DeepLiDAR: Deep Surface Normal Guided Depth Prediction for Outdoor Scene from Sparse LiDAR Data and Single Color Image

Jiaxiong Qiu[1*]   Zhaopeng Cui[2*]   Yinda Zhang[3*]
Xingdi Zhang[1]   Shuaicheng Liu[1,4†]   Bing Zeng[1]   Marc Pollefeys[2,5]
[1] UESTC   [2]ETH Zürich   [3]Google   [4]Megvii Technology   [5]Microsoft

Abstract

In this paper, we propose a deep learning architecture that produces accurate dense depth for the outdoor scene from a single color image and a sparse depth. Inspired by the indoor depth completion, our network estimates surface normals as the intermediate representation to produce dense depth, and can be trained end-to-end. With a modified encoder-decoder structure, our network effectively fuses the dense color image and the sparse LiDAR depth. To address outdoor specific challenges, our network predicts a confidence mask to handle mixed LiDAR signals near foreground boundaries due to occlusion, and combines estimates from the color image and surface normals with learned attention maps to improve the depth accuracy especially for distant areas. Extensive experiments demonstrate that our model improves upon the state-of-the-art performance on KITTI depth completion benchmark. Ablation study shows the positive impact of each model components to the final performance, and comprehensive analysis shows that our model generalizes well to the input with higher sparsity or from indoor scenes.

1. Introduction

Measuring dense and accurate depth for outdoor environment is critically important for various applications, such as autonomous driving and unmanned aerial vehicles. Most of the active depth sensing solutions for indoor environment fail due to strong interference of the passive illumination [11, 41], and stereo methods usually become less accurate for distant areas due to lower resolutions and smaller triangulation angles compared to the close areas [45]. As a result, LiDAR is the dominating reliable solution for the outdoor environment. However, the high-end LiDAR is prohibitively expensive, and the commodity level devices suffer from the notorious low resolution [27] which causes troubles for perception in middle or long range area. Spa-
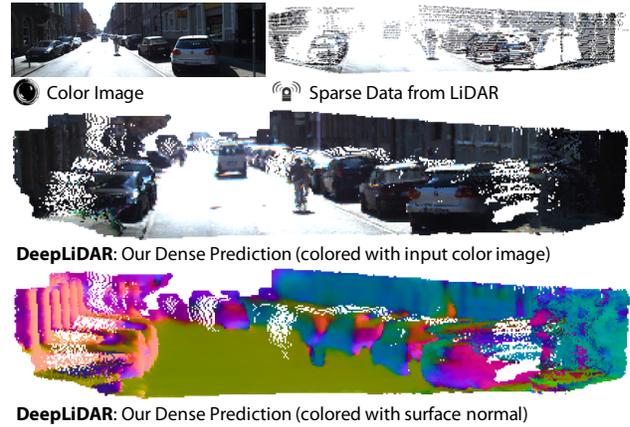
Figure 1. **Our system** takes as input a color image and a sparse depth image from the LiDAR (Row 1), and output a dense depth map (Row 2). Rather than directly producing the complete depth, our model estimates surface normals (Row 3) as the intermediate representation which is helpful to produce accurate depth.

tial and temporal fusion provide denser depth but either requires multiple devices or suffers from dynamic objects and latency. An affordable solution for immediate access of the dense and accurate depth still does not exist.

One promising attempt is to take a sparse but accurate depth from a low-cost LiDAR and make it dense with the help of an aligned color image. With the great success of deep learning, an obvious approach is to directly feed the sparse depth and color image into a neural network and regress for the dense depth. Unfortunately, such a black-box does not work equally well compared to interpretable models, where local depth affinity is learned from color image to interpolate the sparse signal. For indoor scenes, Zhang *et al*. [55] estimated the surface normals as the intermediate representation and solved for depth via a separate optimization, which achieved superior results. However, it is not well studied if the surface normal is a reasonable representation for the outdoor scene and how such system performs.

In this work, we propose an end-to-end deep learning system to produce dense depth from sparse LiDAR data and a color image taken from outdoor on-road scenes leveraging

---

[*]indicates equal contributions.
[†]indicates corresponding author.
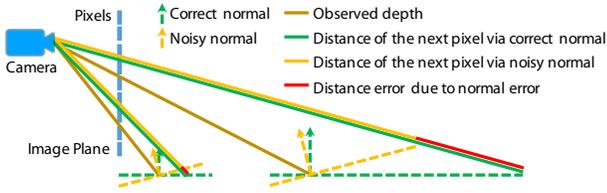
Figure 2. **Sensitivity to noise.** Reconstructing depth from normal becomes more sensitive to the noise/error in the estimated normal when the distance goes up. We show two cases to estimate the depth of the neighboring pixel via correct (green) and noisy (yellow) normal. The further case results in much larger error (red) compared to the closer one even though the surface normal error is the same ($15°$) for two cases.

surface normal as the intermediate representation. We find it non-trivial to make such a system work equally well as in the indoor environment, generally because of the following three challenges:

Data Fusion. How to combine the given sparse depth and dense color image is still an open problem. One common manner is to concatenate them (usually with a binary mask indicating the pixel-wise availability of the LiDAR depth) directly as the network input (i.e. early fusion), in which the network has the best access to all sources of inputs starting from the encoder. However, the result may produce artifacts near the boundaries of the missing values, or merely copy depth from where it is available but fail otherwise. Inspired by the idea of leveraging intermediate affinity, we design an encoder-decoder architecture, namely deep completion unit (DCU), where separate encoders learn affinity from the color image and features from the sparse depth respectively, while the decoder learns to produce dense output. The DCU falls in the style of late fusion architecture but different in that the feature from the sparse depth is summed into the decoder rather than ordinary concatenation. The summation [5] favors the features on both sides in the same domain, and therefore encourages our decoder to learn features more related with depth in order to keep consistent with the feature from the sparse depth. This also saves network parameters as well as inference memory. Empirically, we find DCU benefits both the intermediate surface normal and the final depth estimation.

Sensitivity to Noise. Zhang *et al*. [55] demonstrated that surface normals of indoor scenes are easier to estimate than absolute depth and sufficient to complete the depth given incomplete signals. However, in outdoor scenes, solving depth from normals does not work ubiquitously well especially for the distant area mainly due to the perspective geometry. As shown in Fig. 2, the same surface normal error causes much larger distance error for the horizontal road surface in the far area compared to the close range area. Having these areas hard to be solved from surface normals geometrically, we propose to learn them directly from the raw inputs. Therefore, our model contains two pathways to

estimate dense depth maps from the estimated surface normals and the color image respectively, which are then integrated via automatically learned attention maps. In other words, the attention maps learn to collect better solution for each area from the pathway that is likely to perform better.

Occlusion. As there is almost inevitably a small displacement between the RGB camera and the LiDAR sensor, different depth values are normally mixed with each other along the boundaries due to occlusion when warping LiDAR data to the color camera coordinate, especially for the regions close to the camera (Fig. 5 (b)). Such mixture of depth confuses the model and causes blurry boundaries. Ideally, the model should downgrade the confidence of the sparse depth in these confusing area and learn to fill in using more reliable surroundings. We propose to learn such a confidence mask automatically, which takes the place of the binary availability mask feeding into the surface normal pathway. Even though without ground truth, our model self-supervisely learns this occlusion area containing overlapping sparse depth.

Our full pipeline is shown in Fig. 3. The contributions of this work are as follows. Firstly, we propose an end-to-end neural network architecture that produces dense depth from a sparse LiDAR depth and a color image using the surface normal as the intermediate representation, and demonstrate that surface normal is also a good local depth representation for the outdoor scene. Secondly, we propose a modified encoder-decoder structure to effectively fuse the sparse depth and the dense color image. Thirdly, we investigate the challenges for outdoor scenarios, and design the network to automatically learn a confidence mask for occlusion handling, and attention maps for the integration of depth estimates from both the color image and normals. Lastly, our experiment shows that out model significantly outperforms the state-of-the-art on benchmarks and generalizes well to input sparsity and indoor scenes.

2. Related Work

Depth prediction from sparse samples. Producing dense depth from sparse inputs starts to draw attention when accurate but low-res depth sensors, such as low-cost LiDAR and one-line laser sensors, become widely available. Some methods produced dense depth or disparity via wavelet analysis [16, 30]. Recently, deep learning based approaches were proposed and achieved promising results. Uhrig *et al*. [48] proposed sparsity invariant CNNs to deal with the variant input depth sparsity. Ma *et al*. [34] proposed to feed the concatenation of the sparse depth and the color image into an encoder-decoder deep network, and further extended with self-supervised learning [33]. Jaritz *et al*. [20] combined semantic segmentation to improve the depth completion. Cheng *et al*. [6] learned an affinity matrix to guide the depth interpolation through a recurrent neural net-
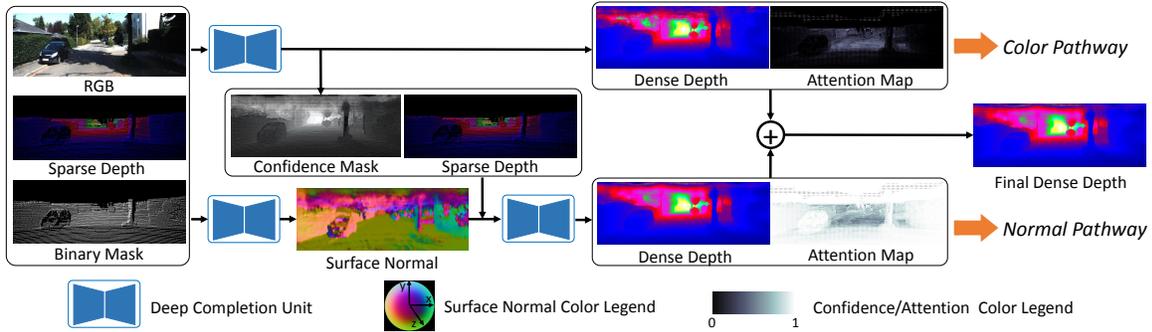
Figure 3. **The pipeline of our model.** Our model consists of two pathways. Both starting from a RGB image, a sparse depth, and a binary mask as the inputs, the surface normal pathway (lower half) produces a pixel-wise surface normal for the scene, which is further combined with the sparse input depth and a confidence mask estimated from the color pathway to produce a dense depth. The color pathway produces a dense depth too. The final dense depth output is the weighted sum of the depths from two pathways using the estimated attention maps.

work. Compared to these work, our model is more physically driven and explicitly exploits surface normals as the intermediate representation.

Depth refinement for indoor environment. In the indoor environment, the quality of the depth from commodity RGB-D sensors is not ideal due to the limitation of the sensing technologies [3, 37, 11]. A lot of works have been proposed to improve the depth using an aligned high-resolution color image. One family of approach is depth super-resolution that targets improving the resolution of the depth image [35, 43, 15, 53, 32, 23, 36, 47]. These methods assume a low-resolution but dense depth map without missing signal. The other family of methods is color image guided depth inpainting, which potentially handles large missing area with arbitrary shape. Traditional methods use color as the guidance to compute local affinity or discontinuity [18, 14, 44, 2, 12, 54, 58, 1]. Even though deep learning has been widely used in image inpainting [49, 38, 29, 52], extension of these networks to color guided depth inpainting is not well studied. Zhang *et al.* [55] proposed to estimate surface normals and solve for depth via a global optimization. However, it is still unclear if normals, as the intermediate representation for depth, still work for the outdoor scenes.

Depth estimation from a single RGB image. There are a lot of works estimating depth from only a single color image. Early methods mainly relied on the hand-crafted features and probabilistic graphical models [42, 21, 22, 24, 31]. With the development of deep learning, many methods [9, 25, 40, 28] based on deep neural networks have been proposed for the single-view depth estimation due to the strong feature representation of deep networks. For example, Eigen *et al.* [9] proposed a multi-scale convolutional network to predict depth from coarse to fine. Laina *et al.* [25] proposed a single-scale but deeper fully convolutional architecture. Liu *et al.* [28] combined the strength of deep CNN and continuous CRF in a unified CNN framework. There are also some methods [8, 26, 4, 39] which exploit surface normals during the depth estimation. Eigen *et al.*

[8] and Li *et al.* [26] proposed architectures to predict depth or normals but independently. Chen *et al.* [4] used sparse surface annotation as supervision for depth estimation but not intermediate representation. Qi *et al.* [39] jointly predicted depth and surface normal based on two-stream CNNs and focused on indoor scenes. Most recently, some unsupervised methods [57, 13, 51] were also proposed. Even though these methods produced plausible depth estimation from a single color image, they do not handle sparse depth as an additional input and are not suitable to recover high-quality depth. Moreover, our method is the first to use surface normals as the intermediate representation for the outdoor depth completion.

3. Method

Our model is an end-to-end deep learning framework that takes an RGB image and a sparse depth image reprojected from LiDAR as inputs and produce a dense depth image. As illustrated in Fig. 3, the whole network mainly consists of two pathways: the color pathway and surface normal pathway. The color pathway takes as input the color image and the sparse depth to output a complete depth. The surface normal pathway first predicts a surface normal image from the input color image and sparse depth, which is then combined together with the sparse depth and a confidence mask learned from the color pathway to produce a complete depth. Each of these two pathways are implemented with a stack of deep completion units (DCU), and the depths from two pathways are then integrated by a learned weighted sum to produce the final complete depth.

3.1. Deep Completion Unit

Zhang *et al.* [55] proposes to remove the incomplete depth from the input when predicting either depth or surface normal in order to get rid of the local optima. However, since the sparse depth is strongly correlated with the dense depth and surface normals, it is certainly non-optimal if the network has no chance to learn from it. Inspired by traditional color image guided inpainting [46, 17, 52], we propose a network architecture to have the encoder to learn the
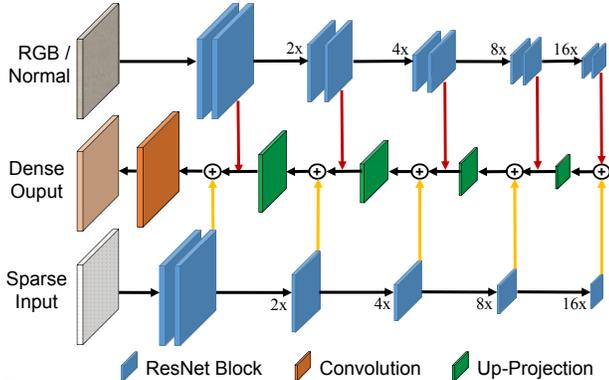
Figure 4. **Detailed architecture of deep completion unit.** Our deep completion unit takes the late fusion strategy, where features from the RGB/normal and sparse depth are combined only in the decoder. Different from [20], we sum the features from both side at each resolution throughout the decoder.

local affinity from color image or surface normals, which is then leveraged by the decoder to conduct interpolation with the features generated from the input sparse depth through another encoder.

The details of our deep completion unit is shown in Fig. 4. Both encoders for RGB/normal and sparse depth consist of a series of ResNet blocks followed by convolution with stride to downsize the feature resolution eventually to $1/16$ of the input. The decoder consists of four up-projection units as introduced in [25] to gradually increase the feature resolutions and to integrate features from both encoders to produce dense output. Since the input sparse depth is strongly related with the decoder output, e.g., surface normal or depth, features from the sparse depth should contribute more in the decoder. As such, we concatenate the features from the RGB/normal but sum the features from the sparse depth onto the features in decoder. As the summation favors the features on both sides in the same domain [5], the decoder is encouraged to learn features more related to depth in order to keep consistent with the feature from the sparse depth. As shown in Fig. 3, we use the DCU to predict either surface normal or depth with the same input but trained with the target ground truth.

3.2. Attention Based Integration

Recovering depth from surface normals does not work ubiquitously well everywhere, and might be sensitive to normal noise in some areas. We propose to generate depth for these areas leveraging priors from the color image rather than geometry from the estimated surface normal. Therefore, our model consists of two pathways in parallel to predict dense depth from the input color image and estimated surface normals respectively. Both pathways also take the sparse depth as input. The final dense depth should be an integration of these two estimated depths, where comparatively more accurate depth measurements are chosen from the right one.

(a) RGB Image

(b) Zoom-in view   (c) Warped depth   (d) Confidence
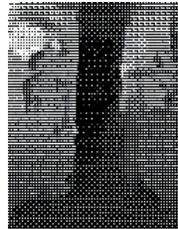
Figure 5. **Occlusion and learned confidence.** (b) shows a zoom-in view of the region marked by the blue box in (a). Due to the displacement between the RGB camera and LiDAR, aligning sparse depth to the color image causes foreground/background depth mixed in the occluded area, like the tree trunk in (c). Our network learns a confidence mask (d) that successfully down-weight the confusing mixed area.

We use an attention mechanism to integrate the depths recovered from two pathways, where the combination of two depths is not fixed but depends on the current context. In particular, we first predict a score map for each of pathways using the last feature map before the output through three convolutions with ReLU. The two score maps from two pathways are then fed into a softmax layer, and converted into a combination weight. The final dense depth output is then calculated as

$$\hat{D} = w_c \cdot \hat{D}_c + w_n \cdot \hat{D}_n, \tag{1}$$

where $\hat{D}_c$ and $\hat{D}_n$ are depths from color and surface normal pathway, and w_c and w_n are the learned combination weights respectively. As it can be seen in Fig. 7, the learned w_c and w_n target on the strong part of their corresponding depth output effectively.

3.3. Confidence Prediction

As mentioned before and shown in Fig. 5, there are ambiguous areas with mixture of foreground and background depth signals due to the displacement between the LiDAR sensor and the color camera. This is usually caused by occlusion, which happens more frequently along the object boundaries in close range. Ideally, we should find these confusing areas and resolve the ambiguity, which however is even more challenging as this requires an accurate 3D geometry estimation near the depth discontinuities. On the contrary, we ask the network to automatically learn a confidence mask to indicate the reliability of the input sparse depth. We replace the simple binary mask, which is an input hard confidence, with the learned confidence mask (m_d) from the color pathway. As shown in Fig. 5, even though without ground truth of such masks, the model could suc-

4

cessfully learn the occlusion area with overlapping sparse depth values (e.g., low weights for tree trunk).

3.4. Loss Function

The loss function of the overall system is defined as:

$$L = \lambda_1 L_d(\hat{D}_n) + \lambda_2 L_d(\hat{D}_c) + \lambda_3 L_d(\hat{D}) + \lambda_4 L_n(N) \quad (2)$$

where L_d defines the loss on the estimated depth, and L_n defines the loss on the estimated surface normal. We use cosine loss following [56] for L_n. For L_d, we use L_2 loss on the estimated depth and a cosine loss on the normal converted from the depth. $\lambda_{1,2,3,4}$ adjusts the weights between terms of the loss function. We adopt a multi-stage training schema for stable convergence. We first set $\lambda_4 = 1$ and all the other weights to zero to only pre-train the surface normal estimation. We then set $\lambda_1 = 0.3, \lambda_2 = 0.3, \lambda_3 = 0.0, \lambda_4 = 0.1$ to further train the color and surface normal pathways. In the end, we set $\lambda_1 = 0.3, \lambda_2 = 0.3, \lambda_3 = 0.5, \lambda_4 = 0.1$ to train the whole system end-to-end. For all the training setting, we use Adam as the optimizer with a starting learning rate of 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is descended to half every 5 epochs.

3.5. Training Data

Due to the lack of the ground-truth normal in the real datasets, we generate a synthetic dataset using an open urban driving simulator Carla [7]. We render 50K training samples including RGB image, sparse depth map, dense depth map, and surface normal image, and the examples are shown in our supplementary materials. For the real data, we use the KITTI depth completion benchmark dataset for fine-tuning and evaluation. The complete surface normal ground truth for the KITTI dataset is computed from the ground-truth dense depth map by local plane fitting [44].

4. Experiments

We perform extensive experiments to verify the effectiveness of our model, including comparison to related work and ablation study. Since one of the major applications of our model is on car-held LiDAR devices, most of the experiments are done on KITTI depth completion benchmark [48]. Nevertheless, we also run our model in indoor environment to verify the generalization capability.

4.1. Comparison to State-of-the-art

Evaluate on KITTI Test Set. We first evaluate our method on the test set of the KITTI depth completion benchmark. The test set contains 1000 data, including color image, sparse LiDAR depth, and transformation between color camera and LiDAR. Ground truth are held, and evaluation can be only done on their server to prevent overfitting. The evaluation server calculates four metrics: root mean squared error (RMSE mm), mean absolute error (MAE mm), root mean squared error of the inverse depth (iRMSE 1/km)

	RMSE	MAE	iRMSE	iMAE
CSPN [6]	1019.64	279.46	2.93	1.15
Spade-RGBsD [20]	917.64	234.81	**2.17**	**0.95**
HMS-Net [19]	841.78	253.47	2.73	1.13
MSFF-Net [50]	836.69	241.54	2.63	1.07
NConv-CNN [10]	829.98	233.26	2.60	1.03
Sparse-to-Dense [33]	814.73	249.95	2.80	1.21
Ours	**758.38**	**226.50**	2.56	1.15

Table 1. **Performance of depth completion on KITTI test set [48].** The evaluation is done via KITTI testing server, and submissions are ranked by RMSE on the leaderboard. Our model outperforms the 2nd [33] on RMSE with a large gap.

	RMSE	MAE	iRMSE	iMAE
Bilateral [44]	2989.02	1200.56	9.67	5.08
Fast [2]	3548.87	1767.80	26.48	9.13
TGV [12]	2761.29	1068.69	15.02	6.28
Zhang *et al.* [55]	1312.10	356.60	4.29	1.41
Ours	**687.00**	**215.38**	**2.51**	**1.10**

Table 2. **Performance of depth completion on KITTI validation set [48].** We compare to non-learning based approaches [44, 2, 12] and Zhang *et al.* [55]. Our method performs the best on all the evaluation metrics.

and mean absolute error of the inverse depth (iMAE 1/km), among which RMSE is the most important indicator and chosen to rank submissions on the leader-board since it measures error directly on depth and penalizes on further distance where depth measurement is more challenging.

The performances of our methods and all the other high ranking methods are listed in Tab. 1. Our method ranks the 1st on the leader-board at the time of submission, outperforming the 2nd with significant improvement. Qualitative comparison with some competing methods [33, 6] are shown in Fig. 6. For each example, we show both the recovered complete depth, together with zoom-in view to high-light some details. In general, our method produces more accurate depth (e.g., the complete car) with better details (e.g., road-side railing). The running time of our model on a single GPU (Nvidia GTX 1080Ti) is 0.07s per image.

Evaluation on KITTI Validation Set. We further compare on the validation set of KITTI benchmark to other related methods that are not on the benchmark, including bilateral filter using color (Bilateral), fast bilateral (Fast), optimization using total variance (TGV), and deep depth completion for indoor scene [55]. Models are trained on the training set only. The quantitative results are shown in Tab. 2. As can be seen, our method significantly outperforms all the other methods. Non-learning based approaches [44, 2, 12] do not perform well possibly because of drastic illumination change and complicate scene structures. Zhang *et al.* [55] performs much better than the above mentioned methods but still far from our model as it does not handle outdoor specific issues.

Qualitative comparisons are shown in Fig. 7. From the highlighted regions, the Bilateral [44] and Fast [2] over-smooth the boundaries and details of objects. In con-
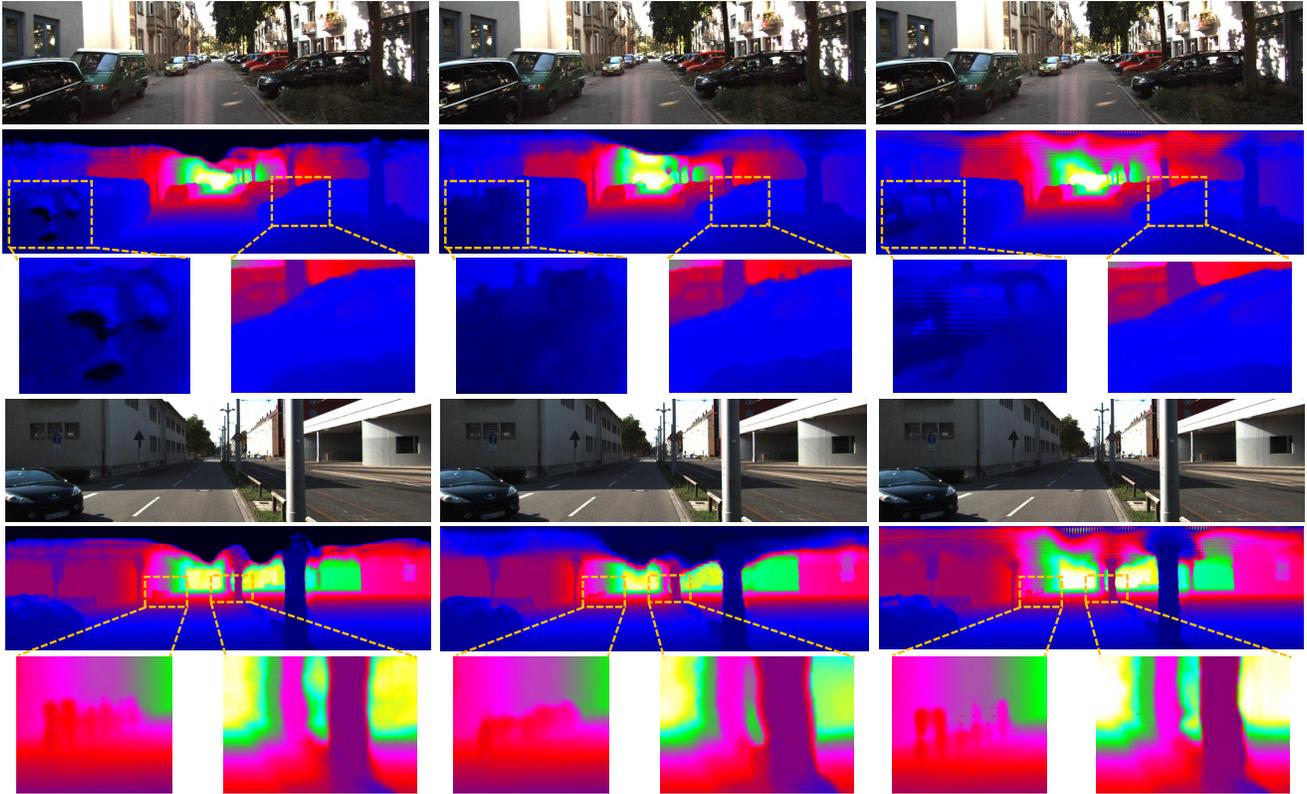
| | (a) Sparse-to-Dense [33] | (b) CSPN [6] | (c) Our method |

Figure 6. **Qualitative comparison on KITTI test set.** We show results of our method and top ranking methods: CSPN [6] and Sparse-to-Dense [33]. For each example, we show color image, dense depth output, and zoom-in view of some local areas. Our model produces more accurate results aligning better with the color image. Our model also preserves thin structures like tree, traffic light, and road lamp.

trast, TGV [12] generates the detailed structures, but noisy smooth surfaces, like roads. Zhang *et al*. [55] performs well on close regions, but worse than our method in far areas and where surface normal estimation fails, e.g., traffic sign and car windows. Our method successfully solves these problems for two reasons. Firstly, we integrate the offline linear optimization into network, which allows end-to-end training for presumably more optimal solutions. From Tab. 3 ("-Attention Integration"), we can see that the depth prediction from our normal pathway is already much better than Zhang *et al*. [55]. Secondly, we further learn a confidence mask to handle occlusion and use the attention based integration to improve the area where normal pathway fails.

4.2. Ablation Study

To understand the impact of each model components on the final performance, we conduct a comprehensive ablation study by disabling each component respectively and show how result changes. Quantitative results are shown in Tab. 3. Performance drops reasonably with each component disabled, and the full model works the best.

Effect of Surface Normal Pathway. To verify if surface normal is a reasonable intermediate depth representation for outdoor scene similar as for the indoor case, we train a model without estimating the normal but directly output

Models	**RMSE**	MAE	iRMSE	iMAE
- Normal Pathway	774.25	258.77	4.65	1.40
- Attention Integration	729.96	239.08	2.74	1.20
- DCU	767.82	246.36	2.69	1.17
- Confidence mask	756.32	272.91	2.70	1.19
Full	**687.00**	**215.38**	**2.51**	**1.10**

Table 3. **Ablation study of depth completion on KITTI validation set.** We re-train our model with each major component disabled and evaluate on KITTI validation set. Our full model with all the components on achieves the best performance.

the complete depth. Under this setting, there is also no attention integration since only one pathway is available. The performance is shown as "-Normal Pathway" in Tab. 3. The performance drops significantly, i.e. RMSE increases about 87mm, compared to our full model. This demonstrates that surface normal is also helpful for outdoor depth completion.

Effect of Attention Based Integration. We then disable the attention based integration to verify the necessity of the two-pathway combination, *i.e.* only considering the depth from the normal pathway. Without this integration, all the evaluation metrics drop (Tab. 3 "-Attention Integration") compared to the full model. Fig. 7 (row w_c, w_n) shows the attention map learned automatically for color pathway and surface normal pathway. It can be seen that surface normal pathway works better (i.e. higher weight) in close range but gets worse when the distance goes up, which is consis-
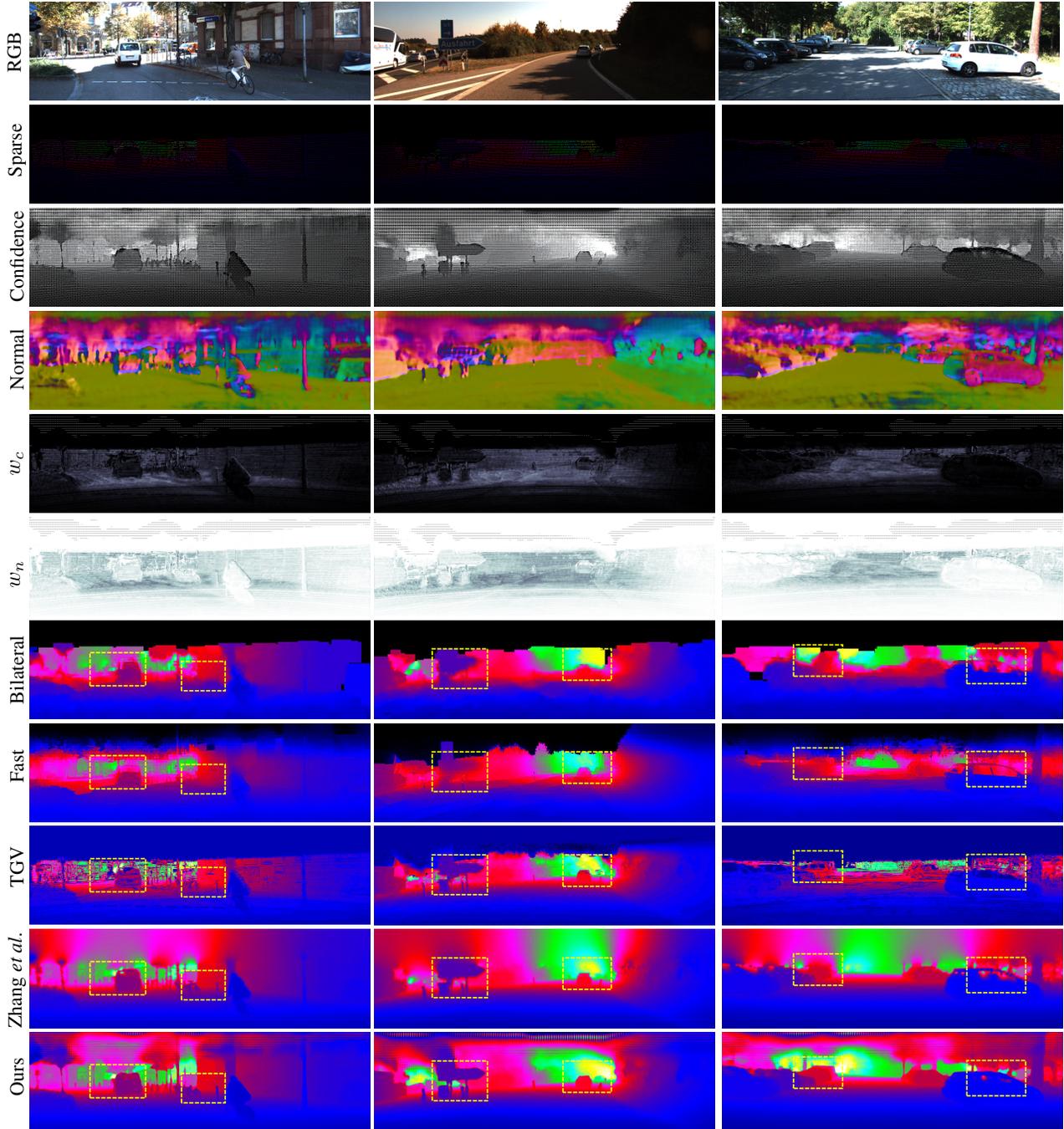
Figure 7. **Qualitative results on KITTI validation set.** From top to bottom are RGB image input, sparse depth input, confidence mask, estimated surface normals, attention map for color pathway, attention map for normal pathway, results of Bilateral [44], Fast [2], TGV [12] Zhang *et al.* [55], and our method. We mark some regions in the results to highlight the difference across methods.

tent with our analysis. In contrast, the color pathway cannot capture accurate details in close range compared to the surface normal pathway but better in far distance. Although the color pathway works better for fewer regions compared to the surface normal pathway, it is critically important to achieve good performance in far area, where large error are more likely to happen.

Effect of Deep Completion Unit. We also replace our deep completion unit to a traditional encoder-decoder architecture with early fusion, where the input color image, sparse depth, and a binary mask are concatenated at the beginning and fed as input to the network. This modification causes significant performance drop even with all the other components of the model enabled (Tab. 3 "-DCU"). Notice that we
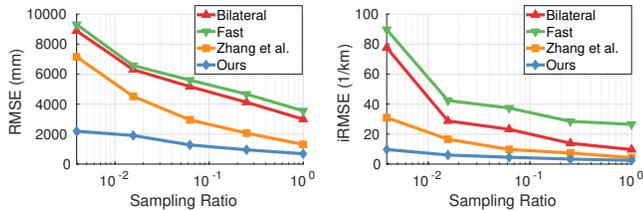
Figure 8. **Performance with different sparsity.** We test our model on input depth with different sparsity by sub-sampling the raw LiDAR. Our method works well even with very sparser depth input, and outperforms the other methods.

sum the features from the sparse depth encoder with the features from decoder rather than the ordinary concatenation. We also tried the concatenation option which however takes more memory and produces slightly worse performance.

Effect of Confidence Mask. Last but not least, we disable the confidence mask by replacing the learned one with a typical binary mask indicating the availability of sparse depth per-pixel. This causes dramatic increase of RMSE by 69mm compared to the full model. In contrast, our full model learns confidence masks according to the inputs, which provide extremely useful information about the reliability of the input sparse depth for the surface normal pathway, as shown in Fig. 5(d) and Fig. 7(Confidence). As can be seen, the area with overlapping depth from foreground and background are generally marked with low confidence. Notice that these areas usually happens on the boundary of the foreground where occlusion happens.

4.3. Generalization Capability

Even though in this paper we especially focus on producing dense depth for car-held LiDAR devices, our model can be considered as a general depth completion approach. Therefore, we investigate the generalization capability of our model under different situations, specifically with different input depth sparsity and in indoor environment.

Robustness Against Depth Sparsity. It is interesting to see if our model could still work on more challenging cases where input depths are even sparser. The raw LiDAR depth provided by the benchmark is roughly 18,400 samples per depth image of 1216 by 352 resolution, i.e. 4.3% of the pixels having depth. We uniformly sub-sample the raw LiDAR depth by ratios of $1/4$, $1/16$, $1/64$, and $1/256$, which correspond to 1.075%, 0.269%, 0.0672%, and 0.0168% of pixels having depth. It worth noting that 0.0168% corresponds to 72 pixels per depth image. This is an extreme hard case where the scene structure is almost missing from the input sparse depth.

The performances of our model and other methods [44, 2, 55] on LiDAR with different sparsity are shown in Fig. 8. We can see the performances are better (i.e. lower RMSE) with more input sparse depth. Our method still performs reasonably well even for the most challenging case (i.e. 0.0168%). Actually our result under this case is still

	RMSE↓	REL↓	$\delta_{1.25}$ ↑	$\delta_{1.25^2}$ ↑	$\delta_{1.25^3}$ ↑
Bilateral [44]	0.479	0.084	92.4	97.6	98.9
TGV [12]	0.635	0.123	81.9	93.0	96.8
Ma *et al*. [34]	0.230	0.044	97.1	99.4	99.8
Zhang *et al*. [55]	0.228	0.042	97.1	99.3	99.7
CSPN [6]	0.117	**0.016**	99.2	**99.9**	**100**
Ours	**0.115**	0.022	**99.3**	**99.9**	**100**

Table 4. **Performance of depth completion on NYU v2 dataset [44].** We compare to non-learning based approaches [44, 12] and deep learning based methods [34, 6, 55]. Our method performs the best on the main RMSE metric (in meter), and performs on-par with the state-of-art method on other metrics.

better than results of the traditional methods [44, 2] with full sparse data (i.e. 4.3%).

Depth Completion in Indoor Scenes. We also evaluate our model for indoor scenes on NYUv2 dataset [44]. Adopting similar experiment setting as [6, 34], we synthetically generate sparse depth via random sampling, train on 50K images sampled from the training set, and evaluate on the official labeled test set (containing 654 images). Images are down-sampled to half resolution and center-cropped to 304×228. The same metrics are adopted, including root mean square error (RMSE), mean absolute relative error (REL), and the percentage δ_t of pixels with both the relative error and inverse of it under a certain threshold t (where $t = 1.25, 1.25^2, 1.25^3$). The quantitative comparisons are listed in Tab. 4. The numbers for Bilateral [44], Ma *et al*. [34], and CSPN [6] are obtained from CSPN [6]. The numbers for the other methods are obtained using their released implementations. Even not designed specifically for indoor environment, our method still achieve comparable or better performance compared to the state-of-the-art (rank top for 4 out of 5 metrics). Please refer to supplementary materials for more qualitative results.

5. Conclusion

In this paper, we propose an end-to-end neural network for depth prediction from Sparse LiDAR data and a single color image. We use the surface normal as the intermediate representation directly in the network and demonstrate it is still effective for outdoor scene similar as the indoor scene. We propose a deep completion unit to better fuse the color image with the sparse input depth. We also analyze specific challenges for the outdoor scene, and provide solutions within the network architecture, such as attention based integration to improve performance in far distance and estimating a confidence mask for occlusion handling. Extensive experiments show that our method achieves the state-of-art performance on the benchmark, and generalizes well to sparser input and indoor scenes.

8

References

[1] J. T. Barron and J. Malik. Intrinsic scene properties from a single rgb-d image. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 17–24, 2013.

[2] J. T. Barron and B. Poole. The fast bilateral solver. In *Proc. of the European Conf. on Computer Vision (ECCV)*, pages 617–632, 2016.

[3] A. Bhandari, A. Kadambi, R. Whyte, C. Barsi, M. Feigin, A. Dorrington, and R. Raskar. Resolving multi-path interference in time-of-flight imaging via modulation frequency diversity and sparse regularization. *CoRR*, 2014.

[4] W. Chen, D. Xiang, and J. Deng. Surface normals in the wild. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, pages 1557–1566, 2017.

[5] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. Multi-view 3d object detection network for autonomous driving. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1907–1915, 2017.

[6] X. Cheng, P. Wang, and R. Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *Proc. of the European Conf. on Computer Vision (ECCV)*, pages 108–125, 2018.

[7] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. CARLA: An open urban driving simulator. In *Proc. of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.

[8] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, pages 2650–2658, 2015.

[9] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2366–2374, 2014.

[10] A. Eldesokey, M. Felsberg, and F. S. Khan. Propagating confidences through cnns for sparse data regression. In *The British Machine Vision Conference (BMVC)*, 2018.

[11] S. R. Fanello, J. Valentin, C. Rhemann, A. Kowdle, V. Tankovich, P. Davidson, and S. Izadi. Ultrastereo: Efficient learning-based matching for active stereo systems. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 6535–6544, 2017.

[12] D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, pages 993–1000, 2013.

[13] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 270–279, 2017.

[14] X. Gong, J. Liu, W. Zhou, and J. Liu. Guided depth enhancement via a fast marching method. *Image and Vision Computing*, 31(10):695–703, 2013.

[15] Y. Han, J.-Y. Lee, and I. So Kweon. High quality shape from a single rgb-d image under uncalibrated natural illumination. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, pages 1617–1624, 2013.

[16] S. Hawe, M. Kleinsteuber, and K. Diepold. Dense disparity maps from sparse disparity measurements. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, pages 2126–2133, 2011.

[17] K. He, J. Sun, and X. Tang. Guided image filtering. In *Proc. of the European Conf. on Computer Vision (ECCV)*, pages 1–14, 2010.

[18] D. Herrera, J. Kannala, J. Heikkilä, et al. Depth map inpainting under a second-order smoothness prior. In *Proc. of the Scandinavian Conference on Image Analysis*, pages 555–566. Springer, 2013.

[19] Z. Huang, J. Fan, S. Yi, X. Wang, and H. Li. Hms-net: Hierarchical multi-scale sparsity-invariant network for sparse depth completion. *arXiv preprint arXiv:1808.08685*, 2018.

[20] M. Jaritz, R. De Charette, E. Wirbel, X. Perrotton, and F. Nashashibi. Sparse and dense data with cnns: Depth completion and semantic segmentation. In *Proc. of International Conf. on 3D Vision (3DV)*, pages 52–60, 2018.

[21] K. Karsch, C. Liu, and S. B. Kang. Depth extraction from video using non-parametric sampling. In *Proc. of the European Conf. on Computer Vision (ECCV)*. Springer, 2012.

[22] K. Karsch, C. Liu, and S. B. Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 36(11):2144–2158, 2014.

[23] M. Kiechle, S. Hawe, and M. Kleinsteuber. A joint intensity and depth co-sparse analysis model for depth map super-resolution. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, pages 1545–1552, 2013.

[24] J. Konrad, M. Wang, and P. Ishwar. 2d-to-3d image conversion by learning depth from examples. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 16–22, 2012.

[25] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *Proc. of International Conf. on 3D Vision (3DV)*, pages 239–248, 2016.

[26] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1119–1127, 2015.

[27] D. Lingfors, J. M. Bright, N. A. Engerer, J. Ahlberg, S. Killinger, and J. Widén. Comparing the capability of low- and high-resolution lidar data with application to solar resource assessment, roof type classification and shading analysis. *Applied Energy*, 205:1216–1230, 2017.

[28] F. Liu, C. Shen, G. Lin, and I. Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 38(10):2024–2039, 2016.

[29] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[30] L.-K. Liu, S. H. Chan, and T. Q. Nguyen. Depth reconstruction from sparse samples: Representation, algorithm, and sampling. *IEEE Trans. on Image Processing (TIP)*, 24(6):1983–1996, 2015.

[31] M. Liu, M. Salzmann, and X. He. Discrete-continuous depth estimation from a single image. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 716–723, 2014.

[32] J. Lu and D. Forsyth. Sparse depth super resolution. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2245–2253, 2015.

[33] F. Ma, G. V. Cavalheiro, and S. Karaman. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. *arXiv preprint arXiv:1807.00275*, 2018.

[34] F. Ma and S. Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *Proc. IEEE International Conf. on Robotics and Automation (ICRA)*, 2018.

[35] O. Mac Aodha, N. D. Campbell, A. Nair, and G. J. Brostow. Patch based synthesis for single depth image super-resolution. In *Proc. of the European Conf. on Computer Vision (ECCV)*, pages 71–84, 2012.

[36] M. Mahmoudi and G. Sapiro. Sparse representations for range data restoration. *IEEE Trans. on Image Processing (TIP)*, 21(5):2909–2915, 2012.

[37] N. Naik, A. Kadambi, C. Rhemann, S. Izadi, R. Raskar, and S. Kang. A light transport model for mitigating multipath interference in TOF sensors. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[38] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016.

[39] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 283–291, 2018.

[40] A. Roy and S. Todorovic. Monocular depth estimation using neural regression forest. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5506–5514, 2016.

[41] S. Ryan Fanello, C. Rhemann, V. Tankovich, A. Kowdle, S. Orts Escolano, D. Kim, and S. Izadi. Hyperdepth: Learning depth from structured light without matching. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5441–5450, 2016.

[42] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1161–1168, 2006.

[43] E. Shabaninia, A. R. Naghsh-Nilchi, and S. Kasaei. High-order markov random field for single depth image super-resolution. *IET Computer Vision*, 2017.

[44] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *Proc. of the European Conf. on Computer Vision (ECCV)*, pages 746–760, 2012.

[45] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag New York, Inc., New York, NY, USA, 1st edition, 2010.

[46] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, pages 839–846, 1998.

[47] I. Tosic and S. Drewes. Learning joint intensity-depth sparse representations. *IEEE Trans. on Image Processing (TIP)*, 23(5):2122–2132, 2014.

[48] J. Uhrig, N. Schneider, L. Schneidre, U. Franke, T. Brox, and A. Geiger. Sparsity invariant cnns. In *Proc. of International Conf. on 3D Vision (3DV)*, 2017.

[49] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. Conditional image generation with pixel-cnn decoders. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4790–4798, 2016.

[50] B. Wang, Y. Feng, and H. Liu. Multi-scale features fusion from sparse lidar data and single image for depth completion. *Electronics Letters*, 2018.

[51] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia. Lego: Learning edge with geometry all at once by watching videos. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 225–234, 2018.

[52] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018.

[53] L.-F. Yu, S.-K. Yeung, Y.-W. Tai, and S. Lin. Shading-based shape refinement of rgb-d images. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1415–1422, 2013.

[54] H.-T. Zhang, J. Yu, and Z.-F. Wang. Probability contour guided depth map inpainting and superresolution using non-local total generalized variation. *Multimedia Tools and Applications*, pages 1–18, 2017.

[55] Y. Zhang and T. Funkhouser. Deep depth completion of a single rgb-d image. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 175–185, 2018.

[56] Y. Zhang, S. Song, E. Yumer, M. Savva, J.-Y. Lee, H. Jin, and T. Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5287–5295, 2017.

[57] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1851–1858, 2017.

[58] Y. Zuo, Q. Wu, J. Zhang, and P. An. Explicit edge inconsistency evaluation model for color-guided depth map enhancement. *IEEE Trans. on Circuits and Systems for Video Technology*, 28(2):439–453, 2018.