Self-Ensembling Depth Completion via Density-aware Consistency

Xuanmeng Zhang[a], Zhedong Zheng[c,d,*], Minyue Jiang[b], Xiaoqing Ye[b]

[a]*ReLER, AAII, University of Technology Sydney*
[b]*Department of Computer Vision Technology, Baidu Inc.*
[c]*Faculty of Science and Technology, University of Macau*
[d]*Institute of Collaborative Innovation, University of Macau*

Abstract

Depth completion can predict a dense depth map by taking a sparse depth map and the aligned RGB image as input, but the acquisition of ground truth annotations is labor-intensive and non-scalable. Therefore, we resort to semi-supervised learning, where we only need to annotate a few images and leverage massive 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 data. Specifically, SEED contains a pair of the teacher and student models, which are given high-density and low-density sparse depth maps as input respectively. The main idea underpinning SEED is to enforce the density-aware consistency by encouraging consistent prediction across different-density input depth maps. One empirical challenge is that the pseudo-depth labels produced by the teacher model inevitably contain wrong depth values, which would mislead the convergence of the student model. To resist the noisy labels, we propose an automatic method to measure the reliability of the gener-

*Zhedong Zheng is the corresponding author. (Email address: zhedongzheng@um.edu.mo)

ated pseudo-depth labels adaptively. By leveraging the discrepancy of prediction distributions, we model the pixel-wise uncertainty map as the prediction variance and rectify the training process from noisy labels explicitly. To our knowledge, we are among the early semi-supervised attempts on the depth completion task. Extensive experiments on both outdoor and indoor 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.

Keywords:
Depth Completion, Semi-supervised Learning, Density-aware Consistency, Uncertainty Estimation.

1. Introduction

Dense and accurate depth perception is critical for subsequential applications, such as simultaneously localization and mapping (SLAM), autonomous driving, and augmented reality (AR). To obtain the depth of the surrounding environment, various depth sensors have been 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 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 autonomous driving vehicles and mobile robots. However, current LiDARs can only obtain sparse depth perceptions because the number of horizontal scan lines is limited, *e.g.*, the 32-line Velodyne scanner. The sparse depth is insufficient for many practical applications such as navigation and planning [1]. Depth maps, as 2.5D representations,

have been widely used in real-world applications such as scene understanding and scene representation [2, 3]. A promising way to recover the dense depth map from the sparse depth input is depth completion [4, 5]. In the past few years, the deep learning-based depth completion algorithms have achieved significant performance. The widely-used image-guided depth completion approaches [6, 7] take a sparse depth map and the 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 cost of acquiring ground truth labels for supervised learning is challenging and not easily scalable [1]. Due to the massive sparse and noisy points captured by the LiDAR, the ground truth labels (dense depth maps) of depth completion are scarce, which are a major limitation of supervised depth completion. Therefore, one problem occurs: how to improve the generalization of the model on massive unlabeled data. In this work, 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 huge success of semi-supervised learning methods, we propose a **SE**lf-**E**nsembling **D**epth (**SEED**) completion framework to bring performance gain by leveraging both the labeled data and unlabeled data. We resort to semi-supervised learning, where we only need to annotate a few training samples and leverage the rest unlabeled data to facilitate model training. Theoretically, a robust depth completion model should make consistent predictions when given sparse depth maps with different densities as input. However, we observe that the predicted depth values 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-density depth maps as inputs. Specifically, we design a self-ensembling paradigm to explore the density-aware consistency of unlabeled data by reducing the prediction gap between high-density and low-density input depth maps. Taking the raw depth and color image pairs as input, the teacher model of SEED first generates pseudo-depth with the unlabeled data as input. Then the student model is supervised by the pseudo-depth when fed a low-density version of the depth map. With the output of the teacher model as supervision, 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.

One empirical challenge is that the generated pseudo-depth maps [3] inevitably contain incorrect predictions, which would mislead the convergence of the student model. Existing semi-supervised methods [8, 9] filter out the low-score pseudo-labeled samples by manually setting the threshold. But these approaches cannot 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 perform uncertainty estimation automatically. By leveraging the distribution discrepancy of outputs, we model the uncertainty as the prediction variance **without introducing extra modules and parameters.** Then we incorporate the uncertainty rectification into the optimization to tackle the problem of noisy pseudo-depth labels. SEED can dynamically filter out unreliable predictions from the teacher model and focus the student model on reliable pseudo annotations according to the uncertainty criterion [10]. Besides, we perform iterative training by updating the parameters of the teacher model with the 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 the student model, then the knowledge learned by the student model is fed back to the teacher model (self-training). Therefore, we call the semi-supervised training algorithm a self-ensembling paradigm. Overall, the contributions are as follows:

- We present a semi-supervised depth completion method (SEED) for a real-world scenario, *i.e.*, limited annotations and massive unlabeled data. (1) To boost the generalization of the model, we propose a self-ensembling framework by enforcing the density-aware consistency on the unlabeled data. (2) To resist the noisy pseudo labels, we propose a variance-based uncertainty estimation method to rectify the learning from unreliable pseudo-depth labels.

- We demonstrate the effectiveness of the proposed approach by evaluating on both outdoor and indoor datasets, *i.e.*, KITTI [4] and NYUv2 [11]. 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.

2. Related Work

Depth Completion. Based on the modality of input data, previous works can be divided into two categories, *i.e.*, depth-only methods [4] and image-guided methods [12, 13, 14]. The depth-based methods [4] only take a sparse depth

map scanned from the LiDAR [15] as the input. In the early years, traditional approaches mainly perform depth completion with classical image processing algorithms. However, with the rising of deep neural networks, convolutional neural network-based models have drawn wide attention due to their extraordinary performance on computer vision tasks. Uhrig *et al.* [4] integrate an effective sparse convolution operator into neural networks to process the sparse depth data. Similarly, Huang *et al.* [16] propose more complex sparsity-invariant layers with multi-scale and hierarchical architectures. To tackle the noisy input data, Eldesokey *et al.* [17] propose to learn the uncertainty of depth maps in a self-supervised manner. The normalized convolutional neural networks [17] improve the interpretability of models and outperform existing Bayesian deep networks by a large margin.

Another line of depth completion approaches are image-guided methods [12], which employ an additional RGB image as the guidance to further improve the performance. Tang *et al.* [18] introduce a novel three-stage multi-scale training framework (BP-Net) with three-stage training strategy, incorporating depth refinement and multi-modal fusion with bilateral propagation. Yan *et al.* [19] model the scene geometry with tri-Perspective view decomposition (TPVD).

Recently, SPN-based approaches [6, 7, 20] have gained a surge of interest due to the extraordinary performance for depth completion. Specifically, they adopt the spatial propagation networks (SPNs) to refine the dense depth map progressively, producing a sequence of output predictions in the refine process. Cheng *et al.* [6] first introduce the convolutional spatial propagation network to learn the affinity of neighboring pixels. To further improve the effectiveness and efficiency of the convolutional spatial propagation network, Cheng *et al.* [7] pro-

pose to learn the number of iterations and the kernel sizes adaptively.

Semi-Supervised Learning. Semi-supervised learning aims to facilitate model training with limited labels. Early works utilize discriminators to calibrate the distributions of unlabeled data and labeled data with an adversarial loss. Pseudo-labels-based methods [9] first generate pseudo-labels on unlabeled data and then perform self-training with strong augmentations. Chen *et al.*[21] introduce a multi-task mean teacher model for semi-supervised shadow detection. Although semi-supervised learning has made significant progress in classification, there is a paucity of literature focusing on depth regression tasks. Kuznietsov *et al.* [22] introduce an image alignment loss to guide the model to predict the photo-consistent depth maps.

Uncertainty Estimation. Understanding whether a model is under-confident or falsely over-confident can help us to evaluate the reliability of the prediction [23]. Kendall *et al.* [23] present a Bayesian framework to map the input data to the aleatoric uncertainty. Recently, several approaches are proposed to estimate the uncertainty of depth regression tasks. Poggi *et al.* [24] make a comprehensive evaluation of uncertainty estimation approaches. Eldesokey *et al.* [17] introduce a probabilistic convolutional network with meaningful statistical interpretability for the prediction. However, it remains unexplored how to employ uncertainty to resist the noisy pseudo-depth labels of the depth completion task. In this paper, we attempt to fill this gap by incorporating uncertainty estimation to rectify the training from unreliable pseudo-depth labels.

3. Approach

3.1. Problem Definition

For depth completion, the samples with ground truth depth annotations are regarded as labeled data, while only the raw sparse depth maps captured by LiDARs without annotations are viewed as unlabeled data. Given both labeled and unlabeled data, we intend to address depth completion in a semi-supervised manner. 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 sets separately. Concretely, x, s, and y represent the RGB image, the input sparse depth map, and the ground truth depth labels respectively. It is worth noting that the sparse input depth s is the scanned raw data captured by LiDAR sensors, x is the RGB image aligned with s, and y is the densely-annotated ground-truth depth map.

3.2. Preliminaries

Spatial propagation networks. Spatial propagation networks (SPNs) [25] have proven to be an effective depth refinement module for depth completion [6, 7, 20]. The SPN-based model mainly contains a U-Net module and a spatial linear propagation module. The U-Net module learns the affinity matrix among neighboring pixels and predicts an initial depth map, while the spatial linear propagation module performs propagation from the initial depth map under the guidance of the learned affinity matrix. After propagating T steps progressively, we can obtain the refined depth map d^T. In the fully-supervised setting, the predicted depth map is supervised by the ground-truth labels with $\mathcal{L}_1$ and $\mathcal{L}_2$ loss:

$$\mathcal{L}_{reconstruct}^{labeled} = \frac{1}{|\mathcal{V}|} \sum_{\rho \in \{1,2\}} \sum_{v \in \mathcal{V}} \left| d_v^{gt} - d_v^T \right|^\rho , \tag{1}$$

where d_v^T is the T-th prediction (final prediction) of point v, and $\mathcal{V}$ represents the valid point set of the ground truth depth map. The combined reconstruction loss is the sum of $\mathcal{L}_1$ loss ($\rho = 1$) and $\mathcal{L}_2$ loss ($\rho = 2$). In this paper, we adopt SPN-based model as the baseline to be compatible with most state-of-the-art depth completion models [6, 7, 20].

3.3. Overview

The brief pipeline of SEED is shown in Fig. 1. SEED contains a pair of the teacher and student models, which are given high-density and low-density sparse depth maps as input respectively. We first follow the common practice of fully-supervised approaches [6, 7, 20] to train the initial teacher model on limited labeled data (Fig. 1 **I**). On the unlabeled data, we aim to evolve both the teacher and student models via the density-consistent mechanism (see Sec. 3.4 and Fig. 1 **II**). Taking the RGB image and the aligned raw depth map (high-density) as input, the teacher model generates the pseudo labels by refining the predictions progressively. The student model fed with low-density depth is trained to be consistent with the prediction of the teacher model. However, the pseudo-depth labels inevitably contain depth values with large errors, which would mislead the convergence of the student model. Therefore, we aim to filter out the unreliable noisy pixels from the pseudo-depth labels. We observe that some pseudo-labels converge at an early stage of the refinement, while predictions on other pixels do not converge to a stable depth value during the refine process. The disconvergence of depth estimation on these pixels indicates that the teacher model is not confident about the predictions and these pseudo-depth labels are not reliable. Therefore, we formulate the uncertainty as the variance of T predicted depth maps (generate in the refine process) to measure the reliability of the pseudo-depth labels. The

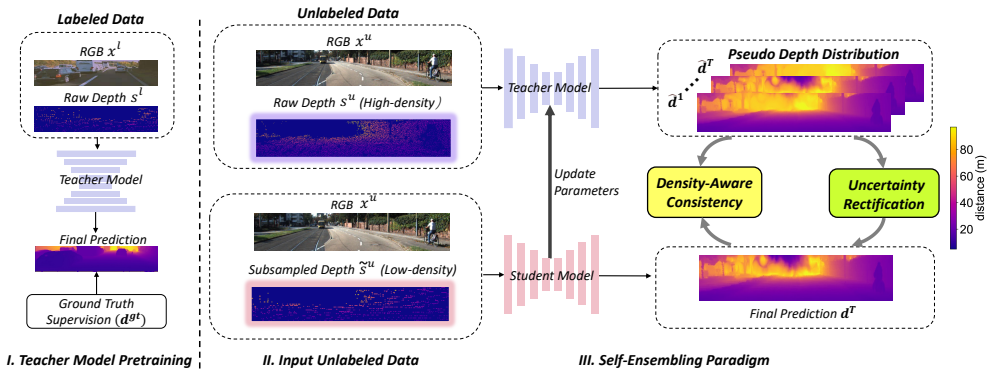

Figure 1: **Overview of SEED**. **I**. The teacher model is first trained on limited labeled data with ground-truth annotations. **II**. For unlabeled data, we train the models in a self-ensembling manner via the density-aware consistency. Given the high-density depth map s^u and RGB image x^u as input, the teacher model generates T predictions in the refine process progressively (from $\hat{d}^1$ to $\hat{d}^T$). **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 depth map $\widetilde{s}^u$ (subsampled) and the corresponding RGB image x^u, the student model is trained to maintain density-aware consistency under the guidance of uncertainty rectification. Finally, we perform iterative bootstrapping by updating the parameters of the teacher model with current student model and re-train a new student.

estimated uncertainty is explicitly involved into the density-aware consistency to rectify the student model training (see Sec. 3.5 and Fig. 1 **III**). Finally, we update the parameters of the teacher model with the current student model and iterate the process to re-train a new student. During inference, only the student model is required to conduct depth completion. In general, we perform the self-ensembling paradigm by distilling the reliable knowledge from the teacher model to the student model, and then push the knowledge the student model learned back to the teacher model.

3.4. Density-Aware Consistency

We propose SEED, a semi-supervised learning framework to improve the generalization of the model on unlabeled and unseen data. Ideally, a robust depth

completion model should output consistent predictions when given depth maps with different densities as input. However, we notice that the predictions inherently fluctuate when the density of input depth changes. As observed in [4], the performance drops significantly with the density of the input depth map decreasing. This observation inspires us to promote the generalization of the model of unlabeled data by encouraging consistent predictions across different-density depth maps. Specifically,we design a teacher-student paradigm to explore the density-aware consistency of 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 method [6, 7, 20]. The widely-used spatial propagation networks (SPNs) are employed to refine the predicted depth maps progressively. Taking the raw depth and the color image as input, SEED can generate pseudo-depth labels on the unlabeled data (see Fig. 1 **II**). 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:

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

where $\hat{d_v^T}$ is the pseudo-depth label (the 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. By reducing the input density of the student model deliberately, we enforce consistent predictions between high-density and low-density input. Considering the teacher model is provided with more input depth information, there exist certain regions where the prediction of the student is not accurate enough while the teacher performs better. By mimicking the output of the teacher model, the student is forced to learn harder

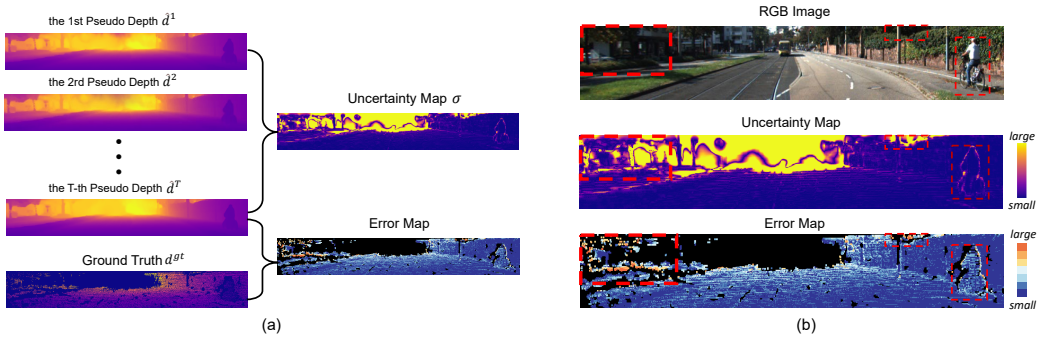

Figure 2: **Visualization of the uncertainty and the error map.** (a) We calculate the uncertainty and the error map of the final prediction produced by the teacher model. (b) We make a comparison between the uncertainty and the error map. For the uncertainty map, the yellow area denotes the high variance in the uncertainty map while the blue region represents the small variance. The error map depicts correct predictions in blue and wrong predictions in the red-color pixels. As shown in the red dotted box, the high uncertainty area has remarkable overlaps with the large-error region.

from pseudo-depth labels, and explore the structural information from the RGB guidance with limited depth information.

3.5. Noise-robust Uncertainty Rectification

The pseudo-depth labels generated by the teacher model inevitably contain incorrect predictions, which can mislead the convergence of the student model. Here we cannot directly use the error map to measure the reliability of pseudo-depth maps, because the error map is calculated as the difference between the predictions and the ground truth labels, which are unavailable to the unlabeled data. Therefore, we resort to the uncertainty [23] to evaluate the reliability of pseudo-depth maps. Specifically, aleatoric uncertainty encompasses the inherent noise present in observations, such as sensor or motion noise. This type of uncertainty persists, regardless of the quantity of collected data. In contrast, epistemic uncertainty addresses the uncertainty associated with model parameters. This uncer-

tainty reflects our lack of knowledge about which model generated the collected data, often referred to as model uncertainty. In this paper, we strive to incorporate both epistemic and aleatoric uncertainty into a unified model. This is particularly relevant for the semi-supervised depth completion task, where both sensor noise / motion noise and insufficient data contribute to uncertainties in prediction results. For the SPN-based baseline [6, 7, 20], the teacher model generates the pseudo labels through a refinement process, producing a sequence of intermediate predictions progressively. During the refine process, the spatial propagation is conducted iteratively with the confidence-incorporated learnable affinity normalization. The intermediate predictions (from 1-st to T-th step) are generated in different steps of the refine process. Ideally, every pixel of the pseudo-depth maps tends to converge to a stable depth value (from $\hat{d}_v^1$ to $\hat{d}_v^T$). However, some hard pixels and regions do not converge after the refinement, which indicates that the teacher model is not confident about the predictions and the pseudo-depth labels are unreliable. Therefore, we propose to model the uncertainty by the prediction variance of the teacher model's output (see Fig. 1 **III**), because the variance between the intermediate outputs (generated in the refine process) 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:

$$\sigma_v = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{d}_v^t - \mu_v)^2}, \quad \mu_v = \frac{1}{T}\sum_{t=1}^{T}\hat{d}_v^t, \tag{3}$$

where $\hat{d}_v^t$ represents the t-th prediction of the teacher model, μ_v denotes the prediction mean of the point v. To further verify the effectiveness of the proposed uncertainty estimation, we also make a comparison between the error map and uncertainty map and qualitatively and quantitatively. For the qualitative compar-

ison, we visualize both the calculated uncertainty map and the error map. As illustrated in the Fig. 2, the red region of the error map denotes the significant error while the bright yellow area denotes high variance in the uncertainty map. We observe that the large-error region (highlighted in the red dotted box) has remarkable overlaps with the high-variance area, which indicates the strong correlation between the errors and the uncertainties. It is worth noting that plenty of highly-uncertain pixels are concentrated in the middle part of the picture, which validates 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). For the quantitative comparison, we sort all pixels of the depth maps in the ascending order of uncertainty and then calculate the average errors. We observe that the pixels with higher uncertainties have larger RMSE (see more details in Sec. 4.3 and Fig. 3).

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 labels and focus the training on reliable pseudo-depth labels. The new loss function is formulated as:

$$\mathcal{L}_{reconstruct}^{unlabeled} = \frac{1}{|\mathcal{V}|} \sum_{\rho \in \{1,2\}} \sum_{v \in \mathcal{V}} \alpha_v |\hat{d}_v^T - d_v^T|^\rho, \tag{4}$$

where α_v is the newly-added weighting factor compared to Eq. 2 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. We adopt the mapping function $\alpha_v = e^{-\sigma_v}$ to adjust the weights according to the uncertainty criteria. Here α_v serves as a dynamic threshold to filter out the noisy pseudo-depth labels automatically. When the uncertainty σ_v equals to zero, it indicates that the teacher model is very confident about the prediction. In this case, 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 can guide the model to neglect noisy pseudo-depth labels ($\alpha_v \rightarrow 0$). After finishing the uncertainty rectification, following the self-training manner, 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 extra modules or Gaussian noise**. Besides, the proposed uncertainty rectification mechanism can also be incorporated with other uncertainty estimation methods [26, 23, 27] (see more details in Sec. 4.3).

4. Experiment

We conduct comprehensive experiments on both indoor and outdoor datasets. We first give a brief description of datasets and evaluation metrics, and then describe the implementation details of the proposed method. After that, we conduct extensive ablation experiments to study the individual component of the proposed framework. Next, we make a comparison with the state-of-the-art methods in both semi-supervised and fully-supervised settings. Finally, to further demonstrate the generalization ability, we also extend our approach to a more challenging task, *i.e.*, domain adaptation.

4.1. Datasets and Evaluation Metrics

KITTI. The KITTI depth completion dataset [4] is a large outdoor autonomous driving dataset. The standard training, validation, and test sets contain 85,898,

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 that there are 1/8 labeled samples while the rest are unlabeled data.

Labeled Datasets	KITTI			NYUv2		
	1/2	1/4	1/8	1/8	1/16	1/32
# Frames^l	41,042	22,048	13,792	6,443	3,016	1,598
# Sequences^l	69	34	17	35	17	8

1,000, and 1,000 frames separately. For the training data, there are 138 recording image sequences in total. Each sequence contains a set of consecutive depth frames and RGB images captured by the sensors and cameras. Following the partition protocols of previous semi-supervised works [9], we divide the training set into two groups, *i.e.*, labeled and unlabeled data, with different proportions. Specifically, we randomly **choose 1/8, 1/4, and 1/2 sequences data from the training dataset as the labeled set** and regard **the remaining sequences as the unlabeled data**. The number of labeled frames and sequences under different partition protocols can be found in Tab. 1.

NYUv2. The NYUv2 dataset [11] is an indoor dataset collected by Microsoft Kinect and consists of RGBD sequences from 464 scenes. The standard training dataset consists of 47,584 RGBD images. In order to make a fair comparison with exisiting approaches[12, 6, 20], we down-size the input frames to the resolution of 320 $\times$ 240, and use the center-crop to the resolution of 304 $\times$ 228. Following early methods [20, 7], we randomly sample 500 points from the depth map as the input sparse depth map. Similar to the semi-supervised partition protocols of KITTI, we divide the standard training sequences into labeled and unlabeled sets. Considering NYUv2 [11] is simpler than KITTI DC [4], we randomly subsample with smaller ratios: **1/16, 1/32, and 1/64 of total training sequences to construct**

the labeled set. The number of labeled frames and sequences in different partition protocols is shown in Tab. 1.

TartanAir. TartanAir [28] is a large-scale virtual dataset collected in the synthetic simulation environments for robot navigation and autonomous driving tasks. There are 1037 sequences in total, which cover a wide range of scenarios in the real world, from the unstructured natural environments to the structured indoor scenes. To conduct the domain adaptation experiments, we select two indoor sequences, *i.e.*, office and office2, to construct the source dataset for domain adaptation experiments (126,924 images in total), and the NYUv2 [11] are chosen as the target dataset.

Evaluation Metrics. To make a fair comparison with existing works [12, 6, 7, 20] on the KITTI benchmark [4], we adopt four widely-used metrics for quantitative evaluation: root mean squared error of the inverse depth (iRMSE), root mean squared error (RMSE), mean absolute error (MAE), and mean absolute error of the inverse depth (iMAE). For the indoor dataset NYUv2 [11], we select three evaluation metrics: the mean absolute relative error (REL), the root mean squared error (RMSE), and δ_τ (the percentage of predicted pixels where the relative error is less than the threshold τ) [29, 12, 6]. These metrics are formulated as follows:

$$\begin{aligned}
&\textbf{RMSE}\text{: } \sqrt{\tfrac{1}{|\mathcal{V}|}\textstyle\sum_{v\in\mathcal{V}}\left|d_v^{gt}-d_v^{T}\right|^2}, && \textbf{MAE}\text{: } \tfrac{1}{|\mathcal{V}|}\textstyle\sum_{v\in\mathcal{V}}\left|d_v^{gt}-d_v^{T}\right|,\\
&\textbf{iRMSE}\text{: } \sqrt{\tfrac{1}{|\mathcal{V}|}\textstyle\sum_{v\in\mathcal{V}}\left|\tfrac{1}{d_v^{gt}}-\tfrac{1}{d_v^{T}}\right|^2}, && \textbf{iMAE}\text{: } \tfrac{1}{|\mathcal{V}|}\textstyle\sum_{v\in\mathcal{V}}\left|\tfrac{1}{d_v^{gt}}-\tfrac{1}{d_v^{T}}\right|,\\
&\textbf{REL}\text{: } \tfrac{1}{|\mathcal{V}|}\textstyle\sum_{v\in\mathcal{V}}\left|\tfrac{d_v^{gt}-d_v^{T}}{d_v^{gt}}\right|, && \boldsymbol{\delta_\tau}\text{: } max\left(\tfrac{d_v^{gt}}{d_v^{T}},\tfrac{d_v^{T}}{d_v^{gt}}\right)<\tau,
\end{aligned}$$

where d_v^{gt} is the ground truth of pixel v. It is worth noting that **we choose RMSE as the primary metric** in all experiments by default.

Table 2: Ablation study of designed components. "PL" means pseudo labels. "DC" means density-aware consistency. "UR" indicates the proposed uncertainty rectification. "IS" represents the iterative self-training.

Group	PL	DC	UR	IS	RMSE$_{KITTI}$ (mm) ↓ 1/8	1/4	1/2
I					884.2	858.8	830.7
II	✓				880.1	856.3	828.3
III	✓	✓			859.9	850.2	820.2
IV	✓	✓	✓		853.5	836.1	810.8
V	✓	✓	✓	✓	851.6	834.2	808.1

4.2. Implementation Details

We adopt a representative spatial propagation network (SPN) framework NLSPN [20] to conduct experiments. If not specified, our approach is implemented based on the NLSPN [20]. For a fair comparison, we follow the previous works [6, 7, 20] to set the number of refine steps $T = 18$. The growth of the prediction performance tends to be flat after about half of the total refine steps. Therefore, we calculate the variance on the second-half sequences of predictions. We adopt an ADAM optimizer 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 dynamic number randomly sampled from the range of [0.5, 1.0].

4.3. Ablation Studies

Effects of Density-aware Consistency. As shown in Tab. 2, we conduct experiments on KITTI [4] under 1/8, 1/4, and 1/2 partition protocols 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 (Group I). As shown in the Tab. 2 Group II, only performing prediction mimicking (pseudo-labeling)

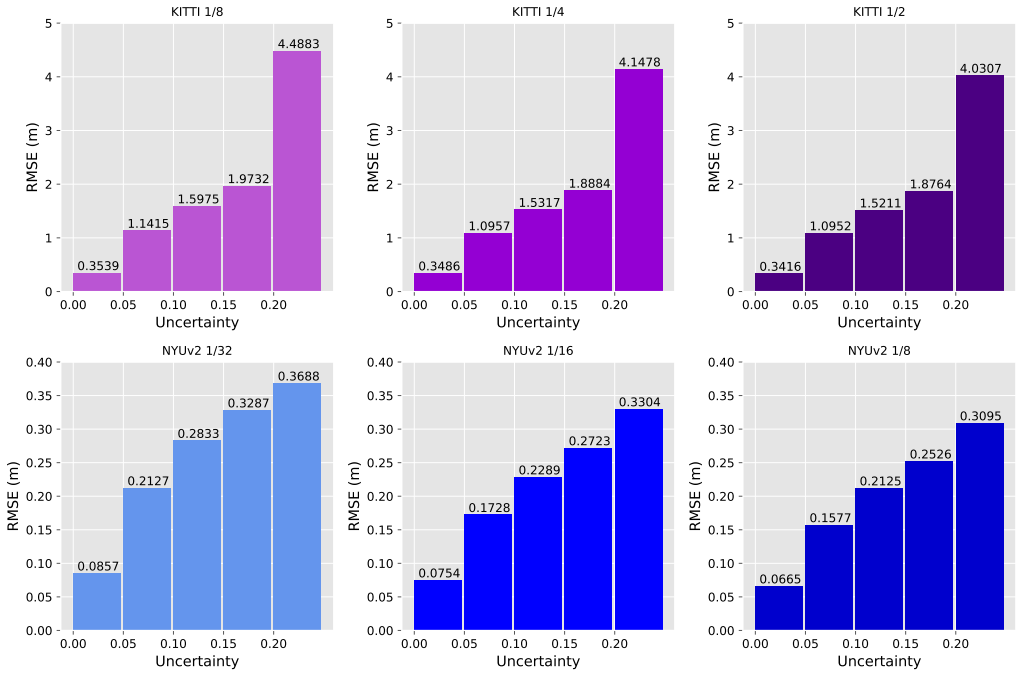

Figure 3: Chart of RMSE with different uncertainties. In all experiments, we observe that the pixels with higher uncertainty have larger errors. Specifically, the pixels with the lowest uncertainty ((0.00, 0.05]), have the smallest RMSE, while the pixels with the highest uncertainty ([0.20, $+\infty$)) have a very large average RMSE.

can lead the student model to converge to the teacher model (the baseline model), which can only bring slight performance improvement (4.1mm, 2.5mm, 2.4mm performance gain on 1/8, 1/4, and 1/2 settings). With the random-sampling depth-aware consistency, SEED can further improve the performance significantly and reduce the RMSE by 20.2mm, 6.1mm, and 8.1mm on 1/8, 1/4, and 1/2 settings respectively (Group III).

Effects of Uncertainty Rectification. We also 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 signifi-

cantly reduce from 859.9mm to 853.5mm, 850.2mm to 836.1mm, and 820.2mm to 810.8mm on 1/8, 1/4, and 1/2 settings respectively (Group IV). Besides, we also observe that the iterative self-training can further improve the performance slightly (Group V).

Quality of Uncertainty Estimation. To evaluate the quality of the estimated uncertainty, we calculate the RMSE with different uncertainties on both KITTI [4] and NYUv2 [11] datasets under different partition protocols. Specifically, we first sort all the pixels in the ascending order of uncertainty and divide the pixels into different groups according to the uncertainty values ([0.00, 0.05), [0.05, 0.10), [0.10, 0.15), [0.15, 0.20), [0.20, $+\infty$)). Then we calculate the average RMSE of each group and draw the corresponding chart. As shown in Fig. 3, we observe that the pixels with higher uncertainty have larger RMSE. The results verify that the proposed uncertainty estimation method is a robust way to measure the reliability of pseudo-labels. Furthermore, we also conduct experiments to compare different uncertainty estimation approaches [23, 26, 27] (see Tab. 3). We follow previous works [24, 27] to set the hyper-parameters for a fair comparison. For the Monte Carlo dropout sampling [23], we perform 8 forwards for the model at test time. In the snapshot ensemble experiments [26], we set the number of snapshots models to 8 and the number of cycles $C = 20$. To perform the bootstrapped ensemble [27], we randomly initialize 8 instances and train each instance on a randomly extracted 25% subsets of the entire training set separately. To estimate uncertainty by image flipping we calculate the difference between the original prediction and the horizontally flipped counterpart prediction. We observe that the Monte Carlo dropout sampling [23] slightly outperforms the baseline, while Snapshot [26], Bootstrap [30], and Flipping achieve better performance. In com-

Table 3: Comparison of different uncertainty estimation methods. We provide a comparison with other uncertainty estimation methods. The experimental results show that the proposed method achieves the best performance.

Method	RMSE_{KITTI} (mm)↓			RMSE_{NYUv2} (m) ↓		
	1/8	1/4	1/2	1/32	1/16	1/8
Baseline	884.2	858.8	830.7	0.118	0.112	0.104
Dropout [23]	881.3	864.5	830.7	0.121	0.114	0.107
Snapshot [26]	879.4	855.7	826.0	0.114	0.109	0.103
Bootstrap [27]	875.3	854.9	824.8	0.115	0.110	0.101
Flipping	856.6	843.3	815.8	0.113	0.107	0.101
Ours	**851.6**	**834.2**	**808.1**	**0.106**	**0.102**	**0.097**

Table 4: Performance with different baselines. We implement our approach with two commonly-used depth completion models: CSPN [6], CSPN++ [7], and NLSPN [20]. Our approach consistently improves the performance of three models under different partition protocols. The results verify that our algorithm can generalize well on different depth completion models.

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	CSPN++	889.2	859.9	844.6	0.1210	0.1137	0.1041
Ours	CSPN++	860.9	841.6	821.7	0.1101	0.1076	0.0998
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

parison, our method outperforms all these methods by a large margin. Besides, all other methods require multiple forwards or multiple model instances for uncertainty estimation. In contrast, our method models the uncertainty by leveraging the outputs of the refine process, which only forwards one time with one model.

Scalability to Different Basic Model Structures. To demonstrate the generalization of our algorithm, we also adopt other state-of-the-art models to conduct experiments on KITTI [4] and NYUv2 [11]. We implement our approach with three commonly-used depth completion models, *i.e.*, CSPN [6], CSPN++ [7] and NLSPN [20]. As shown in Tab 4, SEED consistently improves the performance on all

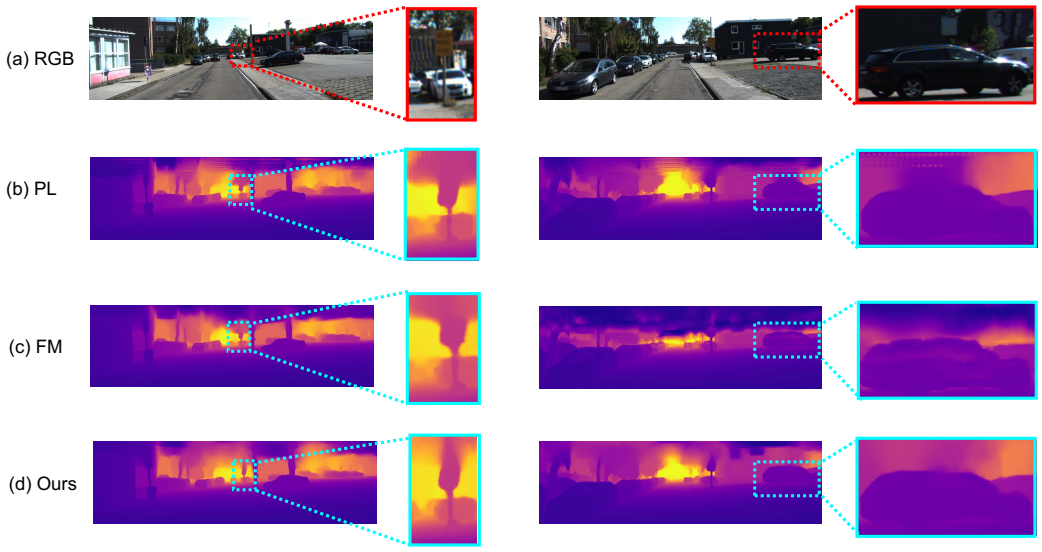

Figure 4: **Qualitative comparison on the KITTI validation dataset** [4]. (a) the color images. (b) Pseudo-Lebeling. (c) FixMatch [9], and (d) our method. As shown in the zoom regions, we observe the proposed method can generate better depth maps with detailed structure.

models for different split protocols. On the KITTI dataset [4], the NLSPN-based model [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 [11]. The results verify that SEED has strong generalization and is compatible with different depth completion models.

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

Comparison with Semi-supervised Methods. On the semi-supervised setting, we compare our method with some recent competitive semi-supervised methods, *i.e.*, Pseudo-labeling, Mean Teacher [31], Temporal Ensemble [30], and FixMatch [9]. We compare them using the same baseline architecture and partition protocols. Table 5 shows the comparison results on KITTI validation set [4] and

Table 5: Comparison with the semi-supervised state-of-the-art methods. "PL" means Pseudo-Labeling "MT" indicates Mean Teacher [31], and "TE" denotes Temporal Ensemble [30]. "FM" represents FixMatch [9].

Method	RMSE$_{KITTI}$ (mm)↓			RMSE$_{NYUv2}$ (m) ↓		
	1/8	1/4	1/2	1/32	1/16	1/8
PL	880.1	856.3	828.3	0.115	0.110	0.102
MT [31]	878.2	857.0	828.8	0.116	0.110	0.101
TE [30]	876.1	854.6	827.9	0.114	0.109	0.101
FM [9]	863.4	850.7	821.0	0.113	0.108	0.100
UDR [32]	878.1	856.3	838.6	0.117	0.115	0.108
CUPL [33]	875.1	870.6	858.0	0.116	0.114	0.104
UDL [34]	860.1	856.8	818.4	0.115	0.110	0.102
Ours	**851.6**	**834.2**	**808.1**	**0.106**	**0.102**	**0.097**

Table 6: The training time comparison with the semi-supervised state-of-the-art methods. "PL" means Pseudo-Labeling "MT" indicates Mean Teacher [31], and "TE" denotes Temporal Ensemble [30]. "FM" represents FixMatch [9].

Method	Time$_{KITTI}$ (hours)			Time$_{NYUv2}$ (hours)		
	1/8	1/4	1/2	1/32	1/16	1/8
PL	24.3	24.3	24.3	8.4	8.4	8.4
MT [31]	28.6	28.6	28.6	11.2	11.2	11.2
TE [30]	29.5	29.5	29.5	12.6	12.6	12.6
FM [9]	24.5	24.5	24.5	8.6	8.6	8.6
Ours	24.7	24.7	24.7	8.7	8.7	8.7

NYUv2 test set [11]. In the Pseudo-labeling experiment, we use the final output of the teacher model as the pseudo-depth labels to train the student model. For Mean Teacher [31], the weights of the teacher model are updated as an exponential moving average with EMA decay set to 0.999 during the training phase. In the Temporal Ensemble implementation [30], the ensembling momentum is set to 0.6. We perform the weak and strong augmentation methods as described in FixMatch [9]. The weak augmentation is a standard flip augmentation strategy that flips both RGB images and depth maps with a probability of 50%. For strong augmentation, we perform RandAugment for RGB images, and randomly inject

noises on the input sparse depth maps. In comparison with other semi-supervised methods, SEED achieves the best performance on all partition protocols. For example, our method achieves 851.6mm RMSE with 1/8 annotations on KITTI dataset [4], which outperforms Mean Teacher [31], Temporal Ensemble [30] and FixMatch [9] by 26.6mm, 24.5mm, and 11.8mm respectively. Compared to other methods, SEED can filter out unreliable predictions from the teacher model and focus the student model on reliable pseudo annotations, which verifies the proposed uncertainty rectification method can effectively tackle the problem of the noise of pseudo-labels. As shown in the right part of Tab. 5, SEED acquires remarkable performance on NYUv2 dataset [11]. We also visualize the recovered depth maps KITTI [4] dataset for qualitative comparison. As shown in the Fig. 4, we find that SEED can preserve the fine structure information near depth boundaries, which demonstrates the effectiveness of the proposed method. We further compare the training time in Tab. 6. We observe the proposed method costs less training time than Mean Teacher [31] and Temporal Ensemble [30], while achieving much better performance than FixMatch and Pseudo-labeling with comparable time cost.

Comparison with Fully-supervised Methods. To further verify the effectiveness of our approach, we also compare our method with fully-supervised state-of-the-art methods. Our goal is to narrow the gap between semi-supervised and fully-supervised learning methods with limited annotations. The detailed quantitative comparison results on the NYUv2 dataset [11] are illustrated in Tab. 7. With only 1/32 and 1/16 ground-truth annotations, SEED consistently performs better than most of the previous works and is even on par with the latest works. In particular, SEED yields close performance to the best performing fully-supervised model

Table 7: Comparisons with the **fully-supervised** state-of-the-art on the NYUv2 test dataset [11]. With only the 1/8 ground-truth annotations, our method yields close performance to the best-performing fully-supervised model.

Method	Labels	RMSE↓ m	REL↓ m	$\delta_{1.25}$ ↑	$\delta_{1.25^2}$ ↑	$\delta_{1.25^3}$ ↑
S2D [12]	47,584	0.123	0.026	99.1	99.9	100.0
DepthCoeff [35]	47,584	0.118	0.013	99.4	99.9	-
CSPN [6]	47,584	0.117	0.016	99.2	99.9	100.0
CSPN++ [7]	47,584	0.116	-	-	-	-
DeepLiDAR [29]	47,584	0.116	0.022	99.3	99.9	100.0
PRNet [36]	47,584	0.104	0.014	99.4	99.9	100.0
TWISE [37]	47,584	0.097	0.013	99.6	99.9	100.0
NLSPN [20]	47,584	0.092	0.012	99.6	99.9	100.0
Ours	**1,598**	0.106	0.015	99.5	99.9	100.0
Ours	**3,016**	0.102	0.014	99.5	99.9	100.0
Ours	**6,443**	0.097	0.013	99.5	99.9	100.0

with only 1/8 ground truth annotations. We also provide the comparison results of experiments on the KITTI benchmark [4]. From Tab. 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 SEED can significantly close the gap between semi-supervised and fully-supervised learning methods.

Table 8: Comparison with the **fully-supervised** state-of-the-art on the KITTI test dataset [4].

Method	Labels	RMSE↓ mm	MAE↓ mm	iRMSE↓ 1/km	iMAE↓ 1/km
CSPN [6]	85,898	1019.64	279.46	2.93	1.15
HMS [16]	85,898	841.78	253.47	2.73	1.13
TWISE [37]	85,898	840.20	195.58	2.08	0.82
DDP [38]	85,898	832.94	203.96	2.10	0.85
S2D [12]	85,898	814.73	249.95	2.80	1.21
3dDepthNet [39]	85,898	798.44	226.27	2.36	1.02
NLSPN [20]	85,898	741.68	199.59	1.99	0.84
Ours	**13,792**	816.75	217.17	2.12	0.91
Ours	**22,048**	794.01	213.52	2.10	0.90
Ours	**41,042**	778.96	211.41	2.08	0.89

Table 9: Comparison with the state-of-the-art semi-supervised methods. As shown in the table, SEED outperforms other methods by a large margin.

Method	RMSE (m) ↓	REL (m) ↓	$\delta_{1.25}$ ↑	$\delta_{1.25^2}$ ↑	$\delta_{1.25^3}$ ↑
Baseline	0.134	0.020	99.0	99.8	99.9
MT [31]	0.128	0.019	99.1	99.8	99.9
TE [30]	0.127	0.018	99.1	99.8	99.9
FM [9]	0.121	0.018	99.2	99.8	99.9
Ours	**0.110**	**0.016**	**99.4**	**99.9**	**100.0**

4.5. Experiments on Domain Adaptation

In this section, we extend our approach to a more challenging scenario: domain adaptation. Domain adaptation deals with scenarios where a model trained on a source distribution is used in a different but related target distribution. More specifically, domain adaptation uses labeled data in a source domain to solve new tasks in a target domain. A case in point is to exploit synthetic data, where the annotations are more accessible compared to the costly labeling of real-world images [40]. In the depth completion, the ground truth annotations are hard to access but easy for synthetic data in simulated environments. Therefore, we choose a synthetic dataset, *i.e.*, TartanAir [28], as the source dataset, and take the NYUv2 [11] as the target dataset. We observe that the depth distributions of different datasets can vary to a large extent, which is very challenging for the depth completion task. To bridge the large domain gap between the synthetic and real-world datasets, we first train the teacher model using the labeled data on the TartanAir [28] and then perform our algorithm on the NYUv2 dataset [11] (unlabeled data).

As shown in Tab. 9, we observe that without SEED, the baseline model only achieves the RMSE with 134.0mm. In contrast, our algorithm can improve the performance by **24mm** and outperforms all other semi-supervised methods [31, 30, 9]. Although a large domain gap exists between the source and target datasets,

Table 10: Comparisons with the state-of-the-art of on fully-supervised methods the NYUv2 test dataset. **Without any labeling** on NYUv2 (target dataset), SEED outperforms most fully-supervised approaches which require full real-data annotations.

Method	Label (NYUv2)	RMSE↓ m	REL↓ m	$\delta_{1.25} \uparrow$	$\delta_{1.25^2} \uparrow$	$\delta_{1.25^3} \uparrow$
S2D [12]	47,584	0.123	0.026	99.1	99.9	100.0
DepthCoeff [35]	47,584	0.118	0.013	99.4	99.9	-
CSPN [6]	47,584	0.117	0.016	99.2	99.9	100.0
CSPN++ [7]	47,584	0.116	-	-	-	-
DeepLiDAR [29]	47,584	0.116	0.022	99.3	99.9	100.0
NLSPN [20]	47,584	0.092	0.012	99.6	99.9	100.0
Ours	0	0.110	0.016	99.4	99.9	100.0

SEED can still efficiently exploit both the labeled and unlabeled data. From Tab. 10, we can see that SEED can even outperform many fully-supervised methods with only annotations on synthetic data. **It is worth noting that we do not use any ground-truth depth annotations on the target dataset (NYUv2).**

5. Conclusion

To explore the feasibility of leveraging unlabeled data, we introduce SEED, a semi-supervised learning algorithm to boost the performance of depth completion. By enforcing the density-aware consistency, we perform self-training on the unlabeled data while ensembling the reliable information of the student and teacher models. We further propose to exploit the uncertainty to resist the noisy pseudo-depth labels in the training process. Extensive experiments demonstrate that density-aware consistency and uncertainty-regularizing optimization can bring significant performance gain. We hope SEED can serve as a solid baseline and pave the way for future work on semi-supervised depth completion.

6. Limitations and Discussion

Although our method can leverage the unlabeled to boost the performance of depth completion, there is still room for improvement. One limitation is that it still requires a certain amount of labeled data to train the initial teacher model To address this issue, we could consider using some synthetic data to mimic other real LiDAR inputs to evolve the training quality for our teacher model pre-training. Moreover, we will explore incorporating extra supervision such as semantic masks, to extend to other real-world scenarios, *e.g.*, view synthesis and instance detection.

References

[1] F. Ma, G. V. Cavalheiro, S. Karaman, Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera, in: ICRA, 2019.

[2] X. Zhang, Z. Zheng, D. Gao, B. Zhang, P. Pan, Y. Yang, Multi-view consistent generative adversarial networks for 3d-aware image synthesis, in: CVPR, 2022.

[3] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, H. Zhao, Depth anything: Unleashing the power of large-scale unlabeled data, in: CVPR, 2024.

[4] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, A. Geiger, Sparsity invariant cnns, in: 3DV, 2017.

[5] K. Wang, L. Zhao, J. Zhang, J. Zhang, A. Wang, H. Bai, Joint depth

map super-resolution method via deep hybrid-cross guidance filter, Pattern Recognition 136 (2023) 109260.

[6] X. Cheng, P. Wang, R. Yang, Learning depth with convolutional spatial propagation network, IEEE transactions on pattern analysis and machine intelligence 42 (10) (2019) 2361–2379.

[7] X. Cheng, P. Wang, C. Guan, R. Yang, Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion, in: AAAI, 2020.

[8] D.-H. Lee, et al., Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, in: ICML Workshop, 2013.

[9] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, C. Raffel, Fixmatch: Simplifying semi-supervised learning with consistency and confidence, in: NeurIPS, 2020.

[10] Z. Zheng, Y. Yang, Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation, International Journal of Computer Vision 129 (4) (2021) 1106–1120.

[11] P. K. Nathan Silberman, Derek Hoiem, R. Fergus, Indoor segmentation and support inference from rgbd images, in: ECCV, 2012.

[12] F. Ma, S. Karaman, Sparse-to-dense: Depth prediction from sparse depth samples and a single image, in: ICRA, 2018.

[13] J. Deng, J. Zhang, Z. Hu, L. Wang, J. Jiang, X. Zhu, X. Chen, Y. Yuan,

C. Wang, Rgb-d salient object ranking based on depth stack and truth stack for complex indoor scenes, Pattern Recognition 137 (2023) 109251.

[14] R. Li, D. Xue, S. Su, X. He, Q. Mao, Y. Zhu, J. Sun, Y. Zhang, Learning depth via leveraging semantics: Self-supervised monocular depth estimation with both implicit and explicit semantic guidance, Pattern Recognition (2023) 109297doi:https://doi.org/10.1016/j.patcog.2022.109297.

[15] Z. Yu, Z. Sheng, Z. Zhou, L. Luo, S.-Y. Cao, H. Gu, H. Zhang, H.-L. Shen, Aggregating feature point cloud for depth completion, in: ICCV, 2023, pp. 8732–8743.

[16] Z. Huang, J. Fan, S. Cheng, S. Yi, X. Wang, H. Li, Hms-net: Hierarchical multi-scale sparsity-invariant network for sparse depth completion, IEEE Transactions on Image Processing (2019) 3429–3441.

[17] A. Eldesokey, M. Felsberg, K. Holmquist, M. Persson, Uncertainty-aware cnns for depth completion: Uncertainty from beginning to end, in: CVPR, 2020.

[18] J. Tang, F.-P. Tian, B. An, J. Li, P. Tan, Bilateral propagation network for depth completion, in: CVPR, 2024.

[19] Z. Yan, Y. Lin, K. Wang, Y. Zheng, Y. Wang, Z. Zhang, J. Li, J. Yang, Tri-perspective view decomposition for geometry-aware depth completion, in: CVPR, 2024.

[20] J. Park, K. Joo, Z. Hu, C.-K. Liu, I.-S. Kweon, Non-local spatial propagation network for depth completion, in: ECCV, 2020.

[21] Z. Chen, L. Zhu, L. Wan, S. Wang, W. Feng, P.-A. Heng, A multi-task mean teacher for semi-supervised shadow detection, in: CVPR, 2020, pp. 5611–5620.

[22] Y. Kuznietsov, J. Stuckler, B. Leibe, Semi-supervised deep learning for monocular depth map prediction, in: CVPR, 2017.

[23] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, in: NeurIPS, 2017.

[24] M. Poggi, F. Aleotti, F. Tosi, S. Mattoccia, On the uncertainty of self-supervised monocular depth estimation, in: CVPR, 2020.

[25] S. Liu, S. De Mello, J. Gu, G. Zhong, M.-H. Yang, J. Kautz, Learning affinity via spatial propagation networks, in: NeurIPS, 2017.

[26] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, K. Q. Weinberger, Snapshot ensembles: Train 1, get m for free, in: ICLR, 2017.

[27] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in: NeurIPS, 2017.

[28] W. Wang, D. Zhu, X. Wang, Y. Hu, Y. Qiu, C. Wang, Y. Hu, A. Kapoor, S. Scherer, Tartanair: A dataset to push the limits of visual slam, in: IROS, 2020.

[29] J. Qiu, Z. Cui, Y. Zhang, X. Zhang, S. Liu, B. Zeng, M. Pollefeys, Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image, in: CVPR, 2019.

[30] S. Laine, T. Aila, Temporal ensembling for semi-supervised learning, in: ICLR, 2017.

[31] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: NeurIPS, 2017.

[32] C. Zhang, G. Li, Y. Qi, S. Wang, L. Qing, Q. Huang, M.-H. Yang, Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection, in: CVPR, 2023, pp. 16271–16280.

[33] S. Chen, T. Ye, J. Bai, E. Chen, J. Shi, L. Zhu, Sparse sampling transformer with uncertainty-driven ranking for unified removal of raindrops and rain streaks, in: ICCV, 2023, pp. 13106–13117.

[34] Q. Ning, W. Dong, X. Li, J. Wu, G. Shi, Uncertainty-driven loss for single image super-resolution, NeurIPS (2021) 16398–16409.

[35] S. Imran, Y. Long, X. Liu, D. Morris, Depth coefficients for depth completion, in: CVPR, 2019.

[36] B.-U. Lee, K. Lee, I. S. Kweon, Depth completion using plane-residual representation, in: CVPR, 2021.

[37] S. Imran, X. Liu, D. Morris, Depth completion with twin surface extrapolation at occlusion boundaries, in: CVPR, 2021.

[38] Y. Yang, A. Wong, S. Soatto, Dense depth posterior (ddp) from single image and sparse range, in: CVPR, 2019.

[39] R. Xiang, F. Zheng, H. Su, Z. Zhang, 3ddepthnet: Point cloud guided depth completion network for sparse depth and single color image, in: CVPR, 2020.

[40] N. Araslanov, S. Roth, Self-supervised augmentation consistency for adapting semantic segmentation, in: CVPR, 2021.