] as the baseline model. For a fair comparison, we follow the previous works [5, 6, 18, 38] to set the number of refine steps $T = 18$. In practice, the growth of the prediction performance tends to be flat after about half of total refine steps. Therefore, we calculate the variance on the second half sequences of predictions. In our experiments, we adopt an ADAM optimizer [25] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and the initial learning rate of 0.001. For the unlabeled data, we take the original raw input as the high-density depth and generate the low-density version by randomly sub-sampling. The sub-sample ratio is a range of random numbers less than 1.

4.3. Ablation Studies

Impact of Density-aware Consistency. As shown in Tab. 2, we conduct experiments on KITTI [45] under 1/8, 1/4, and 1/2 partition protocols (see Tab. 1) respectively. We first train the supervised baseline model on the labeled data with reconstruction loss in Eq. 1. For the baseline with supervised learning, the RMSE for 1/8, 1/4, and 1/2 annotated settings is 884.2mm, 858.8mm, and 830.7mm. With the proposed depth-aware consistency, SEED can improve the performance significantly and reduce the RMSE by 24.3mm, 8.6mm and 10.5mm for 1/8, 1/4, and 1/2 settings respectively.

Method	Uncertainty	Function	RMSE$_{KITTI}$ (mm) ↓ 1/8	 1/4	 1/2
Baseline	-	-	884.2	858.8	830.7
Ours	UFILP	INV	856.2	843.3	815.8
Ours	UFILP	EXP	856.6	844.4	818.7
Ours	UVAR	INV	852.3	835.2	809.2
Ours	UVAR	EXP	851.6	834.2	808.1

Table 3. Different design choices of uncertainty rectification. The UFLIP represents the image-flipping uncertainty and UVAR denote the proposed prediction variance-based uncertainty. INV and EXP denote the functions $\alpha_v = \frac{1}{\sigma_v+1}$ and $\alpha_v = e^{-\sigma_v}$ respectively.

Impact of Uncertainty Rectification. We study the effectiveness of the proposed uncertainty rectification. From Tab. 2, we find that our algorithm can better handle the noisy pseudo-depth annotations. Specifically, the RMSE can significantly reduce from 859.9mm to 853.5mm, 850.2mm to 836.1mm, and 820.2mm to 810.8mm for 1/8, 1/4 and 1/2 settings respectively. Furthermore, we explore different design choices of the uncertainty rectification, *i.e.*, the variations of uncertainty estimation and the mapping function. For the uncertainty estimation, SEED models the uncertainty as the prediction variance of the output distribution. Inspired by [15, 39], an alternative strategy is designed to estimate uncertainty by image flipping. This strategy requires predicting two depth maps: d and $\overleftarrow{d}$ for original input x and the horizontally flipped counterpart $\overleftarrow{x}$. The uncertainty estimated by image flipping can be formulated as the difference between d and $\overset{\rightleftarrows}{d}$ (the back-flipped $\overleftarrow{d}$): $\sigma_{flip} = |\, d - \overset{\rightleftarrows}{d} \,|$. We denote UFLIP as the image-flipping based uncertainty and UVAR as the proposed prediction-variance based uncertainty. As shown in Tab. 3, both the UVAR and UFLIP are compatible with our framework, while the UVAR performs better than UFLIP. Therefore, it shows that the proposed variance-based uncertainty UVAR can better measure the reliability and stability of the predicted pseudo-depth annotations. When calculating the weighting factor, we also conduct experiments with another inversely correlated function $\alpha_v = \frac{1}{\sigma_v+1}$. We deploy INV to represent the function $\alpha_v = \frac{1}{\sigma_v+1}$ and EXP denote the function $\alpha_v = e^{-\sigma_v}$. As illustrated in Tab. 3, we observe that both the two mapping functions are compatible with the proposed uncertainty rectification, which shows the robustness of our algorithm.

Impact of Iterative Training. We empirically study the impact of iterative self-training in our semi-supervised framework. Tab. 2 shows that the iterative learning can further boost performance. We also study the impact of the iteration number in Tab. 4. We observe that performance

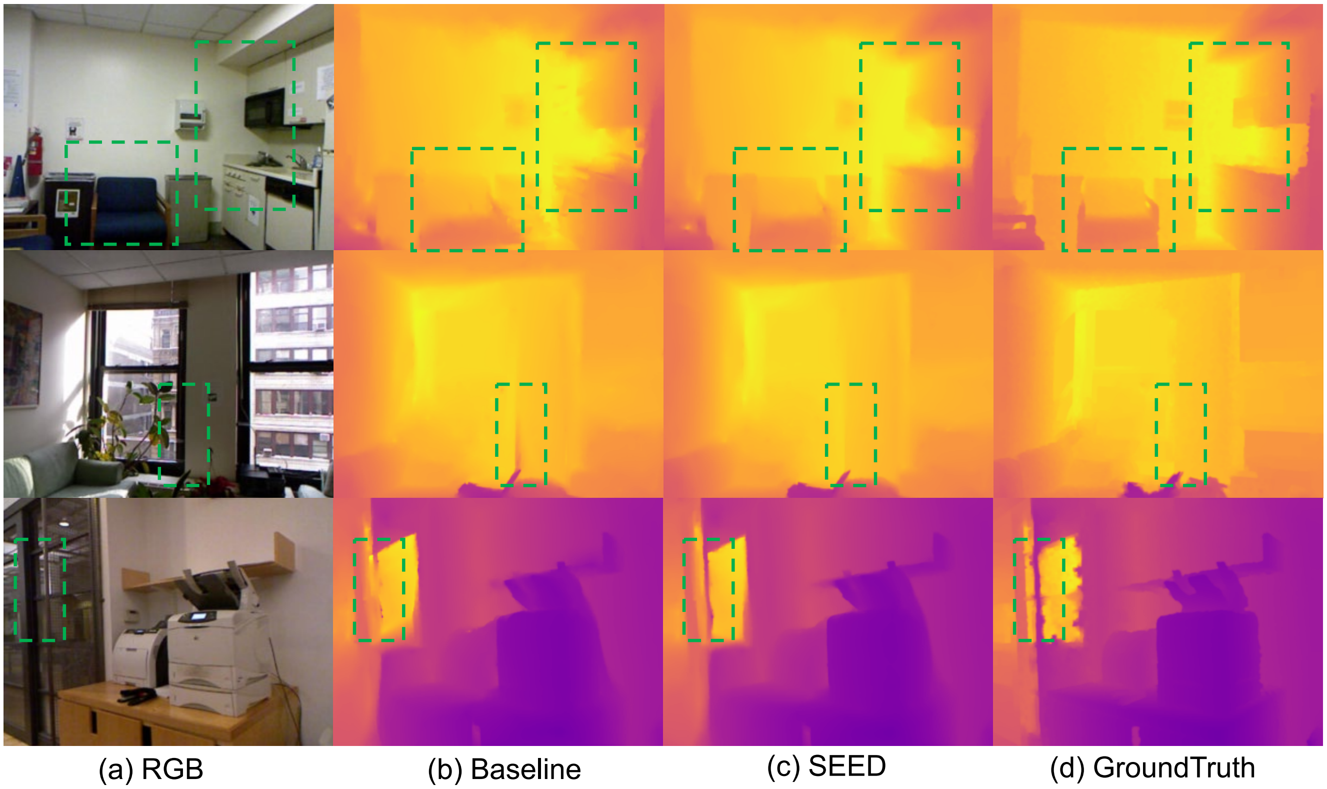

Figure 3. **Depth completion results on the NYUv2 [37] dataset**. Our approach (SEED) achieves more detailed and precise depth completion results. As illustrated in the green box, our method can preserve the small structure information near the depth boundaries.

Iteration	RMSE$_{KITTI}$ (mm) ↓			RMSE$_{NYUv2}$ (m) ↓		
	1/8	1/4	1/2	1/32	1/16	1/8
1	853.5	836.1	810.8	0.1090	0.1046	0.0982
2	852.9	834.8	808.4	0.1066	0.1025	0.0971
3	851.6	834.2	808.1	0.1060	0.1016	0.0970

Table 4. The impact of iterative training. We observe that more iterations generally boost performance, and the improvement will gradually flatten out.

Method	Model	RMSE$_{KITTI}$ (mm) ↓			RMSE$_{NYUv2}$ (m) ↓		
		1/8	1/4	1/2	1/32	1/16	1/8
Baseline	CSPN	904.2	868.2	844.6	0.1247	0.1168	0.1083
Ours	CSPN	885.5	854.6	833.6	0.1149	0.1090	0.1042
Baseline	NLSPN	884.2	858.8	830.7	0.1181	0.1122	0.1036
Ours	NLSPN	851.6	834.2	808.1	0.1060	0.1016	0.0970

Table 5. Performance with different baselines. We implement our approach with two commonly-used depth completion models: CSPN [6] and NLSPN [38]. Our approach consistently improves the performance on two models under different partition protocols. The results verify that our algorithm is compatible to different depth completion models.

gradually converges to a stable value with the number of iterations increasing.

Impact of Different Baselines. To demonstrate the generalization of our algorithm, we also adopt other state-of-the-art baseline models to conduct experiments on KITTI [45] and NYUv2 [37]. We implement our approach with two commonly-used depth completion models, *i.e.*, CSPN [6] and NLSPN [38]. As shown in Table 5, SEED consistently improves the performance on two models for different split protocols. On the KITTI dataset [45], SEED based on NLSPN [6] significantly reduces the RMSE by 32.6mm, 24.6mm, and 22.6mm for 1/8, 1/4, and 1/2 labeled settings respectively, while on CSPN [6], the performance gain is 18.7mm, 13.6mm, and 11.0mm. Besides, we also observe similar improvement in NYUv2 dataset [37]. The results verify that SEED has strong generalization and is compatible to different depth completion models.

4.4. Comparison with the State-of-the-arts

We compare SEED with the state-of-the-art semi-supervised and fully-supervised methods on two datasets.

On the semi-supervised setting, we compare our method with some recent competitive semi-supervised methods including Mean Teacher [43], Temporal Ensemble [28], and FixMatch [41]. We compare them using the same base-

Method	RMSE$_{KITTI}$ (mm)↓			RMSE$_{NYUv2}$ (m) ↓		
	1/8	1/4	1/2	1/32	1/16	1/8
MT [43]	878.2	857.0	828.8	0.116	0.110	0.101
TE [28]	876.1	854.6	827.9	0.114	0.109	0.101
FM [41]	863.4	850.7	821.0	0.113	0.108	0.100
Ours	**851.6**	**834.2**	**808.1**	**0.106**	**0.102**	**0.097**

Table 6. Comparison with the semi-supervised state-of-the-art methods. ”MT” indicates Mean Teacher [43], and ”TE” denotes Temporal Ensemble [28]. ”FM” represents FixMatch [41]. our approach performs best on all partition protocols and outperforms other methods by a large margin.

Method	Label	RMSE↓ m	REL↓ m	$\delta_{1.25}$ ↑	$\delta_{1.25^2}$ ↑	$\delta_{1.25^3}$ ↑
S2D [35]	Full	0.123	0.026	99.1	99.9	100.0
DepthCoeff [22]	Full	0.118	0.013	99.4	99.9	-
CSPN [6]	Full	0.117	0.016	99.2	99.9	100.0
CSPN++ [5]	Full	0.116	-	-	-	-
DeepLiDAR [40]	Full	0.116	0.022	99.3	99.9	100.0
PwP [48]	Full	0.112	0.018	99.5	99.9	100.0
FCFR [32]	Full	0.106	0.015	99.5	99.9	100.0
ACMNet [53]	Full	0.105	0.012	99.6	99.9	100.0
PRNet [29]	Full	0.104	0.014	99.4	99.9	100.0
TWISE [21]	Full	0.097	0.013	99.6	99.9	100.0
NLSPN [38]	Full	0.092	0.012	99.6	99.9	100.0
Ours (semi-supervised)	**1/32**	0.106	0.015	99.5	99.9	100.0
Ours (semi-supervised)	**1/16**	0.102	0.014	99.5	99.9	100.0
Ours (semi-supervised)	**1/8**	0.097	0.013	99.5	99.9	100.0

Table 7. Comparisons with the **fully-supervised** state-of-the-art on the NYUv2 [37] test dataset. With only the 1/8 ground-truth annotations, our method yields close performance to the best performing fully-supervised model.

line architecture and partition protocols. Table 7 shows that the comparison results on KITTI [45] validation and NYUv2 [37] test set. In comparison with other SOTA methods, SEED performs best on all partition protocols. For example, our method achieves 851.6mm RMSE with 1/8 annotations on KITTI dataset [45], which outperforms Mean-Teacher [43], Temporal Ensemble [28] and FixMatch [41] by 26.6mm 24.5mm and 11.8mm respectively. As shown in the right part of Table 6, SEED acquires remarkable performance on NYUv2 dataset [37]. We also visualize the predicted results of NYUv2 in Fig. 3. SEED can preserve the small structure information near the depth boundaries, which demonstrate the effectiveness of the proposed method.

To further verify the effectiveness of our approach, we also compare our method with fully-supervised state-of-the-art methods. The detailed quantitative comparison results are illustrated in Table 7. On NYUv2 [37] dataset, with only 1/32 and 1/16 groundtruth annotations, SEED consistently performs better than most of previous works and even on par with the latest work [32,53]. In particular, SEED yields close performance to the best performing fully-supervised model with only 1/8 ground truth annotations. We also provide the comparison results of experiments on KITTI benchmark [45]. From Table 8, we observe that the SEED achieves 816.75mm, 794.01mm, and 778.96mm under the 1/8, 1/4 and 1/2 partition protocols respectively. The results verify that our algorithm can significantly narrow the gap between semi-supervised and fully-supervised learning methods.

Method	Label	RMSE↓ mm	MAE↓ mm	iRMSE↓ 1/km	iMAE↓ 1/km
CSPN [6]	Full	1019.64	279.46	2.93	1.15
HMS [19]	Full	841.78	253.47	2.73	1.13
TWISE [21]	Full	840.20	195.58	2.08	0.82
DDP [51]	Full	832.94	203.96	2.10	0.85
NCNN [10]	Full	829.98	233.26	2.60	1.03
S2D [35]	Full	814.73	249.95	2.80	1.21
3dDepthNet [46]	Full	798.44	226.27	2.36	1.02
PwP [48]	Full	777.05	255.17	2.42	1.13
NLSPN [38]	Full	741.68	199.59	1.99	0.84
Ours (semi-supervised)	**1/8**	816.75	217.17	2.12	0.91
Ours (semi-supervised)	**1/4**	794.01	213.52	2.10	0.90
Ours (semi-supervised)	**1/2**	778.96	211.41	2.08	0.89

Table 8. Comparison with the **fully-supervised** state-of-the-art on the KITTI test dataset [45].

5. Conclusion and Discussion

To explore the feasibility of leveraging unlabeled data for depth completion, we introduce SEED, a semi-supervised learning algorithm that achieves remarkable performance on both indoor and outdoor datasets. By enforcing the density-aware consistency, we tackle input depth maps across different densities while ensembling the reliable information of the teacher and student models. We further propose to exploit the uncertainty to resist the noisy pseudo-depth labels in training process. Extensive experiments demonstrate that the density-aware consistency and uncertainty-regularizing optimization can bring significant performance gain. However, our method is still inferior to the state-of-the-art fully-supervised approaches in terms of prediction accuracy. In the future, we believe that incorporating extra information, such as geometry constraints and segmentation masks, can extend the proposed method to more complex real-world scenarios.

Broader Impacts. This work will contribute to a wide range of applications involving perception and recognition, such as autonomous driving and robot navigation, We do not foresee obvious undesirable ethical and social impacts at this moment.

References

[1] Ronald T Azuma. A survey of augmented reality. *Presence: Teleoperators & Virtual Environments*, 6(4):355–385, 1997. 1

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019. 2

[3] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, 2021. 5

[4] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Semi-supervised deep learning with memory. In *ECCV*, pages 268–283, 2018. 2

[5] Xinjing Cheng, Peng Wang, Chenye Guan, and Ruigang Yang. Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In *AAAI*, volume 34, pages 10615–10622, 2020. 1, 2, 3, 4, 5, 6, 8

[6] Xinjing Cheng, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2361–2379, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[7] Nathaniel Chodosh, Chaoyang Wang, and Simon Lucey. Deep convolutional compressed sensing for lidar depth completion. In *ACCV*, pages 499–513. Springer, 2018. 2

[8] Abdelrahman Eldesokey, Michael Felsberg, Karl Holmquist, and Michael Persson. Uncertainty-aware cnns for depth completion: Uncertainty from beginning to end. In *CVPR*, pages 12014–12023, 2020. 3

[9] Abdelrahman Eldesokey, Michael Felsberg, and Fahad Shahbaz Khan. Propagating confidences through cnns for sparse data regression. In *BMVC*, 2018. 1

[10] Abdelrahman Eldesokey, Michael Felsberg, and Fahad Shahbaz Khan. Confidence propagation through cnns for guided sparse depth regression. *IEEE transactions on pattern analysis and machine intelligence*, 42(10), 2019. 8

[11] Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016. 2, 4

[12] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059. PMLR, 2016. 2

[13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361, 2012. 1

[15] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, pages 270–279, 2017. 6

[16] Vitor Guizilini, Jie Li, Rares Ambrus, Sudeep Pillai, and Adrien Gaidon. Robust semi-supervised monocular depth estimation with reprojected distances. In *CoRL*, pages 503–512. PMLR, 2020. 2

[17] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007. 1

[18] Mu Hu, Shuling Wang, Bin Li, Shiyu Ning, Li Fan, and Xiaojin Gong. Penet: Towards precise and efficient image guided depth completion. In *ICRA*, 2021. 1, 2, 3, 4, 6

[19] Zixuan Huang, Junming Fan, Shenggan Cheng, Shuai Yi, Xiaogang Wang, and Hongsheng Li. Hms-net: Hierarchical multi-scale sparsity-invariant network for sparse depth completion. *IEEE Transactions on Image Processing*, pages 3429–3441, 2019. 8

[20] Wei Chih Hung, Yi Hsuan Tsai, Yan Ting Liou, Yen Yu Lin, and Ming Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. In *BMVC*, 2019. 1, 2

[21] Saif Imran, Xiaoming Liu, and Daniel Morris. Depth completion with twin surface extrapolation at occlusion boundaries. In *CVPR*, pages 2583–2592, 2021. 8

[22] Saif Imran, Yunfei Long, Xiaoming Liu, and Daniel Morris. Depth coefficients for depth completion. In *CVPR*, 2019. 8

[23] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, 2017. 2, 4

[24] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, pages 7482–7491, 2018. 2

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[26] Vladimir Kolmogorov and Ramin Zabih. Computing visual correspondence with occlusions using graph cuts. In *ICCV*, volume 2, pages 508–515, 2001. 1

[27] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *CVPR*, pages 6647–6655, 2017. 2

[28] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017. 7, 8

[29] Byeong-Uk Lee, Kyunghyun Lee, and In So Kweon. Depth completion using plane-residual representation. In *CVPR*, pages 13916–13925, 2021. 8

[30] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICMLW*, volume 3, 2013. 2

[31] Huei-Yung Lin and Jing-Ren Wu. 3d reconstruction by combining shape from silhouette with stereo. In *CVPR*, pages 1–4, 2008. 1

[32] Lina Liu, Xibin Song, Xiaoyang Lyu, Junwei Diao, Mengmeng Wang, Yong Liu, and Liangjun Zhang. Fcfr-net: Feature fusion based coarse-to-fine residual learning for depth completion. In *AAAI*, volume 35, pages 2136–2144, 2021. 1, 2, 3, 4, 6, 8

[33] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *ICLR*, 2021. 5

[34] Fangchang Ma, Guilherme Venturelli Cavalheiro, and Sertac Karaman. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In *ICRA*, pages 3288–3295. IEEE, 2019. 1

[35] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *ICRA*, 2018. 2, 5, 6, 8

[36] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1, 2

[37] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2, 5, 6, 7, 8

[38] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In-So Kweon. Non-local spatial propagation network for depth completion. In *ECCV*. European Conference on Computer Vision, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[39] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *CVPR*, pages 3227–3237, 2020. 3, 6

[40] Jiaxiong Qiu, Zhaopeng Cui, Yinda Zhang, Xingdi Zhang, Shuaicheng Liu, Bing Zeng, and Marc Pollefeys. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In *CVPR*, 2019. 6, 8

[41] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 2, 5, 7, 8

[42] Jie Tang, Fei-Peng Tian, Wei Feng, Jian Li, and Ping Tan. Learning guided convolutional network for depth completion. *IEEE Transactions on Image Processing*, 30:1116–1129, 2020. 1, 2, 4, 5

[43] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, pages 1195–1204, 2017. 2, 7, 8

[44] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *CVPR*, pages 6243–6252, 2017. 1

[45] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *3DV*, pages 11–20, 2017. 1, 2, 4, 5, 6, 7, 8

[46] Rui Xiang, Feng Zheng, Huapeng Su, and Zhe Zhang. 3ddepthnet: Point cloud guided depth completion network for sparse depth and single color image. In *CVPR*, 2020. 8

[47] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, pages 10687–10698, 2020. 5, 6

[48] Yan Xu, Xinge Zhu, Jianping Shi, Guofeng Zhang, Hujun Bao, and Hongsheng Li. Depth completion from sparse lidar data with depth-normal constraints. In *ICCV*, 2019. 8

[49] Zheyuan Xu, Hongche Yin, and Jian Yao. Deformable spatial propagation networks for depth completion. In *ICIP*, pages 913–917. IEEE, 2020. 3

[50] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *CVPR*, pages 1281–1292, 2020. 3

[51] Yanchao Yang, Alex Wong, and Stefano Soatto. Dense depth posterior (ddp) from single image and sparse range. In *CVPR*, 2019. 8

[52] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pages 4320–4328, 2018. 2

[53] Shanshan Zhao, Mingming Gong, Huan Fu, and Dacheng Tao. Adaptive context-aware multi-modal network for depth completion. *IEEE Transactions on Image Processing*, 2021. 2, 5, 8

[54] Zhedong Zheng and Yi Yang. Unsupervised scene adaptation with memory regularization in vivo. *IJCAI*, 2020. 2

[55] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision (IJCV)*, 2021. 2