Sparse Pseudo-LiDAR Depth Assisted Monocular Depth Estimation

Shuwei Shao, Zhongcai Pei, Weihai Chen*, Qiang Liu, Haosong Yue and Zhengguo Li, *Fellow, IEEE*

***Abstract*—Monocular depth estimation has attracted extensive attention and made great progress in recent years. However, the performance still lags far behind LiDAR-based depth completion algorithms. This is because the completion algorithms not only utilize the RGB image, but also have the prior of sparse depth collected by LiDAR. To reduce this performance gap, we propose a novel initiative that incorporates the concept of pseudo-LiDAR into depth estimation. The pseudo-LiDAR depends only on the camera and thus achieves a lower cost than LiDAR. To emulate the scan pattern of LiDAR, geometric sampling and appearance sampling are proposed. The former measures the vertical and horizontal azimuths of 3D scene points to establish the geometric correlation. The latter helps determine which "pseudo-LiDAR rays" return an answer and which do not. Then, we build a sparse pseudo-LiDAR-based depth estimation framework. Extensive experiments show that the proposed method surpasses previous state-of-the-art competitors on the KITTI, NYU-Depth-v2 and SUN RGB-D datasets. The source code and trained models will be publicly available upon acceptance.**

***Index Terms*—Monocular Depth Estimation, LiDAR, Pseudo-LiDAR**

I. Introduction

Monocular depth estimation is a classical research topic in computer vision and is critical for a wide range of applications, for example, autonomous driving, 3D reconstruction and scene understanding [1]–[4]. Given that there are an infinite number of 3D scenes that can be projected onto the same 2D scene, such a task is ill-posed and inherently ambiguous. Because of the explosion of deep learning, significant advances have been made in this field [5]–[7]. Nevertheless, the performance is still far behind the counterparts, namely, LiDAR-based depth completion algorithms [8]–[11].

The depth completion algorithms take as inputs the RGB image and the sparse depth map captured by LiDAR, which provides an accurate sparse depth prior of the surrounding environment. Although highly precise, LiDAR is very expensive. Motivated by the success of utilizing pseudo-LiDAR in other relevant tasks [12]–[14], we propose a pseudo-LiDAR based framework to recover a high-quality depth map. The pseudo-LiDAR depends only on the camera and thus costs less than LiDAR. Because the original input to our framework is still a single image, it belongs to the monocular depth estimation. An illustration of our depth prediction, sparse depth maps from the proposed pseudo-LiDAR and LiDAR is shown in Fig. 1.

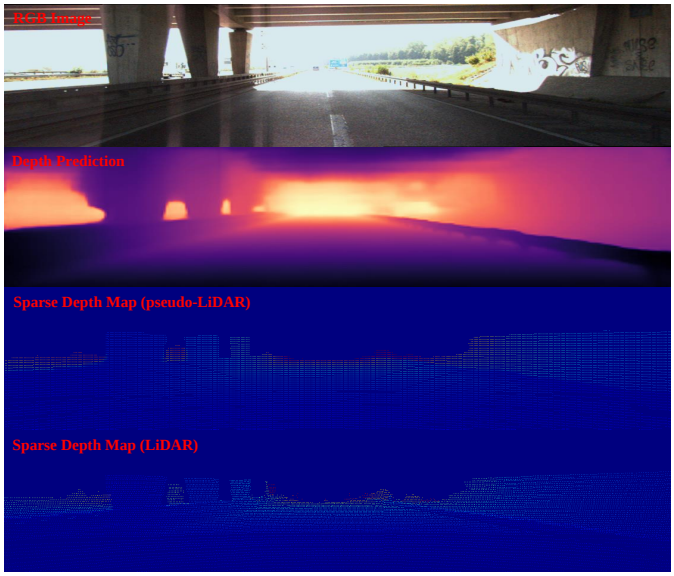

Fig. 1. **Illustration of the depth prediction and sparse depth maps from the proposed pseudo-LiDAR and LiDAR**.

Previous pseudo-LiDAR-based methods, such as [13], use dense depth maps or point clouds, which we refer to as dense pseudo-LiDAR. Based on our observations and experiments, we argue that sparse pseudo-LiDAR may be a better choice for depth estimation, since not only are the depth values vital, but their distribution patterns also matter. We binarize the sparse depth map collected by LiDAR and use the resultant binary mask to sparsify the dense depth map. In this case, the sparsified depth map only has depth values at partial pixels with specific distribution patterns and a pixel with depth has the same value as a pixel of the dense depth map at the same location, but performance is boosted significantly. An intuitive comparison is presented in Fig.2. It is not surprising given that the sparse depth map from LiDAR contains massive amounts of valuable information apart from the depth values. For example, geometric correlation, pixels with depth values near the same arc have similar vertical azimuths in the 3D space, and semantic relation, pixels in highly reflective/transparent regions or the sky tend to miss depth values. However, the scan pattern of LiDAR is not available at test time.

To address this problem, we introduce the geometric sam-

This work was supported by the National Natural Science Foundation of China under grant 61620106012 and in part by the Agency for Science, Technology and Research of Singapore under the Robotics Horizontal Technology Coordinating Office Project under Grant C221518005.

Shuwei Shao, Zhongcai Pei, Weihai Chen, Qiang Liu and Haosong Yue are with the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. (email: swshao@buaa.edu.cn, peizc@buaa.edu.cn, whchen@buaa.edu.cn, 476582539@qq.com and yuehaosong@buaa.edu.cn)

Zhengguo Li is with the SRO department, Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way, Singapore. (ezgli@i2r.a-star.edu.sg)

*(corresponding author: Weihai Chen.)

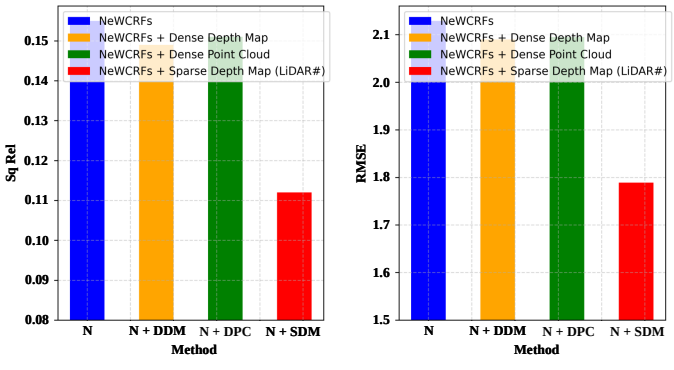

Fig. 2. **Intuitive comparison on the Eigen split of the KITTI dataset (cap: 0-80m). N: NeWCRFs [7]; DDM: dense depth map; DPC: dense point cloud; SDM: sparse depth map. (LiDAR#) means that we binarize the sparse depth map collected by LiDAR, where 1 stands for where there is a depth value and 0 stands for where there is no depth value, and leverage the resultant binary mask to sparsify the dense depth map predicted from NeWCRFs by replacing the Swin-Large backbone [15] with Swin-Tiny. In this case, the difference between sparse and dense depth maps is only the number of pixels with depth values. For "N + DDM" and "N + SDM", the input is the RGB image and depth map. For "N + DPC", the input is the RGB image and dense point cloud converted from the dense depth map.**

pling and appearance sampling to emulate the scan pattern of LiDAR. The geometric sampling[1] measures the vertical and horizontal azimuths of every back-projected 3D scene point and samples points in light of the azimuths. This enables the geometric correlation to be established among sampled points. The appearance sampling effectively says 'for a particular view, what is the likelihood of a "pseudo-LiDAR ray" in this direction returning an answer?', owing to absorption, specular reflection, or diffusion, which is achieved in a learning-based manner. The input is the RGB image and dense depth map. The ground-truth is a soft map acquired by binarizing the sparse depth map from LiDAR and then blurring it with a simple Gaussian filter. In contrast to the binary mask, the denser soft map makes the learning of appearance sampling easier.

The full pipeline of our sparse pseudo-LiDAR-based depth estimation framework is demonstrated in Fig. 3. To summarize, the contributions of this work are as follows:

- We introduce a novel initiative for monocular depth estimation that integrates the strengths of depth completion algorithms utilizing a pseudo-LiDAR, which can provide a critical depth prior of the surrounding environment at a low cost.
- We additionally find that the distribution pattern of depth values has a heavy impact on performance, and propose a new sparse pseudo-LiDAR by simulating the scan pattern of LiDAR with the developed geometric sampling and appearance sampling.
- Extensive experiments indicate the efficacy of our framework with state-of-the-art performance on the KITTI [16], NYU-Depth-v2 [17] and SUN RGB-D [18] datasets [2].

[1]We assume that the camera coordinate system coincides with the LiDAR coordinate system.

[2]While the NYU-Depth-v2 and SUN RGB-D are not collected by LiDAR, our sparse pseudo-LiDAR can be applied to these two datasets also.

II. Related Work

A. Monocular Depth Estimation

Monocular depth estimation aims to predict the depth map from a single image. Saxena *et al.* [19] introduced one of the first learning-based studies, where the Markov random field is used to regress the depth map. Following this, Eigen *et al.* [5] proposed combining global coarse-scale and local fine-scale convolutional neural networks (CNNs). Since then, numerous follow-up studies have been proposed. Laina *et al.* [20] leveraged fully convolutional residual networks and a reverse Huber loss. Cao *et al.* [21] and Fu *et al.* [22] reformulated depth estimation from the perspective of classification, predicting the optimal depth interval rather than the exact depth. Qi *et al.* [23] built a geometric neural network to jointly estimate the depth map and surface normal. Lee *et al.* [6] introduced multi-scale planar layers to provide effective guidance of encoded features for the desired depth prediction. Yin *et al.* [24] suggested the concept of a virtual normal, which allowed robust imposing of a geometric constraint on the depth map. Yang *et al.* [25] made use of the Vision Transformer (ViT) [26] as the encoder. Bhat *et al.* [27] generalized the ordinal regression network [22] by acquiring adaptive bins according to the image content. Patil *et al.* [28] proposed a piecewise planarity prior to leverage information from coplanar pixels. Shao *et al.* [29] used the ensemble strategy to integrate strengths from both CNN and Transformer. Yuan *et al.* [7] developed neural conditional random field (CRF) modules to capture long-range correlation in the decoder. Li *et al.* [31], Jiao *et al.* [32], Qiao *et al.* [33], Zhang *et al.* [34] and Zhang *et al.* [35] further improved the performance by incorporating auxiliary scene class or semantic information. In contrast to these studies, we develop a sparse pseudo-LiDAR-based framework to obtain a high-quality depth map.

B. Pseudo-LiDAR

Recently, pseudo-LiDAR has been widely adopted in 3D object detection. These studies leveraged a depth estimation network to estimate dense depth maps or point clouds, which were utilized to reduce the performance gap between image-based and LiDAR-based methods [12]–[14]. To be specific, Weng *et al.* [36] performed monocular depth estimation and transformed the input image into a pseudo-LiDAR point cloud for detection. Park *et al.* [37] developed an end-to-end framework to better utilize the strengths of pseudo-LiDAR. Li *et al.* [38] proposed utilizing a binary neural network to achieve timely depth prediction in pseudo-LiDAR. Simonelli *et al.* [39] equipped the pseudo-LiDAR-based methods with a confidence prediction module to boost the performance. Obviously, the pseudo-LiDAR has a lower cost than LiDAR because there is a large cost difference between camera and LiDAR. In other fields too, pseudo-LiDAR has gradually received increasing attention. Wang *et al.* [40] utilized pseudo-LiDAR to improve self-supervision in scene flow estimation. Deng *et al.* [41] used pseudo-LiDAR to establish the spatial connection of the points for visual odometry recovery.

Taking inspiration from these relevant studies, a novel initiative that incorporates pseudo-LiDAR into monocular depth

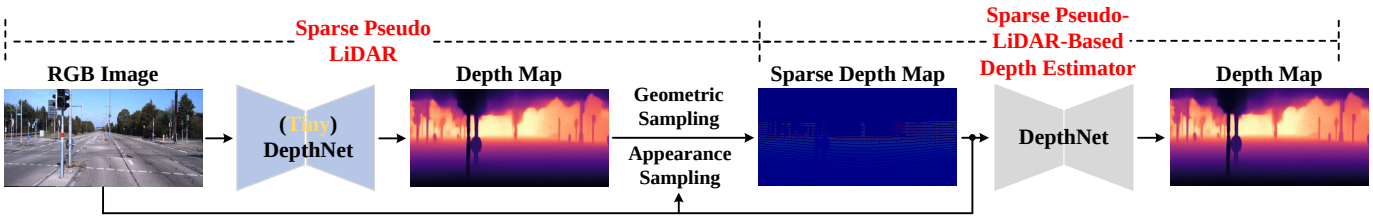

Fig. 3. **Overview of the proposed sparse pseudo-LiDAR-based depth estimation framework**. The framework contains two parts: sparse pseudo-LiDAR generation and sparse pseudo-LiDAR-based depth estimation. Before feeding into the second part, we concatenate the RGB image and the sparse depth map along the channel dimension and use a 7×7 convolution kernel to align the number of channels to 3. Details about the structure of both DepthNets can be found in the methodology section.

estimation is introduced in this work. Moreover, the signal of our pseudo-LiDAR is sparse by leveraging geometric sampling and appearance sampling to sparsify the dense pseudo-LiDAR.

C. Spatial Sampling Strategy

The spatial sampling of an unknown function has been extensively studied in widespread fields, which include spatial statistics [42], computer graphics [43], image processing [44], and remote sensing [45]. Irregular sampling strategies, *e.g.*, keypoint-based sampling [46] and Poisson disk sampling [47], are prone to demonstrate better spatial coverage and uniformity compared to regular sampling on an integral lattice. Nevertheless, implementing the Poisson disk sampling practically can be challenging because of the high computational complexity. Caflisch [48] further developed Monte Carlo methods to refine the Poisson disk sampling by utilizing quasi-random sequences and sampling patterns. These quasi-random sequences, also known as low-discrepancy sequences [49], are deterministically constructed to ensure a uniform coverage of the sampling space. In contrast, we introduce the geometric sampling and appearance sampling to emulate the scan pattern of LiDAR.

III. Methodology

In this section, we elaborate on our sparse pseudo-LiDAR-based depth estimation framework. The whole pipeline consists of two parts: sparse pseudo-LiDAR generation and sparse pseudo-LiDAR-based depth estimation.

A. Problem Formulation

Given a training set $D = \{(\mathbf{r}, \mathbf{d})\}_1^N$ containing N RGB image and depth ground-truth pairs, prior supervised methods train a depth network to map an RGB image $\mathbf{r}$ into a depth map $\hat{\mathbf{d}}$,

$$D_\theta : \mathbf{r} \rightarrow \hat{\mathbf{d}}, \tag{1}$$

where D_θ denotes the depth network. Unlike previous studies, our paradigm leverages the sparse depth map $\overline{\mathbf{d}}$ from pseudo-LiDAR and RGB image to achieve a high-quality depth map,

$$D_\theta : \mathbf{r}, \overline{\mathbf{d}} \rightarrow \hat{\mathbf{d}}. \tag{2}$$

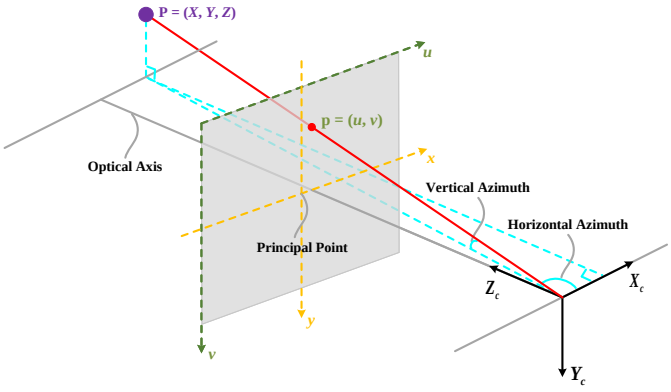

Fig. 4. **Illustration of the vertical and horizontal azimuths**.

B. Sparse Pseudo-LiDAR Generation

To generate the pseudo-LiDAR, we adopt a DepthNet modified from NeWCRFs [7] to predict dense depth maps, where we replace the Swin-Large [15] with Swin-Tiny, and introduce geometric sampling and appearance sampling to simulate the LiDAR scan pattern. The details are presented below.

Geometric sampling. We assume that the camera coordinate system coincides with the LiDAR coordinate system. Consider a dense depth map $\hat{\mathbf{d}}$, where a pixel is defined as $\mathbf{p} = (u, v)$. For each $\mathbf{p}$, the corresponding 3D scene point in the camera coordinate system can be back-projected as

$$\mathbf{P} = \hat{\mathbf{d}}\,(\mathbf{p})\,\mathbf{K}^{-1} h\,(\mathbf{p}) \tag{3}$$

with

$$\mathbf{K} = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix}, \tag{4}$$

where $\mathbf{P} = (X, Y, Z)$, $h\,(\cdot)$ denotes the homogeneous coordinate, f_x and f_y stand for the focal lengths along the x and y axes, respectively and (c_x, c_y) denotes the principal point. We calculate the vertical azimuth (Fig. 4) of $\mathbf{P}$, which is the angle between the straight line connecting the camera center to $\mathbf{P}$ and the plane $\boldsymbol{X}_c\boldsymbol{Z}_c$

$$\alpha_v = \arctan\left(\frac{Y}{\sqrt{X^2 + Z^2}}\right). \tag{5}$$

Suppose that the vertical field of view and the vertical azimuth resolution of pseudo-LiDAR are $[\alpha_v^{\min}, \alpha_v^{\max}]$ and σ_v, respec-

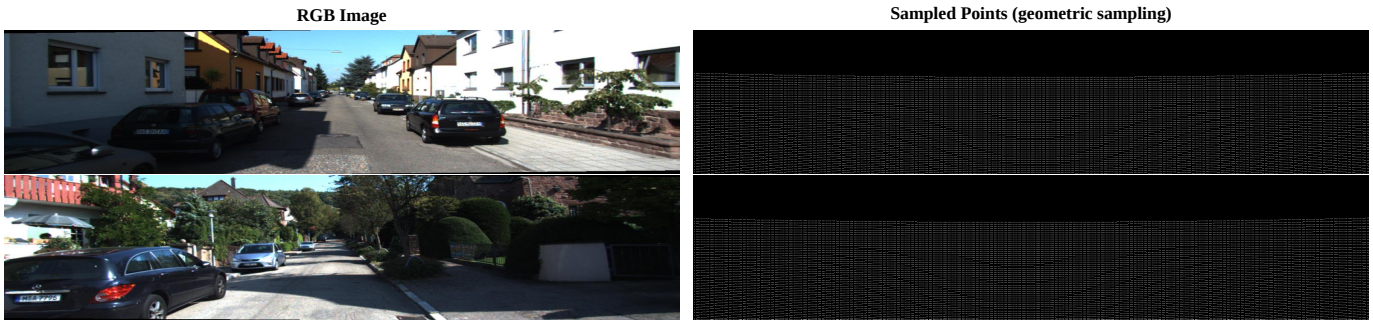

Fig. 5. **Illustration of the sampled points for geometric sampling.**

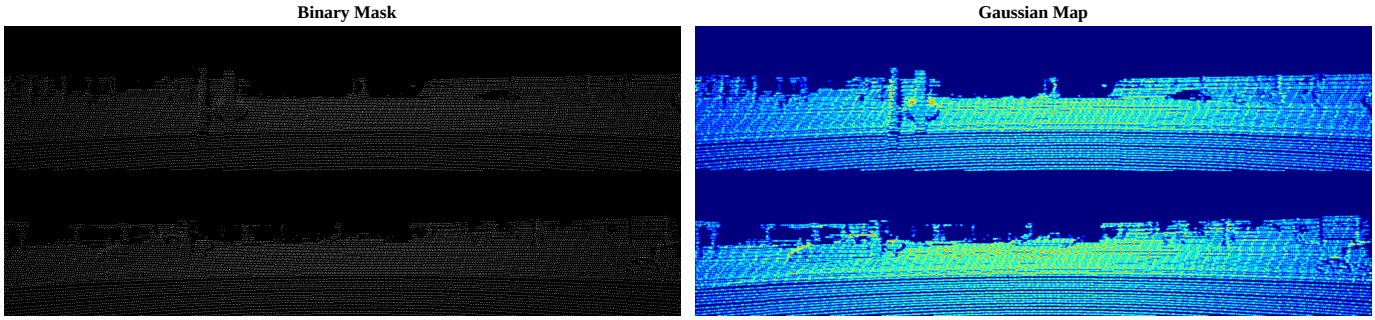

Fig. 6. **Illustration of the binary masks and the corresponding Gaussian maps**.

tively. Then, the vertical azimuths of the "pseudo-LiDAR rays" are calculated as

$$\alpha_v^{rays} = \left\{\alpha_v^{ray\_0}, \ldots, \alpha_v^{ray\_i}, ...\right\}, \tag{6}$$

with

$$\alpha_v^{ray\_i} = \alpha_v^{\min} + i\left(\alpha_v^{\max} - \alpha_v^{\min}\right)/\sigma_v. \tag{7}$$

To acquire the 3D scene points captured by each ray, the L1 distance between α_v and $\alpha_v^{ray\_i}$ is measured

$$dist = \left|\alpha_v - \alpha_v^{ray\_i}\right|, \tag{8}$$

and we treat points with the smallest distances along the vertical direction as the capture points, defined as $\mathbf{p}_v$.

Although the horizontal field of view of LiDAR is 360°, its scan pattern in the horizontal direction is not continuous. To sample points along the horizontal direction, we calculate the horizontal azimuth (Fig. 4) of $\mathbf{P}$, which denotes the angle between the projection of the straight line from the camera center to $\mathbf{P}$ on the plane $\boldsymbol{X}_c\boldsymbol{Z}_c$ and the $\boldsymbol{X}_c$ axis

$$\alpha_u = \arctan\left(\frac{Z}{X}\right). \tag{9}$$

From α_u, we can acquire the maximum and minimum horizontal azimuths of 3D scene points, defined as $\alpha_u^{\max}$ and $\alpha_u^{\min}$, respectively. Suppose that the horizontal azimuth resolution of pseudo-LiDAR is σ_u. Then, the number of horizontal azimuths to use is obtained by

$$num = \left(\alpha_u^{\max} - \alpha_u^{\min}\right)/\sigma_u. \tag{10}$$

We compute the indices of these horizontal azimuths

$$idxs = \{idx\_0, ..., idx\_j, ...\} \tag{11}$$

with

$$idx\_j = j\left(W - 1\right)/num, \tag{12}$$

where W is the width of dense depth map. According to the indices, we can acquire the corresponding horizontal azimuths and points, defined as $\mathbf{p}_u$. The intersection of $\mathbf{p}_v$ and $\mathbf{p}_u$ is taken as the set of sampling points for geometric sampling, which are illustrated in Fig. 5.

Appearance sampling. During the process of scanning, LiDAR may fail to receive the return signal owing to absorption, specular reflection or diffusion which can result in missing depth values. Nevertheless, this allows the semantic correlation to be included in its scan pattern, as these regions are usually from highly reflective/transparent objects or sky. The appearance sampling helps determine which "pseudo-LiDAR rays" return an answer and which do not and is achieved in a learning-based manner.

We binarize the sparse depth map from LiDAR, where 1 indicates where there is a depth value and 0 indicates no depth value, and the resultant binary mask $\mathbf{B}$ indicates which regions are visible to LiDAR, but also raises the problem of how to predict them. For example, due to the quite low resolution of LiDAR, the Velodyne LiDAR employed to collect the KITTI dataset [16] has only 64 beams. This results in the fact that only some of the pixels with a value of 0 in the binary mask are from regions hard to return signals.

To mitigate the problem, we densify the binary mask to a soft map using a simple Gaussian filter $\mathcal{G}$, which we refer to as the Gaussian map $\mathbf{G}$, as shown in Fig. 6,

$$\mathbf{G} = \mathcal{G} * \mathbf{B}, \tag{13}$$

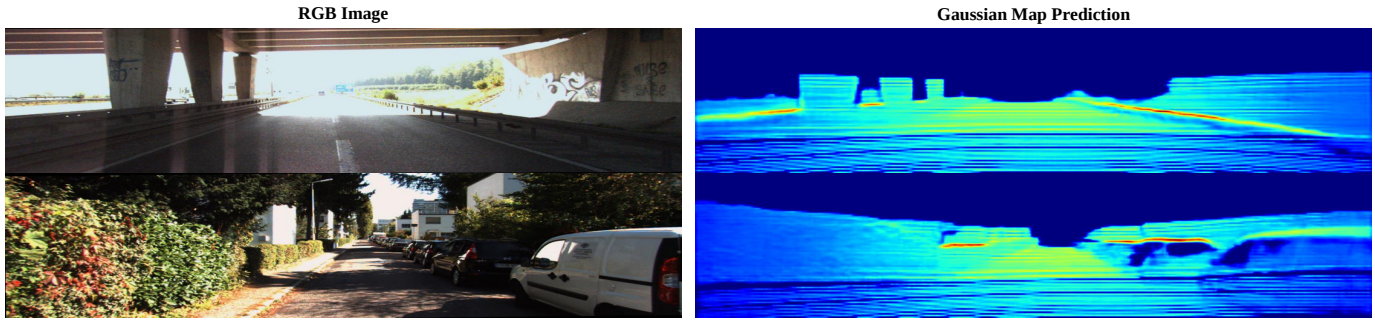

Fig. 7. **Illustration of the predicted Gaussian maps**. The car window and sky tend to be given low values.

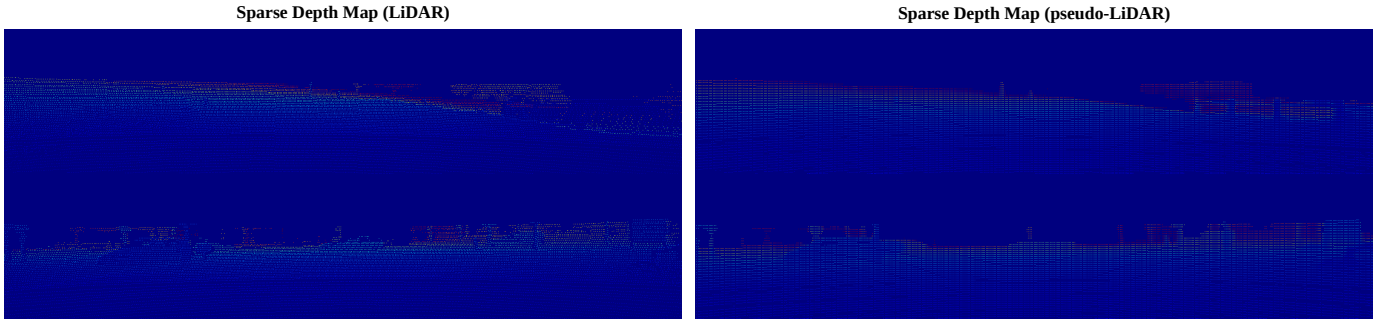

Fig. 8. **Illustration of the sparse depth maps collected by the proposed pseudo-LiDAR and LiDAR on the KITTI dataset**.

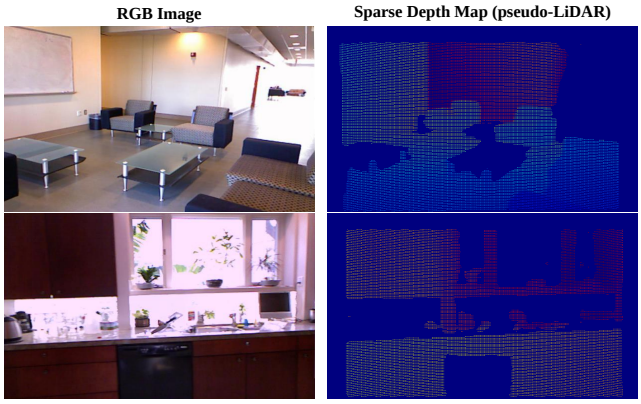

Fig. 9. **Illustration of the sparse depth maps collected by the proposed pseudo-LiDAR on the NYU-Depth-v2 dataset**.

where $*$ is the convolution operation. In contrast to the sparse binary mask, the denser Gaussian map enjoys two strengths: (i) it better abstracts the visible property of LiDAR by filling the gaps between neighboring sample points but does not wrongly fill up invisible regions, and (ii) it is easier to be predicted since it contains continuous values as opposed to 0 or 1 in the binary mask.

A network is used to predict the Gaussian map, whose input is the RGB image and dense depth map. We apply $\mathcal{L}_G$ to enforce the Gaussian map prediction $\hat{\mathbf{G}}$ to approximate $\mathbf{G}$,

$$\mathcal{L}_G = \sum_{\mathbf{p}} \left| \mathbf{G}\left(\mathbf{p}\right) - \hat{\mathbf{G}}\left(\mathbf{p}\right) \right|. \tag{14}$$

Based on $\hat{\mathbf{G}}$, we generate a binary mask with a threshold t

$$\hat{\mathbf{B}} = \hat{\mathbf{G}} > t. \tag{15}$$

The points with a value of 1 in $\hat{\mathbf{B}}$ are treated as the sampling points for appearance sampling. As demonstrated in Fig. 7, we visualize the Gaussian map predictions. Highly reflective or transparent regions, such as car windows and sky, tend to be given low values. In addition, the boundaries of some objects are clearly highlighted.

We make use of the coincident sampling points of geometric sampling and appearance sampling to sparsify the dense depth map, and the resultant sparse depth map is used as the output of our pseudo-LiDAR, as presented in Figs. 8 and 9. It should be noted that while sparse pseudo-LiDAR generation requires depth ground-truth during the training phase, it does not need depth ground-truth once the training is completed. This allows it to be applied to unseen scenarios.

C. Sparse Pseudo-LiDAR-Based Depth Estimation

Different from previous learning paradigms, we utilize the sparse depth map collected by our sparse pseudo-LiDAR and RGB image to infer depth. Because the original input to the pipeline is a single RGB image, our paradigm belongs to the monocular depth estimation. Here, we simply concatenate the sparse depth map and RGB image along the channel dimension, while using two separate branches for sparse depth map and RGB image feature extraction as some depth completion algorithms such as [9] may yield higher accuracy. Then, we utilize a single 7×7 convolution kernel to align the number of channels to 3 and feed it to the depth network. To supervise

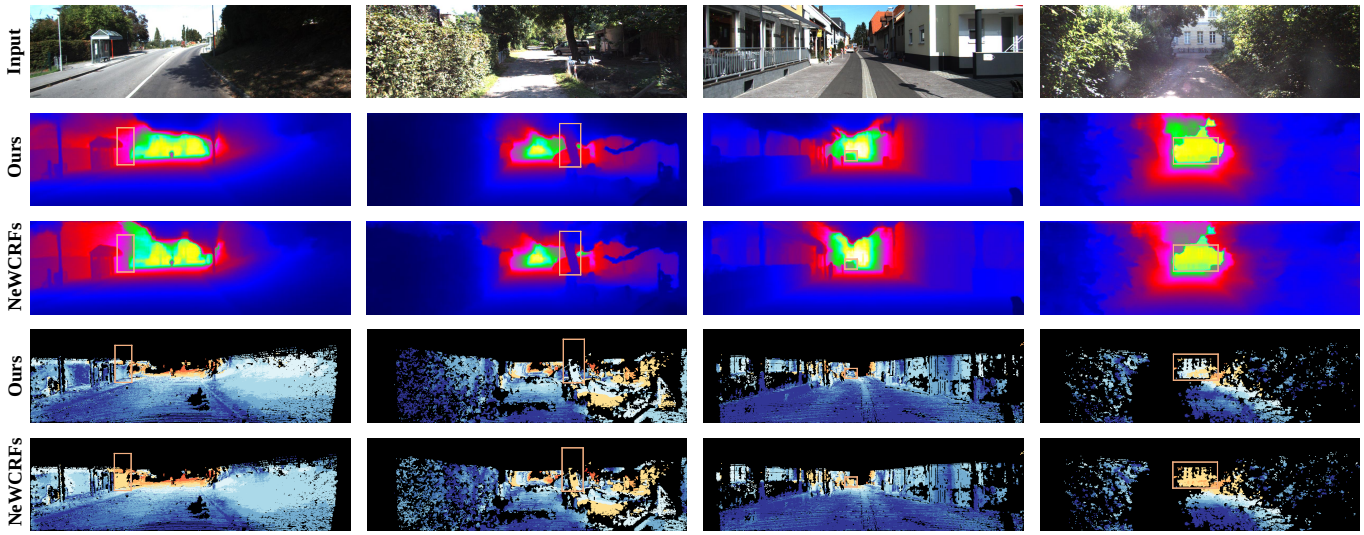

Fig. 10. **Qualitative results on the KITTI online benchmark**. The boxes in orange indicate the focused regions. The second and third rows show depth maps, and the fourth and fifth rows show error maps, where the large errors are in orange or red.

the depth prediction, we use a scaled scale-invariant (SSI) loss introduced by Lee *et al.* [6],

$$\mathcal{L}_{SSI} = \kappa \sqrt{\frac{1}{|\mathrm{T}|}\sum_{\mathbf{p}} \left(\mathrm{g}\left(\mathbf{p}\right)\right)^2 - \frac{\eta}{|\mathrm{T}|^2}\left(\sum_{\mathbf{p}} \mathrm{g}\left(\mathbf{p}\right)\right)^2}, \quad (16)$$

where $\mathbf{g}\left(\mathbf{p}\right) = \log \hat{\mathbf{d}}\left(\mathbf{p}\right) - \log \mathbf{d}\left(\mathbf{p}\right)$, $\mathbf{d}$ represents the depth ground-truth, and $\mathbf{T}$ is a set of pixels with valid ground-truth values. κ and η are set to 10 and 0.85 following [6].

D. Network Architecture.

The whole pipeline consists of two depth networks based on an encoder-decoder structure with skip connections. The depth network for sparse pseudo-LiDAR build adopts a lightweight network for practical applications, with Swin-Tiny [15] as the encoder. The depth network for depth estimation uses Swin-Large [15] as the encoder. The neural window full-connected CRFs [7] constitute the decoder of both depth networks, which are capable of capturing significant long-range correlation. The network to predict the Gaussian map follows a design similar to the first depth network. In the encoding stage, we utilize Swin-Tiny. For the decoding stage, we utilize the decoder of the depth network with a small modification. In the final prediction layer, only the sigmoid activation function is used to constrain the output because the Gaussian map values are between 0 and 1.

IV. Experiments

We conduct extensive experiments on multiple challenging datasets for outdoor and indoor scenarios, including KITTI [16], NYU-Depth-v2 [17] and SUN RGB-D [18]. Note that although the NYU-Depth-v2 and SUN RGB-D are not collected by LiDAR, the proposed sparse pseudo-LiDAR can also be applied to these two datasets. The acquisition sensors for these two datasets, for example, Kinect, also suffer from highly reflective/transparent objects such as mirrors, therefore losing depth values in these regions.

A. Datasets and Evaluation Metrics

KITTI dataset is collected from outdoor scenarios utilizing equipment such as camera and LiDAR mounted on a moving vehicle, along with the stereo images and corresponding 3D laser scans. The image resolution is around 1241×376 pixels. We employ two widely used data splits, Eigen split [5] and official split. The former consists of 23488 training image pairs and 697 test images, and the latter consists of 42949 training pairs, 1000 validation pairs and 500 test images. Note that the ground-truth depth maps of the test images are not available for the official split, and the test results are generated by an online server.

NYU-Depth-v2 dataset provides indoor images and depth maps at a resolution of 640×480 pixels. Following previous studies, we employ the official split to validate the proposed method. The split uses 249 scenes for training and 654 images from 215 scenes for testing.

SUN RGB-D dataset is captured from indoor scenes with four sensors and broad scene diversity, composed of approximately 10K images. We only use this dataset for the generalization study and do not use it for training. We evaluate pre-trained models on the official test set of 5050 images.

Evaluation metrics. We adopt the evaluation metrics as in previous studies [31]:

- SILog: $\sqrt{\frac{1}{|\mathbf{T}|}\sum_{\mathbf{g}\in\mathbf{T}}\left(\mathbf{g}\right)^2 - \frac{1}{|\mathbf{T}|^2}\left(\sum_{\mathbf{g}\in\mathbf{T}}\mathbf{g}\right)^2}$;
- Abs Rel: $\frac{1}{|\mathbf{T}|}\sum_{\hat{\mathbf{d}}\in\mathbf{T}}\left|\hat{\mathbf{d}}-\mathbf{d}\right|/\mathbf{d}$;
- Sq Rel (eigen split): $\frac{1}{|\mathbf{T}|}\sum_{\hat{\mathbf{d}}\in\mathbf{T}}\left\|\hat{\mathbf{d}}-\mathbf{d}\right\|^2/\mathbf{d}$;
- Sq Rel (official split): $\frac{1}{|\mathbf{T}|}\sum_{\hat{\mathbf{d}}\in\mathbf{T}}\left\|\hat{\mathbf{d}}-\mathbf{d}\right\|^2/\mathbf{d}^2$;
- RMSE: $\sqrt{\frac{1}{|\mathbf{T}|}\sum_{\hat{\mathbf{d}}\in\mathbf{T}}\left\|\hat{\mathbf{d}}-\mathbf{d}\right\|^2}$;
- RMSE log: $\sqrt{\frac{1}{|\mathbf{T}|}\sum_{\hat{\mathbf{d}}\in\mathbf{T}}\left\|\log\hat{\mathbf{d}}-\log\mathbf{d}\right\|^2}$;
- iRMSE: $\sqrt{\frac{1}{|\mathbf{T}|}\sum_{\hat{\mathbf{d}}\in\mathbf{T}}\left\|1/\hat{\mathbf{d}}-\mathbf{1}/\mathbf{d}\right\|^2}$;

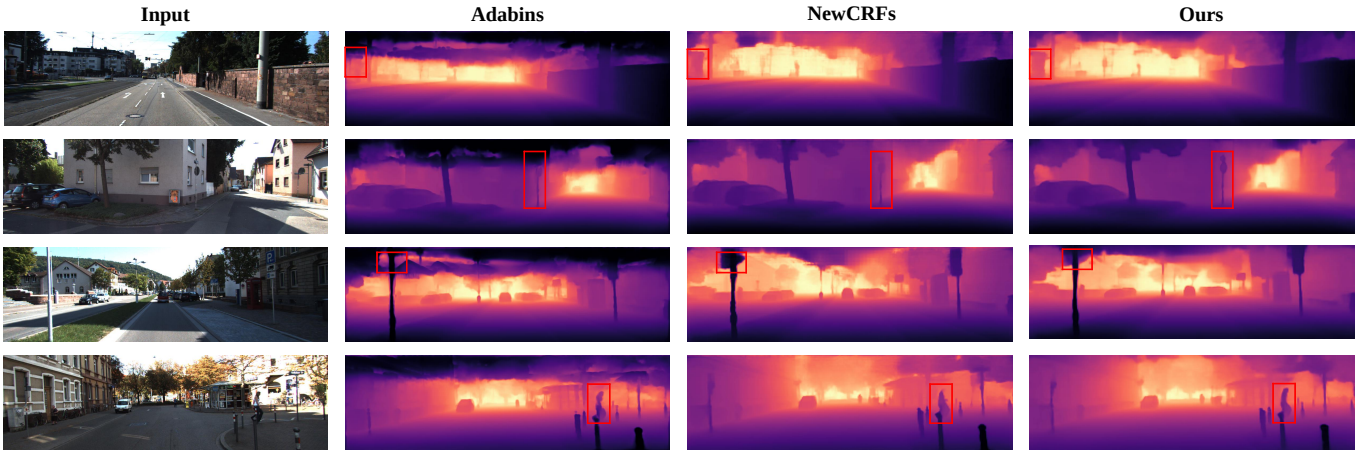

Fig. 11. **Qualitative results on the Eigen split of the KITTI dataset**. The boxes in red highlight the focused regions.

Method	Cap	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	Abs Rel $\downarrow$	RMSE $\downarrow$	Sq Rel $\downarrow$	RMSE log $\downarrow$
P3Depth [28]	0-80m	0.953	0.993	0.998	0.071	2.842	0.270	0.103
TransDepth [25]	0-80m	0.956	0.994	**0.999**	0.064	2.755	0.252	0.098
BTS [6]	0-80m	0.954	0.992	0.998	0.061	2.834	0.261	0.099
PWA [50]	0-80m	0.958	0.994	**0.999**	0.060	2.604	0.221	0.093
Adabins [27]	0-80m	0.964	0.995	**0.999**	0.058	2.360	0.190	0.088
TEDepth‡ [29]	0-80m	0.968	0.996	**0.999**	0.056	2.223	0.174	0.084
NeWCRFs [7]	0-80m	0.974	**0.997**	**0.999**	0.052	2.129	0.155	0.079
Ours	0-80m	**0.977**	**0.997**	**0.999**	**0.050**	**2.034**	**0.142**	**0.076**
TransDepth [25]	0-50m	0.963	0.995	0.999	0.061	1.992	0.185	0.091
BTS [6]	0-50m	0.962	0.994	0.999	0.058	1.995	0.183	0.090
PWA [50]	0-50m	0.965	0.995	0.999	0.057	1.872	0.161	0.087
TEDepth [29]†	0-50m	0.972	0.997	0.999	0.054	1.664	0.135	0.080
P3Depth [28]	0-50m	0.974	0.997	0.999	0.055	1.651	0.130	0.081
Ours	0-50m	**0.980**	**0.998**	**1.000**	**0.049**	**1.541**	**0.112**	**0.073**

TABLE I

PERFORMANCE COMPARISON ON THE EIGEN SPLIT OF KITTI DATASET. ‡ STANDS FOR ENSEMBLE-BASED APPROACH. TEDEPTH INTEGRATES 3 DEPTH ESTIMATION MODELS, WITH CONFORMER [51], RESNEXT101 [52] AND VOLO [53] AS BACKBONES, RESPECTIVELY. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

- $\log_{10}$: $\frac{1}{|\mathbf{T}|}\sum_{\hat{\mathbf{d}}\in\mathbf{T}}\left\|\log_{10}\hat{\mathbf{d}} - \log_{10}\mathbf{d}\right\|^2$;
- $\delta < thr$: % of $\hat{\mathbf{d}}$ satisfies $\left(\max\left(\frac{\hat{\mathbf{d}}}{\mathbf{d}}, \frac{\mathbf{d}}{\hat{\mathbf{d}}}\right) = \delta < thr\right)$ for $thr = 1.25, 1.25^2, 1.25^3$.

B. Implementation Details

The proposed framework is implemented in PyTorch [54] and trained on NVIDIA TITAN RTX GPUs. We use the Adam optimizer [54] where $\beta 1 = 0.9, \beta 2 = 0.99$ and a batch size of 8. The learning rate is scheduled through polynomial decay from 2e-5 to 2e-6. For the KITTI dataset, we set the vertical field of view $[\alpha_v^{\min}, \alpha_v^{\max}]$, vertical azimuth resolution σ_v and horizontal azimuth resolution σ_u to [-2.812, 22.804], 0.4 and 0.16, respectively. The parameter settings roughly match the Velodyne HDL-64E LiDAR for KITTI acquisition. For the NYU-Depth-v2 and SUN RGB-D, we remain σ_v and σ_u unchanged and adjust $[\alpha_v^{\min}, \alpha_v^{\max}]$ to $[-22.804, 22.804]$ because the sensors collecting these two datasets have a larger vertical field of view. The Gaussian filter size is 5×5 and the threshold t is 0.02. The entire training schedule contains two stages. We first train the depth network and Gaussian map network for 20 epochs, which are utilized to construct our sparse pseudo-LiDAR. The collected sparse depth maps can be saved offline to reduce the memory consumption and training time of the second stage. Then, we input the RGB image and sparse depth map to train the depth estimator for 20 epochs. The proposed framework is the same in training and inference phases.

C. Comparison to the State-of-the-art Competitors

KITTI. We first compare the proposed method against the competing approaches on the Eigen split. As demonstrated in Table I, the proposed method outperforms existing approaches, including ensemble-based or not. In addition, noticeable improvement is achieved on Sq Rel and RMSE by improving the NeWCRFs by 6.5 % and 4.5 %, respectively, when the maximum depth is capped at 80m. For the 50m cap, our

Method	dataset	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	SILog ↓	Abs Rel	iRMSE ↓	RMSE ↓	Sq Rel ↓
DORN [22]	validation	0.913	0.985	0.995	12.22	11.78	11.68	3.80	3.03
BTS [6]	validation	0.938	0.987	0.996	10.67	7.51	8.10	3.37	1.59
BA-Full [55]	validation	0.938	0.988	0.997	10.64	8.25	8.47	3.30	1.81
NeWCRFs [7]	validation	0.968	**0.995**	0.998	8.31	5.54	6.34	2.55	0.89
Ours	validation	**0.971**	**0.995**	**0.999**	**7.92**	**5.31**	**6.02**	**2.46**	**0.82**
PAP† [34]	online test	-	-	-	13.08	10.27	13.95	-	2.72
P3Depth [28]	online test	-	-	-	12.82	9.92	13.71	-	2.53
DORN [22]	online test	-	-	-	11.77	8.78	12.98	-	2.23
BTS [6]	online test	-	-	-	11.67	9.04	12.23	-	2.21
BA-Full [55]	online test	-	-	-	11.61	9.38	12.23	-	2.29
PackNet-SAN [56]	online test	-	-	-	11.54	9.12	12.38	-	2.35
PWA [50]	online test	-	-	-	11.45	9.05	12.32	-	2.30
ViP-DeepLab† [33]	online test	-	-	-	10.80	8.94	11.77	-	2.19
NeWCRFs [7]	online test	-	-	-	10.39	8.37	11.03	-	1.83
Ours	online test	-	-	-	**10.14**	**8.16**	**10.84**	-	**1.75**

TABLE II
PERFORMANCE COMPARISON ON THE OFFICIAL SPLIT OF KITTI DATASET. NOTE THAT THE SQ REL IS COMPUTED IN A DIFFERENT WAY HERE. THE SILOG IS THE MAIN RANKING METRIC. † INDICATES THAT THE MODEL IS TRAINED USING ADDITIONAL SEMANTIC LABELS. "-" MEANS NOT APPLICABLE.

Method	Cap	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	Abs Rel ↓	RMSE ↓	Sq Rel ↓	$\log_{10} \downarrow$
TRL† [35]	0-10m	0.815	0.962	0.992	0.144	0.501	-	-
VNL [24]	0-10m	0.875	0.976	0.994	0.108	0.416	-	0.048
DAV [57]	0-10m	0.882	0.980	0.996	0.108	0.412	-	-
BTS [6]	0-10m	0.871	0.977	0.995	0.113	0.407	0.066	0.049
Long *et al.* [58]	0-10m	0.890	0.982	0.996	0.101	0.377	-	0.044
PWA [50]	0-10m	0.892	0.985	0.997	0.105	0.374	-	0.045
TransDepth [25]	0-10m	0.900	0.983	0.996	0.106	0.365	-	0.045
Adabins [27]	0-10m	0.903	0.984	0.997	0.103	0.364	-	0.044
P3Depth [28]	0-10m	0.898	0.981	0.996	0.104	0.356	-	0.043
TEDepth‡ [29]	0-10m	0.907	0.987	0.998	0.100	0.349	-	0.043
Jiao *et al.*† [32]	0-10m	0.915	0.983	0.996	0.100	0.333	-	0.040
NeWCRFs [7]	0-10m	0.922	**0.992**	0.998	0.095	0.334	0.045	0.041
BinsFormer⋆ [31]	0-10m	0.925	0.989	0.997	0.094	0.330	-	0.040
Ours	0-10m	**0.933**	0.991	**0.999**	**0.088**	**0.313**	**0.040**	**0.038**

TABLE III
PERFORMANCE COMPARISON ON THE NYU-DEPTH-V2 DATASET. HERE TEDEPTH INTEGRATES 5 DEPTH ESTIMATION MODELS, WITH CONFORMER [51], VOLO [53], RESNET101 [59], R50+VIT-B/16 [26] AND DENSENET161 [60] AS BACKBONES, RESPECTIVELY. ⋆ INDICATES THAT THE MODEL IS TRAINED USING AUXILIARY SCENE CLASS INFORMATION.

method goes well beyond the P3Depth. Fig. 11 presents a qualitative depth comparison. It can be seen that the NeWCRFs struggles with difficult object boundaries, *e.g.*, humans and signposts, while our method is capable of estimating these smaller details.

We then compare against the competing approaches on the official split. The results on the validation set and test set are listed in Table II, where the results on the test set are generated by an online server. As can be seen, our method outperforms previous approaches again. Besides, we notice that even compared to the approaches using additional semantic labels, such as ViP-DeepLab, our method is able to exceed them by a large margin, which emphasizes the effectiveness of the developed sparse pseudo-LiDAR. We visualize some depth predictions and corresponding error maps from the online server in Fig. 10.

NYU-Depth-v2. To demonstrate the generality of the proposed sparse pseudo-LiDAR, we evaluate our method on the indoor dataset NYU-Depth-v2 collected by Kinect. The results are summarized in Table III. It can be seen that our method exceeds existing leading approaches, such as BinsFormer using auxiliary scene class information, on most metrics. Fig. 12 displays a qualitative depth comparison. Obviously, our method is more robust to texture variations and preserves fine-grained details.

D. Generalization Ability

We study the generalization performance of our method to verify that it does learn transferable discriminative features rather than just memorizing the statistics of training data. The

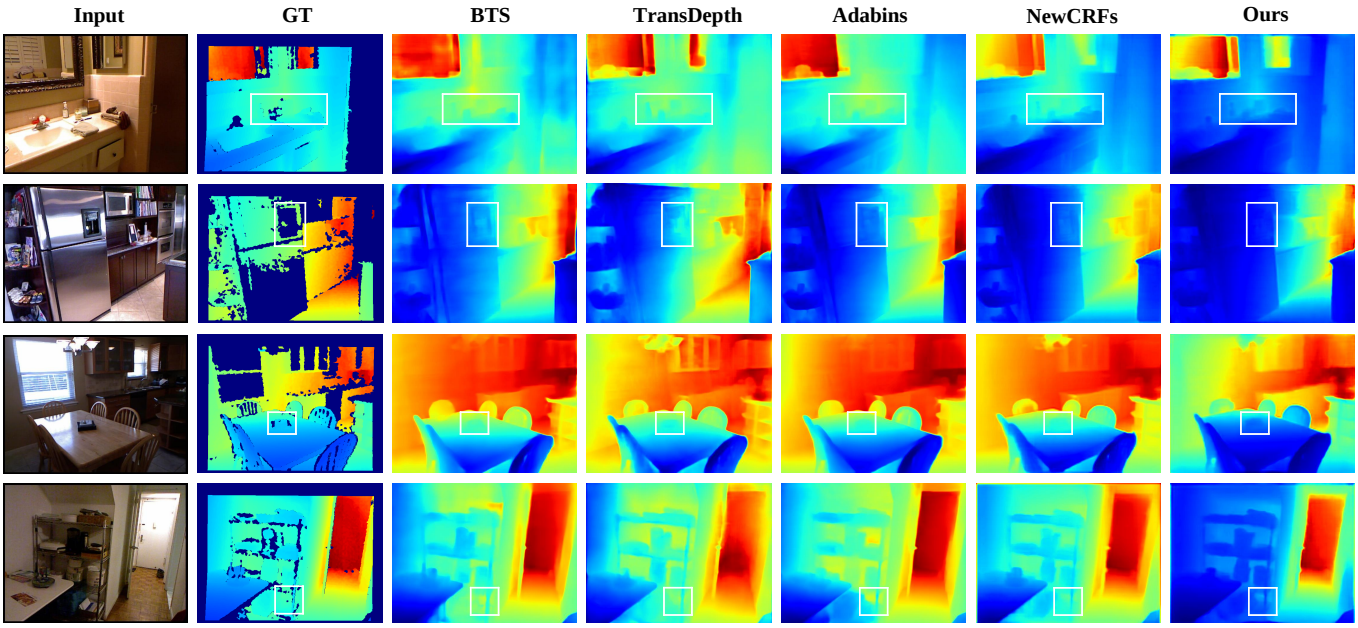

Fig. 12. **Qualitative results on the NYU-Depth-v2 dataset**. The boxes in white highlight the focused regions.

Method	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	Abs Rel $\downarrow$	RMSE $\downarrow$	$\log_{10} \downarrow$
Chen [61]	0.757	0.943	0.166	0.494	0.071
VNL [24]	0.696	0.912	0.183	0.541	0.082
BTS [6]	0.740	0.933	0.172	0.515	0.075
Adabins [27]	0.771	0.944	0.159	0.476	0.068
P3Depth [28]	0.698	-	0.178	0.541	-
NeWCRFs [7]	0.798	0.967	0.151	0.424	0.064
BinsFormer⋆ [31]	0.805	0.963	0.143	0.421	0.061
Ours	**0.817**	**0.970**	**0.135**	**0.408**	**0.060**

TABLE IV
RESULTS ON THE SUN RGB-D DATASET WITH MODELS TRAINED FROM THE NYU-DEPTH-V2 DATASET.

ID	DP	GS	AS	$\delta < 1.25 \uparrow$	Abs Rel $\downarrow$	RMSE $\downarrow$	Sq Rel $\downarrow$
1				0.974	0.052	2.129	0.155
2	✓			0.975	0.053	2.090	0.149
3	✓	✓		0.976	0.052	2.053	0.147
4	✓		✓	0.976	0.052	2.052	0.148
5	✓	✓	✓	**0.977**	**0.050**	**2.034**	**0.142**

TABLE V
ABLATION STUDY. DP: DENSE PSEUDO-LIDAR; GS: GEOMETRIC SAMPLING; AS: APPEARANCE SAMPLING.

models are trained on the NYU-Depth-v2 dataset but evaluated on the SUN RGB-D dataset, without any fine-tuning. The results are listed in Table IV, which indicate that our method indeed generalizes better across different cameras.

E. Ablation Study

To demonstrate the contribution of individual components to the overall performance, we present an ablative analysis on the KITTI dataset, as depicted in Table V.

We start from the baseline (ID 1) [7]. By first introducing the dense pseudo-LiDAR, which in fact is equivalent to depth refinement, we observe a slight performance improvement on most metrics (ID 2). Although the performance is improved, it

Method	$\delta < 1.25 \uparrow$	Abs Rel $\downarrow$	RMSE $\downarrow$	Sq Rel $\downarrow$
Baseline [7]	0.974	0.052	2.129	0.155
DP	0.975	0.053	2.090	0.149
DP + US	0.974	0.053	2.120	0.155
DP + KS	0.969	0.057	2.228	0.174
DP + UGS	0.976	0.052	2.077	0.150
Ours	**0.977**	**0.050**	**2.034**	**0.142**

TABLE VI
COMPARISON OF DIFFERENT SAMPLING MANNERS. DP: DENSE PSEUDO-LIDAR; US: UNIFORM SAMPLING; KS: KEYPOINT-BASED SAMPLING; UGS: UNCERTAINTY-GUIDED SAMPLING.

Method	$\delta < 1.25 \uparrow$	Abs Rel $\downarrow$	RMSE $\downarrow$	Sq Rel $\downarrow$
Baseline [7]	0.974	0.052	2.129	0.155
Ours with SP (BTS [6])	0.976	0.051	2.050	0.146
Ours with SP (TransDepth [25])	0.976	0.051	2.047	0.145
Ours	**0.977**	**0.050**	**2.034**	**0.142**

TABLE VII
EFFECT OF DIFFERENT DEPTH NETWORKS IN THE PROPOSED SPARSE PSEUDO-LIDAR. SP (BTS [6]: BTS IS USED AS THE DEPTH NETWORK IN OUR SPARSE PSEUDO-LIDAR. SP (TRANSDEPTH [25]): TRANSDEPTH IS USED AS THE DEPTH NETWORK IN OUR SPARSE PSEUDO-LIDAR.

Method	Abs Rel $\downarrow$	Parameters (M) $\downarrow$	Inference Time (s) $\downarrow$
TEDepth [29]	0.100	535	0.557
NeWCRFs [7]	0.095	270	0.052
BinsFormer⋆ [31]	0.094	255	0.216
Ours (SP)	-	**181**	**0.038**
Ours (DE)	**0.088**	270	0.052

TABLE VIII
COMPARISON OF MODEL PARAMETERS AND INFERENCE TIME ON THE NYU-DEPTH-V2 DATASET. SP: SPARSE PSEUDO-LIDAR; DE: DEPTH ESTIMATOR.

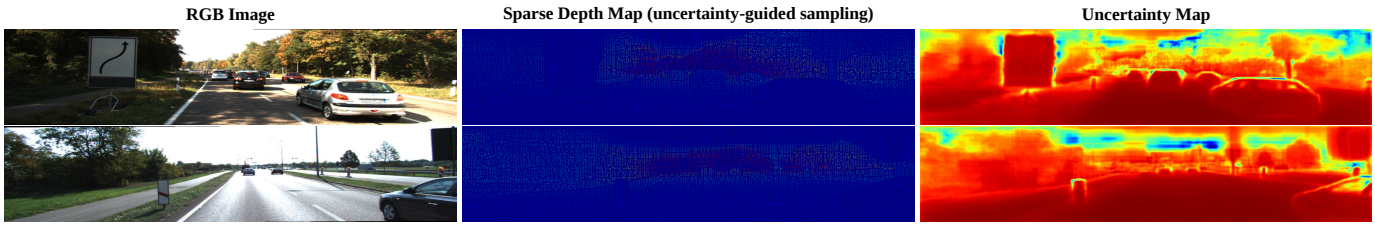

Fig. 13. **Illustration of the uncertainty-guided sampling**.

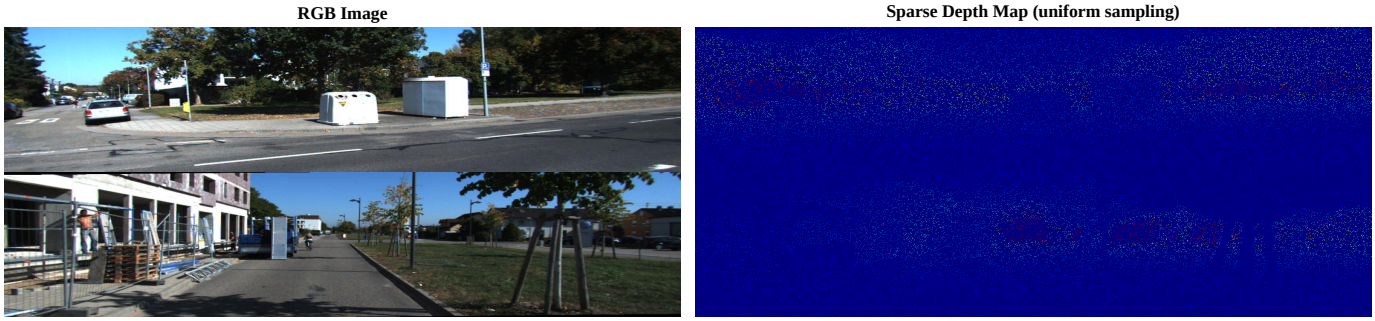

Fig. 14. **Illustration of the uniform sampling**.

Fig. 15. **Illustration of the keypoint-based sampling**.

is limited. Then, we append the geometric sampling (ID 3) or the appearance sampling (ID 4) to mimic the scan pattern of LiDAR, again with a consistent performance improvement. We finally combine the dense pseudo-LiDAR, geometric sampling and appearance sampling, and construct the proposed sparse pseudo-LiDAR (ID 5) thereby, achieving the best results. The reason behind may be that geometric and appearance samplings bring additional valuable information such as, geometric correlation, semantic correlation and highlighted boundaries.

In addition, we develop more sampling manners including uniform sampling, keypoint-based sampling, and uncertainty-guided sampling to further indicate the strengths of geometric sampling and appearance sampling.

Uniform sampling. The sparse depth map captured by LiDAR has roughly 18400 samples per map. In this case, we uniformly sample 18400 points from the dense depth map (Fig. 14).

Keypoint-based sampling. We adopt the scale-invariant feature transform algorithm [46] to extract keypoints, which are then used to sample the dense depth map (Fig. 15).

Uncertainty-guided sampling. To acquire the uncertainty of dense depth map, we modify the prediction layer of DepthNet to jointly predict the dense depth map and the corresponding uncertainty map. We adopt $\mathcal{L}_U$ to supervise the uncertainty prediction $\hat{\mathbf{U}}(\mathbf{p})$,

$$\mathcal{L}_U = \sum_{\mathbf{p}} \left| \hat{\mathbf{U}}(\mathbf{p}) - \mathbf{U}(\mathbf{p}) \right| \tag{17}$$

with

$$\mathbf{U}(\mathbf{p}) = 1 - \exp\left(-\frac{\left| \hat{\mathbf{d}}(\mathbf{p}) - \mathbf{d}(\mathbf{p}) \right|}{b} \right), \tag{18}$$

where b is a coefficient that controls the tolerance for error and is set as 0.2. We use the 4×4 sliding window to choose the point with the lowest uncertainty in each window and use these points as sampling points (Fig. 13).

The results are demonstrated in Table VI. It is not surprising that the uncertainty-guided sampling performs better than the other two sampling methods as it selects points with accurate depth values through uncertainty. Nonetheless, the uncertainty-guided sampling is inferior to the proposed sampling manners, indicating that accurate depth values in pseudo-LiDAR are alone not enough, but their distribution patterns matter as well.

We further explore the effect of different depth networks in the proposed sparse pseudo-LiDAR. As shown in Table VII,

we use the BTS and TransDepth with relatively lower accuracy as the depth network in the proposed sparse pseudo-LiDAR. The results indicate that both variants exceed the baseline by a large margin and our framework is robust to the variation of depth networks in sparse pseudo-LiDAR.

F. Model Parameters and Inference Time

We compare the proposed framework against several well-behaved approaches including TEDepth, NeWCRFs and BinsFormer in terms of inference time and the number of model parameters. As can be seen from Table VIII, our depth estimator shares an identical number of parameters and inference time with NeWCRFs. When our depth estimator is equipped with the designed sparse pseudo-LiDAR, it has fewer parameters and less inference time than TEDepth. Although it has more parameters than BinsFormer, it is 58% faster than BinsFormer in terms of inference time. In addition, the proposed framework is much more performant than the compared approaches. Therefore, the proposed framework achieves a better balance between performance, number of parameters and inference time.

V. Conclusion

A novel perspective is proposed in this work for monocular depth estimation by introducing the concept of pseudo-LiDAR. Furthermore, we propose geometric sampling and appearance sampling to simulate the scan pattern of LiDAR. The former allows geometric relation to be built, and the latter tells us the likelihood of a "pseudo-LiDAR ray" returning an answer. Extensive experiments on the KITTI, NYU-Depth-v2 and SUN RGB-D datasets show that the proposed method outperforms previous state-of-the-art competitors. We hope our novel initiative will encourage more sophisticated pseudo-LiDAR-based depth estimation achievements.

References

[1] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011.

[2] Po-Yi Chen, Alexander H Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2624–2632, 2019.

[3] Oskar Natan and Jun Miura. End-to-end autonomous driving with semantic depth cloud mapping and multi-agent. *IEEE Transactions on Intelligent Vehicles*, 2022.

[4] Tianze Gao, Huihui Pan, and Huijun Gao. Monocular 3d object detection with sequential feature association and depth hint augmentation. *IEEE Transactions on Intelligent Vehicles*, 7(2):240–250, 2022.

[5] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, pages 2366–2374, 2014.

[6] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

[7] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3916–3925, June 2022.

[8] Qiang Liu, Haosong Yue, Zhanggang Lyu, Wei Wang, Zhong Liu, and Weihai Chen. Sehlnet: Separate estimation of high-and low-frequency components for depth completion. In *2022 International Conference on Robotics and Automation*, pages 668–674. IEEE, 2022.

[9] Mu Hu, Shuling Wang, Bin Li, Shiyu Ning, Li Fan, and Xiaojin Gong. Penet: Towards precise and efficient image guided depth completion. In *2021 IEEE International Conference on Robotics and Automation*, pages 13656–13662. IEEE, 2021.

[10] 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.

[11] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *European Conference on Computer Vision*, pages 120–136. Springer, 2020.

[12] Rui Qian, Divyansh Garg, Yan Wang, Yurong You, Serge Belongie, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. End-to-end pseudo-lidar for image-based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5881–5890, 2020.

[13] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019.

[14] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. In *International Conference on Learning Representations*, 2019.

[15] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10012–10022, October 2021.

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012.

[17] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012.

[18] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 567–576, 2015.

[19] Ashutosh Saxena, Sung H Chung, Andrew Y Ng, et al. Learning depth from single monocular images. In *Advances in Neural Information Processing Systems*, volume 18, pages 1–8, 2005.

[20] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth International Conference on 3D Vision*, pages 239–248. IEEE, 2016.

[21] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(11):3174–3182, 2017.

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

[23] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018.

[24] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5684–5693, 2019.

[25] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 16269–16279, October 2021.

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.

[27] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021.

[28] Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. P3depth: Monocular depth estimation with a piecewise planarity prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1610–1621, 2022.

[29] Shuwei Shao, Ran Li, Zhongcai Pei, Zhong Liu, Weihai Chen, Wentao Zhu, Xingming Wu, and Baochang Zhang. Towards comprehensive monocular depth estimation: Multiple heads are better than one. *IEEE Transactions on Multimedia*, 2022.

[30] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *arXiv preprint arXiv:2203.14211*, 2022.

[31] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022.

[32] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *Proceedings of the European Conference on Computer Vision*, September 2018.

[33] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3997–4008, June 2021.

[34] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4106–4115, 2019.

[35] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *Proceedings of the European Conference on Computer Vision*, pages 235–251, 2018.

[36] Xinshuo Weng and Kris Kitani. Monocular 3d object detection with pseudo-lidar point cloud. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, Oct 2019.

[37] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3142–3152, October 2021.

[38] Changcai Li, Haitao Meng, Gang Chen, and Long Chen. Real-time pseudo-lidar 3d object detection with geometric constraints. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems*, pages 3298–3303. IEEE, 2022.

[39] Andrea Simonelli, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, and Elisa Ricci. Are we missing confidence in pseudo-lidar methods for monocular 3d object detection? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3225–3233, October 2021.

[40] Chaokang Jiang, Guangming Wang, Yanzi Miao, and Hesheng Wang. 3d scene flow estimation on pseudo-lidar: Bridging the gap on estimating point motion. *IEEE Transactions on Industrial Informatics*, 2022.

[41] Huiying Deng, Guangming Wang, Zhiheng Feng, Chaokang Jiang, Xinrui Wu, Yanzi Miao, and Hesheng Wang. Pseudo-lidar for visual odometry. *arXiv preprint arXiv:2209.01567*, 2022.

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

[43] Robert Bridson. Fast poisson disk sampling in arbitrary dimensions. *SIGGRAPH sketches*, 10(1):1, 2007.

[44] Yuval Eldar, Michael Lindenbaum, Moshe Porat, and Yehoshua Y Zeevi. The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing*, 6(9):1305–1315, 1997.

[45] David S Early and David G Long. Image reconstruction and enhanced resolution imaging from irregular samples. *IEEE Transactions on Geoscience and Remote Sensing*, 39(2):291–302, 2001.

[46] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[47] Michael McCool and Eugene Fiume. Hierarchical poisson disk sampling distributions. In *Graphics Interface*, volume 92, pages 94–105, 1992.

[48] Russel E Caflisch. Monte carlo and quasi-monte carlo methods. *Acta Numerica*, 7:1–49, 1998.

[49] Harald Niederreiter. *Random number generation and quasi-Monte Carlo methods*. SIAM, 1992.

[50] Sihaeng Lee, Janghyeon Lee, Byungju Kim, Eojindl Yi, and Junmo Kim. Patch-wise attention network for monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1873–1881, 2021.

[51] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: local features coupling global representations for visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 367–376, October 2021.

[52] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017.

[53] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[54] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Advances in Neural Information Processing Systems Workshop Autodiff*, 2017.

[55] Shubhra Aich, Jean Marie Uwabeza Vianney, Md Amirul Islam, and Mannat Kaur Bingbing Liu. Bidirectional attention network for monocular depth estimation. In *2021 IEEE International Conference on Robotics and Automation*, pages 11746–11752. IEEE, 2021.

[56] Vitor Guizilini, Rares Ambrus, Wolfram Burgard, and Adrien Gaidon. Sparse auxiliary networks for unified monocular depth prediction and completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11078–11088, 2021.

[57] Lam Huynh, Phong Nguyen-Ha, Jiri Matas, Esa Rahtu, and Janne Heikkilä. Guiding monocular depth estimation using depth-attention volume. In *European Conference on Computer Vision*, pages 581–597. Springer, 2020.

[58] Xiaoxiao Long, Cheng Lin, Lingjie Liu, Wei Li, Christian Theobalt, Ruigang Yang, and Wenping Wang. Adaptive surface normal constraint for depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 12849–12858, October 2021.

[59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[60] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[61] Xiaotian Chen, Xuejin Chen, and Zheng-Jun Zha. Structure-aware residual pyramid network for monocular depth estimation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 694–700, 2019.

Shuwei Shao received a B.Eng. from Xidian University, Xi'an, Shanxi, China, in 2019. He is currently pursuing a Ph.D. degree in the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. His current research interests include depth estimation, image registration and 3D reconstruction.

Zhongcai Pei received B.Eng., M.Eng. and Ph.D. degrees from the Harbin Institute of Technology, Harbin,China, in 1991, 1994, and 1997, respectively. He has been with the School of Automation Science and Electronic Engineering, Beihang University, as an Associate Professor from 2000 and as a Professor since 2006. He has published over 50 technical papers in referred journals and conference proceedings and filed more than 20 patents. His research interests include bionic CPG mechanisms and control, computer vision, parallel mechanism and robotics.

Weihai Chen received a B.Eng. degree from Zhejiang University, China, in 1982, and the M.Eng. and Ph.D. degrees from Beihang University, Beijing, China, in 1988 and 1996, respectively. He is currently a Professor with the School of Automation Science and Electrical Engineering, Beihang University. He has published more than 200 technical articles in referred journals and conference proceedings and filed more than 30 patents. His research interests include computer vision, image processing, automation, bioinspired robotics, and precision mechanisms. He is also a member of the Technical Committee on Manufacturing Automation of the IEEE Robotics and Automation Society. He was a recipient of the 2017 Best Paper Award for the IEEE Transactions on Industrial Electronics and the 2018 IET Premium Award for the *IET Intelligent Transport Systems*.

Qiang Liu received his B.Eng. degree in Automation from Beihang University in 2020. He is currently pursuing his Master's degree in the School of Automation Science and Electrical Engineering at Beihang University. His research interests lie in the areas of computer vision.

Haosong Yue received a B.S. degree in control science and engineering from the University of Science and Technology Beijing, Beijing, China, in 2009, and a Ph.D. degree in control science and engineering from Beihang University, Beijing, China, in 2015. He is currently an Associate Professor at the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. His research interests include robot vision and control, simultaneous localization and mapping, and depth completion.

Zhengguo Li (F'23) received a B.Sci (Applied Mathematics). and M.Eng. (Automatic Control) from Northeastern University, Shenyang, China, in 1992 and 1995, respectively, and a Ph.D. degree (Automatic Control) from Nanyang Technological University, Singapore, in 2001.

His research interests include video processing and delivery, computational photography, switched and impulsive control, sensor fusion and physics-driven deep learning. He has co-authored one monograph, more than 200 journal/conference papers including more than 60 IEEE Transaction papers, and eleven granted USA patents, including normative technologies on scalable extension of H.264/AVC and HEVC. He has been actively involved in the development of H.264/AVC and HEVC since 2002. He had three informative proposals adopted by the H.264/AVC and three normative proposals adopted by the HEVC. Currently, he is with the Agency for Science, Technology and Research, Singapore.

Zhengguo served as a TPC Chair of IEEE IECON 2023 and 2020, a special session chair of IEEE ICASSP 2022, a Technical Brief Co-Founder of SIGGRAPH Asia, and leading Workshop Chair of IEEE ICME in 2013. He was an Associate Editor of IEEE Signal Processing Letters from 2014-2016, IEEE Transactions on Image Processing from 2016-2020, IEEE Transactions on Circuits and Systems for Video Technology from 2020-2023, APSIPA Transactions on Signal and Information Processing from 2022-2025, an editor of MDPI Sensors from 2022-2024, and a Senior Area Editor of IEEE Transactions on Image Processing from 2020-2024.