IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 September 2022, accepted 8 October 2022, date of publication 19 October 2022, date of current version 4 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3215546

RESEARCH ARTICLE

NNNet: New Normal Guided Depth Completion From Sparse LiDAR Data and Single Color Image

JIADE LIU AND CHEOLKON JUNG, (Member, IEEE)
School of Electronic Engineering, Xidian University, Xi'an 710071, China
Corresponding author: Cheolkon Jung (zhengzk@xidian.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61872280 and Grant 62111540272.

ABSTRACT In this paper, we propose new normal guided depth completion from sparse LiDAR data and single color image, named NNNet. Sparse depth completion often uses normal maps as a constraint for model training. However, direct construction of a normal map from the color image causes a lot of noise in the normal map and reduces the model performance. Thus, we use a new normal map as an intermediate constraint to promote the fusion of multi-modal features. We generate the new normal map from the sparse LiDAR depth data to use it as a constraint for network training. The new normal map is generated by converting the input depth into a grayscale image, constructing a normal map, replacing the Z channel of the normal map with the original depth, and finally adding a mask. Based on the new normal map, we construct an end-to-end network NNNet for sparse depth completion guided by its corresponding color image. NNNet consists of two branches. The one branch generates the new normal map from the depth image and its corresponding color image, while the other branch constructs a dense depth image from the sparse depth and the predicted new normal map. The two branches fully merge the features through skip connection. In loss function, we use L2 loss to ensure that the new normal map plays a restrictive role. Finally, we generate the dense depth image by refining it with a spatial propagation network. Experimental results show that the new normal map provides effective constraints for sparse depth completion. Moreover, NNNet achieves 724.14 in terms of RMSE and outperforms most of the current state-of-the-art methods.

INDEX TERMS Sparse depth completion, convolutional neural network, deep learning, LiDAR, normal map, refinement.

I. INTRODUCTION

In recent years, autonomous driving has become a hot issue of high concern. For autonomous driving and robotics, dense and accurate depth images are of great significance. In indoor scenes, due to the low degree of passive lighting interference, the depth camera can obtain dense and accurate depth images. However, in outdoor scenes, dense depth perception mainly relies on stereo vision or LiDAR sensors. Stereo vision [1], [2] has many limitations in practice, and the depth measurement accuracy is low in long-distance areas. Recently, LiDAR provides the most accurate depth values, which is widely used in autonomous driving and robotics. However, the depth image projected by even the most high-end LiDAR camera is still highly sparse, and there is noise around object boundaries. Depth completion aims at generating a dense and accurate depth image from a sparse depth data with the guidance of its corresponding high-resolution color (RGB) image. The general pipeline of an autonomous vehicle can be split into 4 modules: sensing module, perception module [3], path planning module [4] and control module [5]. Sparse depth completion is a preliminary step for environmental perception. The acquisition of dense and accurate depth maps improves the accuracy of the perception module, while supporting the subsequent path planning and control.

Up to the present, many methods for sparse depth completion have been proposed. With the great success of deep learning, deep neural networks are used to deal with this problem. The sparse depth completion methods based on deep neural networks are divided into two categories: Depth-only

The associate editor coordinating the review of this manuscript and approving it for publication was Utku Kose.

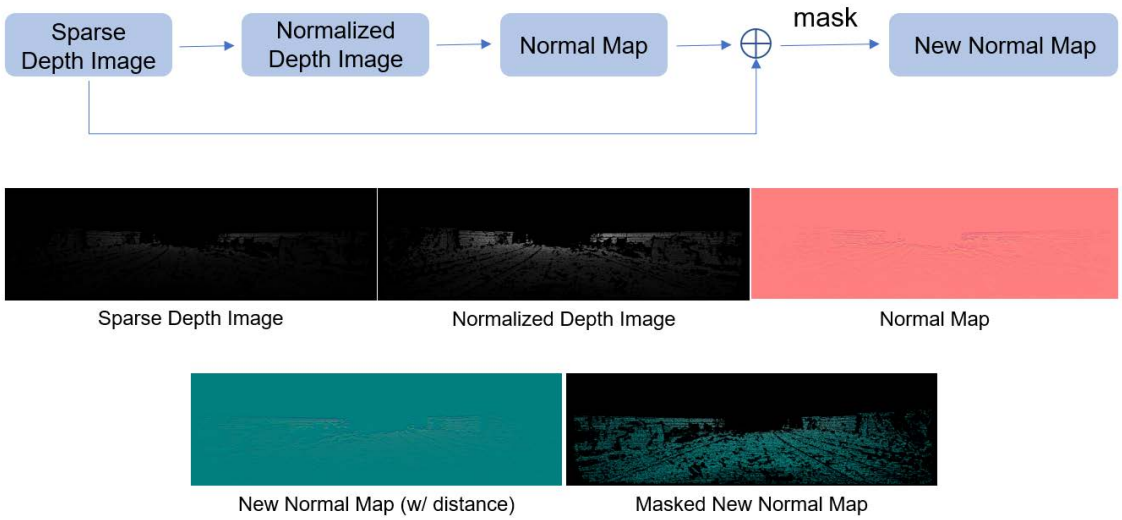


FIGURE 1. Illustration of the new normal map construction. In the first step, we process the depth image into a grayscale image to amplify the relative relationship between pixels. In the second step, we construct a normal map from the grayscale image. In the third step, we replace the Z channel of the normal map with the original depth image. Finally, we add a mask to generate the new normal map.

methods [6], [7], [8] and RGB-guided schemes [9], [10], [11]. Due to the sparseness of the input LiDAR depth, the depth-only methods have difficulties in recovering semantically consistent boundaries and depth for small and thin objects. In recent studies, the RGB-guided schemes have attracted more and more attention because it can provide richer structural information and semantic information. However, the fusion of multimodal features also causes difficulties and challenges [8], [10]. A common approach to RGB-guided depth completion is to use normal maps as a constraint for model training. Zhang et al. [12] estimated surface normal as an intermediate representation, and achieved outstanding performance in indoor scenes. Qiu et al. [10] further proposed an end-to-end network that used surface normal as an intermediate representation and generated dense and accurate depth images in outdoor road scenes. However, we have found that constructing a normal map directly from the color image causes a lot of noise in the normal map or need additional datasets to pre-train the model. The normal map constructed from the depth image is smoother and more accurate. Inspired by this phenomenon, we consider the possibility of constructing a normal map from the sparse depth image as shown in Fig. 1.

We analyze the information contained in the three channels of the normal map. The X channel contains the relative left and right information of the image, the Y channel contains the relative top and bottom information of the image, and the Z channel contains the relative front and back information of the image. We consider replacing the Z channel of the normal map with the original depth value can not only save the relative front and rear information of the image, but also add the accurate front and rear information of the image. Therefore, we construct a new normal map from the sparse depth image. The new normal map represents the enhanced relative relationship between depth points while retaining accurate depth values, thus providing effective constraints for model training. Based on the new normal map, we propose a depth completion network based on convolutional neural networks (CNNs), named NNNet. Specifically, one branch of the network model constructs a new normal map from the RGB and depth images. As an intermediate constraint, the new normal map promotes the fusion of multi-modal features. The other branch constructs a dense depth image from the sparse depth image and the predicted new normal map. Since the number of the input channels in the two branches are the same, the structure of the two branches is the same that forms a pseudo-siamese network and is conducive to the fusion of features between the branches. The two branches fully merge the features through skip connections. Finally, we generate the dense depth image by refining it with a spatial propagation network. Experimental results demonstrate that NNNet achieves better performance on the test set (RMSE: 724.14) than the validation set (RMSE: 757.05), which indicates its good generalization ability. Fig. 2 illustrates the network structure of NNNet.

Compared with existing methods, the main contributions of this paper are summarized as follows:

- We propose a new normal map for sparse depth completion. We generate the new normal map from the ground truth dense depth image and use it as a constraint for network training. The new normal map represents the enhanced relative relationship between depth points while retaining accurate depth values. Thus, the new normal map provides an effective constraint for network training.
- We construct a depth completion network based on CNN that integrates the features of the input color and depth images with the help of the new normal map, named NNNet. NNNet consists of two branch-CNNs and one optimization module. The two branch-CNNs have the

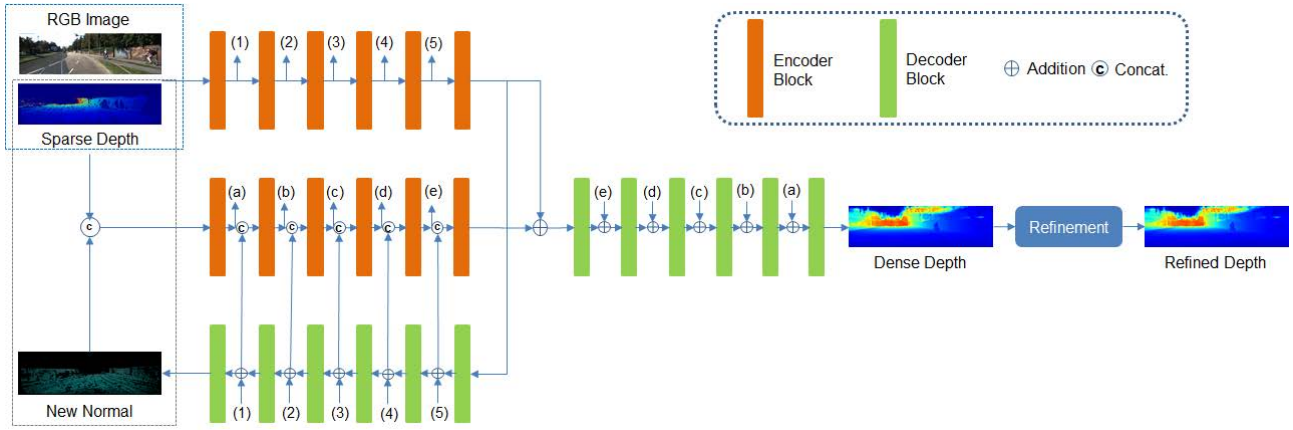


FIGURE 2. Network structure of NNNet. NNNet consists of three parts: The first part takes the RGB image and the sparse depth image as input and generates the new normal map. The second part takes the new normal map and the sparse depth image as input, and estimates its dense depth image. The first and second parts perform feature fusion through skip connections, while the third part refines the dense depth image based on a refinement module.

same structure that forms a pseudo-siamese network, which is conducive to the fusion of features between the branches.

II. RELATED WORK

This section introduces the related work of depth completion, including depth estimation from a single RGB image, depth completion of a single sparse depth image, and sparse depth completion based on RGB image guidance.

A. DEPTH ESTIMATION FROM A SINGLE RGB IMAGE

In recent years, depth estimation from a single RGB image has attracted considerable research interest. Saxena et al. [13] carried out early research on estimating depth from a single RGB image. Karsch et al. [14] tried to use the whole depth images in the training set to produce more consistent image level prediction. Eigen et al. [15] proposed a multi-scale model for depth prediction. Laina et al. [16] proposed a deeper fully convolutional architecture on a single scale. Ladicky et al. [17] proposed using semantic information to improve the accuracy of depth estimation. Chen et al. [18] proposed using sparse surface annotation to supervise model training. Liu et al. [19] designed a deep convolution neural network to learn unary and paired terms. Qi et al. [20] focused on depth estimation of indoor scenes, designed a two branch convolutional neural network to jointly predict depth and surface normals. Although these methods can produce reasonable depth estimation, they are not suitable for restoring high-precision depth. By using the new normal as an intermediate constraint, NNNet successfully integrates the multi-dimensional information of RGB image and the depth information in sparse depth image, and has achieved good performance in predicting high-precision depth.

B. DEPTH COMPLETION OF A SINGLE SPARSE DEPTH IMAGE

Due to the demands for low-cost LiDAR, predicting a dense and accurate depth image from a single sparse depth image has attracted much attention. To achieve this, researchers use sparse depth images or low-resolution depth images as input to reconstruct high-resolution depth images. In the early stage, researchers used compressed sensing theory [21] or wavelet analysis [22], [23] to generate dense depth images. Fast [24] assumed that smooth regions in an image are closely related internally and claimed that there are consistency or gradual changes in computing these regions. TGV [25] formulated a convex optimization problem for depth upsampling using higher order regularization. Moreover, Ku et al. [26] transformed sparse depth images into dense ones by a series of operators including dilation, hole closure, hole filling, and blurring. The early work mainly focused on bilateral filtering or global energy optimization. In recent years, deep learning is used for sparse depth completion. Uhrig et al. [6] proposed sparsity invariant CNNs to deal with depth images at different degrees of sparsity. Eldesokey et al. [27] generated a full depth image and a confidence map with normalized convolution to predict dense depth images. Jaritz et al. [28] applied semantic segmentation to depth completion. Chodosh et al. [21] combined compressive sensing and deep learning into depth prediction. Cheng et al. [11] guided depth interpolation through a recurrent neural network by using an affinity matrix. From the perspective of depth super-resolution, some methods use dictionary learning to deal with this problem. There are also some methods that use pairs of low-resolution and high-resolution depth image databases [29] or self-similarity search [30] to generate high-resolution depth images. Riegler et al. [31] proposed a deep network to produce a high-resolution depth image robust to depth discontinuities and used a variational model to refine the depth image. Unlike the depth super-resolution methods, NNNet deals with highly sparse depth images and does not require additional data sets or manual operation.

C. IMAGE-GUIDED SPARSE DEPTH COMPLETION

Image-guided sparse depth completion methods often achieve good performance because RGB images can provide

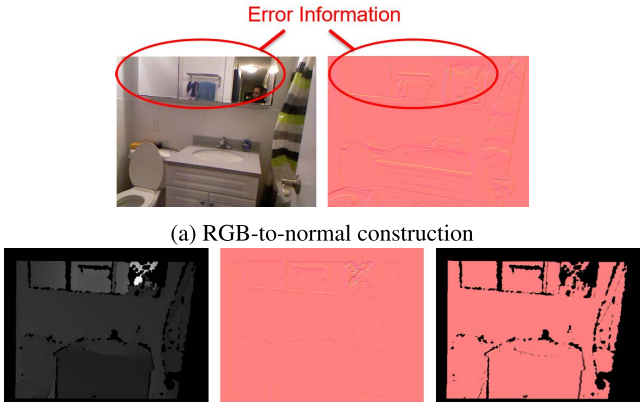


FIGURE 3. Normal map comparison. The RGB-to-normal construction easily causes wrong information, while the depth-to-normal construction provides smooth and accurate information.

more structural and semantic information. The early work mainly focused on bilateral filtering [32] or global energy optimization [33]. Later, edge information guidance [34], image guidance [32] and surface normal guidance [35] were used. Ma et al. [36] fed the concatenation of the sparse depth image and the color image into an encoder-decoder network, and further extended with self-supervised learning [7]. Qiu et al. [10] used surface normal as guidance in outdoor scenes and recovered dense depth image from sparse LiDAR data. Huang et al. [37] proposed three sparsity invariant operations to solve this problem. Eldesokey et al. [8] combined their confidence propagation with RGB information to predict dense depth images. Gansbeke et al. [38] used a two branch convolutional neural network to predict depth and learn an uncertainty to fuse two results.

Although RGB image guidance provides rich information for spare depth completion, it still needs to efficiently fuse RGB image features and sparse depth features which is a new challenge for researchers. In the past methods, it is often only a simple fusion of image features using skip connections. Inspired by the previous work that used surface normal as an intermediate constraint, we propose to construct new normal maps for sparse depth completion. On the basis of the new normal maps, we design a network architecture that effectively integrates multi-modal features for sparse depth completion.

III. PROPOSED METHOD

We first construct a new normal map for NNNet from the groundtruth dense depth image as illustrated in Fig. 1. The new normal map constructed from the depth image contains more information than the original depth image. NNNet is an end-to-end model which is divided into three parts as shown in Fig. 2. The first part takes the RGB image and the sparse depth image as input and generates the new normal map. The second part takes the new normal map and the sparse depth image as input, and estimates its dense depth image. The first and second parts perform feature fusion through skip connections, while the third part refines the dense depth image based on a refinement module.

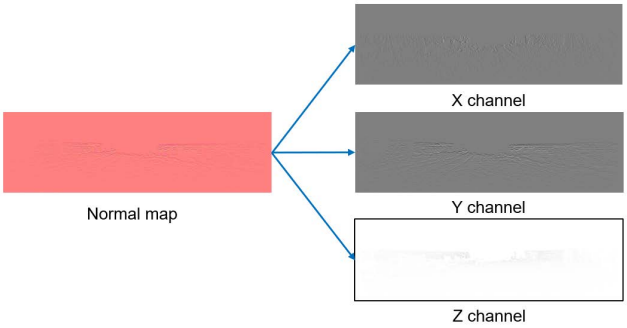


FIGURE 4. Each channel analysis in the normal map. The X channel of the normal map contains the left and right information of the image, the Y channel contains the top and bottom information of the image, and the Z channel contains the relative frontal and back information of the image.

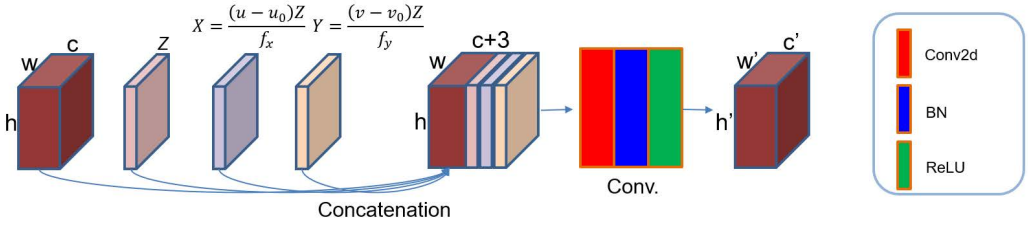
A. NEW NORMAL CONSTRUCTION

Inspired by the use of normal maps as intermediate constraints in the previous work [10], we explore the normal map construction from depth images, instead of RGB images. As shown in Fig. 3a, we compare two normal maps generated by RGB and depth images. The normal map generated directly from the RGB image contains errors. They are from the textures of RGB image, resulting in wrong normal planes. Zhang et al. [12] used CNNs to obtain a more accurate normal map from RGB images, but it requires additional data sets to pre-train their models. Thus, we try to construct a normal map directly from the depth image. As shown in Fig. 3b, the normal map constructed from the depth image is smoother and more accurate.

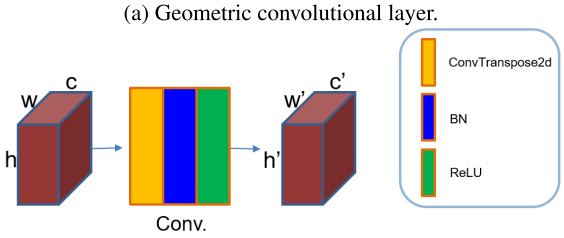
On the basis of the above experimental results, we continue to analyze each channel in the normal map. As shown in Fig. 4, X channel of the normal map contains the left and right information of the image, Y channel contains the top and bottom information of the image, and Z channel contains the relative front and back information of the image. We can find that the relative relationship between the front and back contained in the Z channel is weak, and the original accurate depth values are lost. Therefore, we propose to construct a new normal map. As shown in Fig. 1, we get a grayscale depth image from the input depth image to strengthen the relative relationship between the pixels. Then, we construct a normal map from the grayscale image. Next, we change the Z channel of the normal map to the original depth values, i.e. new normal map. Thus, the new normal map preserves the relative back-to-back relationship between pixels while keeping the accurate original depth values in the new normal map. Finally, we add a mask from the groundtruth dense depth image to the new normal map.

B. NETWORK ARCHITECTURE

On the basis of the new normal map, we propose NNNet for sparse depth completion. The network architecture of NNNet



(a) Geometric convolutional layer.



(b) Deconvolutional layer.

FIGURE 5. Details of the geometric convolutional layer and deconvolutional layer. The geometric convolutional layer proposed by Hu et al. [39] concatenates the features and position information of the pixels as the input of the convolutional layer, and its effectiveness for sparse depth feature extraction has been verified. The deconvolution layer is followed by one BN layer and one ReLU layer.

is divided into two branches. The one generates a new normal map from the sparse depth image and the RGB image, while the other generates a dense depth image from the sparse depth image and the new normal map. Since the number of input channels is the same, the network structure of the two branches is completely the same forming a pseudo-siamese network, which promotes the feature fusion of two branches. NNNet is an encoder-decoder structure. The encoder consists of one convolutional layer and five geometric convolutional layers. The geometric convolutional layer proposed by Hu et al. [39] is effective for extracting sparse depth features. As shown in Fig. 5a, the geometric convolutional layer inputs the extracted features and the coordinates of the pixels into the convolutional layer to obtain new features. The geometric convolutional layers include one BN layer and one ReLU layer. Correspondingly, the decoder consists of five deconvolution layers and one convolution layer. As shown in Fig. 5b, the deconvolutional layer is followed by one BN layer and one ReLU layer. We use skip connections to fully integrate the features between different branches. NNNet runs in two stages. During the first stage of operation, NNNet extracts features from the RGB image and the sparse depth to construct a new normal map. Among them, the encoded feature and the decoded feature are merged by skip connections. In the second stage, NNNet fuses the decoding features of the first branch and the coding features of the second branch, thus effectively promoting the feature fusion of RGB image and sparse depth image. The new normal map is used as an intermediate constraint for network training to fuse the features of RGB and sparse depth images.

C. REFINEMENT

It has been reported that the predicted dense depth image may not retain effective values of the input depth image [11]. In response to this phenomenon, a large number of methods have been proposed. Chen et al. [43] proposed CSPN++ to deal with this issue. Hu et al. [39] further proposed an dilated and accelerated CSPN++ based on CSPN++. Thus, we use the basic refinement module that the depth value at a pixel is optimized by those around it. The basic refinement module is defined as:

$$D_i^{t+1} = W_{ii} D_i^0 + \sum_{j \epsilon N(i)} W_{ji} D_j^t \tag{1}$$

where D^0 is the dense depth image, D^t is the depth image obtained after t iterations, and W_{ji} is the affinity between pixel j and pixel i. For pixel i, in each iteration, we obtain information from the value of surrounding pixels $N(i)$ for refinement. As shown in Eq. (1), the refinement module learns the correlation among pixels and improves the prediction accuracy of the depth map. The refinement module is implemented by the spatial propagation network (SPN). SPN is proposed by Liu et al. [44] to learn local affinities. We use both Chen et al.'s method [43] and Hu et al.'s method [39] to implement the refinement module.

D. LOSS FUNCTION

We train NNNet in two stages. Corresponding to the two stages, we have two different loss functions. The loss function of the first stage L_{stage1} is defined as follows:

$$L_{stage1} = \lambda_1 \| (D_{pred1} - D_{gt}) \odot mask \|^2 + \lambda_2 \| (N_{pred1} - N_{gt}) \odot mask \|^2 \tag{2}$$

where λ is the hyper-parameter, and we empirically set λ_1 and λ_2 to 0.8 and 0.2, respectively; D_{pred1} is the predicted dense depth image; D_{gt} is the true value of the depth image; N_{pred1} is the new normal map predicted by the network; N_{gt} is the true value of the new normal map; and $\odot$ is an element-wise multiplication. As shown in Eq. (2), we do not use the

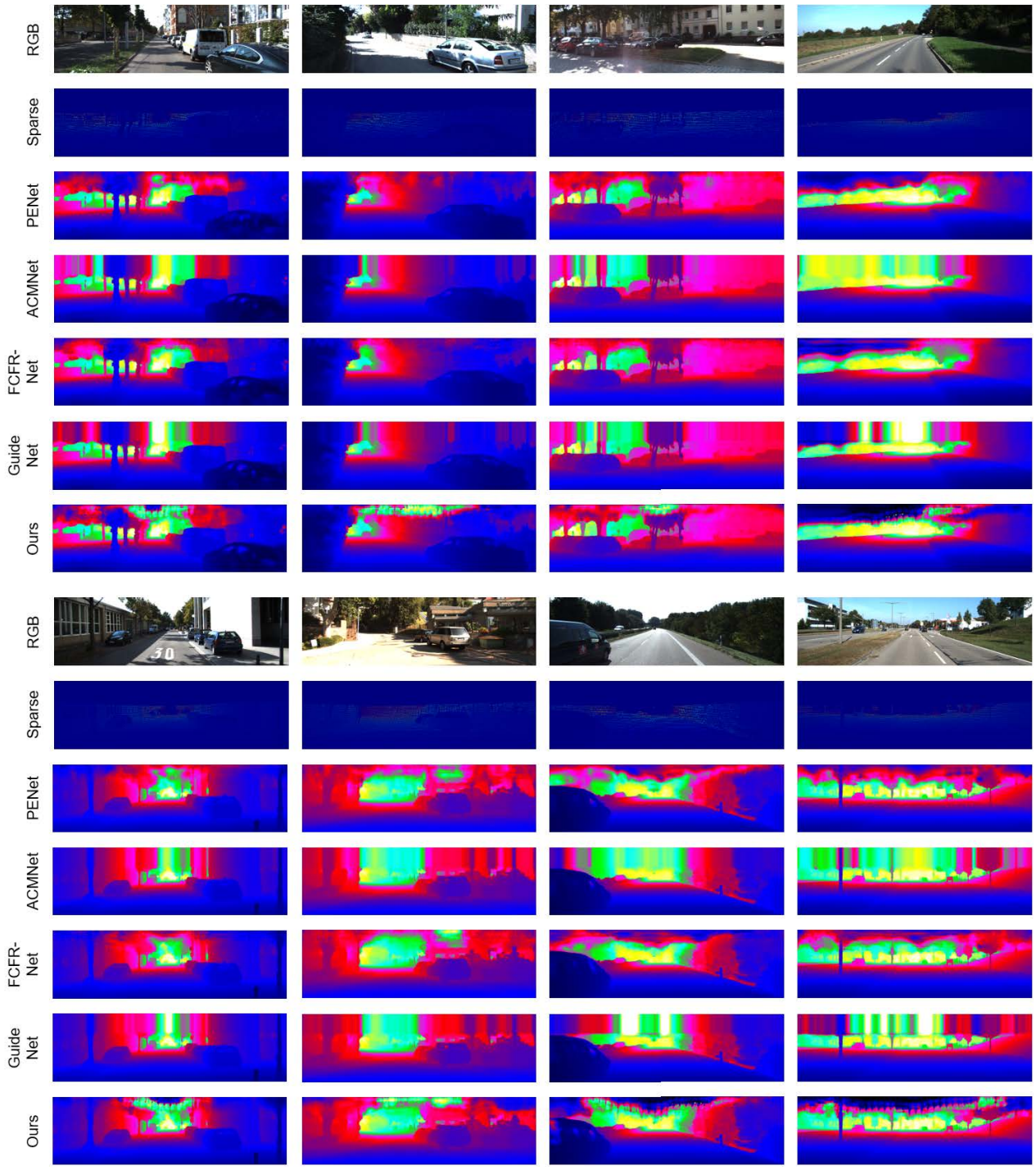


FIGURE 6. Visual comparison with state-of-the-art methods on KITTI test set. Top to bottom: RGB image, sparse depth image, PENet [39], ACMNet [40], FCFR-Net [41], GuideNet [42], and NNNet (Ours).

refinement module in the first stage, but use the depth map and the new normal map for model training.

The loss function of the second stage L_{stage2} is defined as follows:

$$L_{stage2} = \|(D_{pred2} - D_{gt}) \bigodot mask\|^2 \tag{3}$$

where D_{pred2} is the refined depth image optimized by the refinement module. As shown in Eq. (3), we use the refinement module in the second stage and only use the depth map to train the entire model.

IV. EXPERIMENTAL RESULTS

We perform various experiments and ablation studies to verify the effectiveness of NNNet. We have uploaded the model parameters of NNNet to the KITTI depth completion

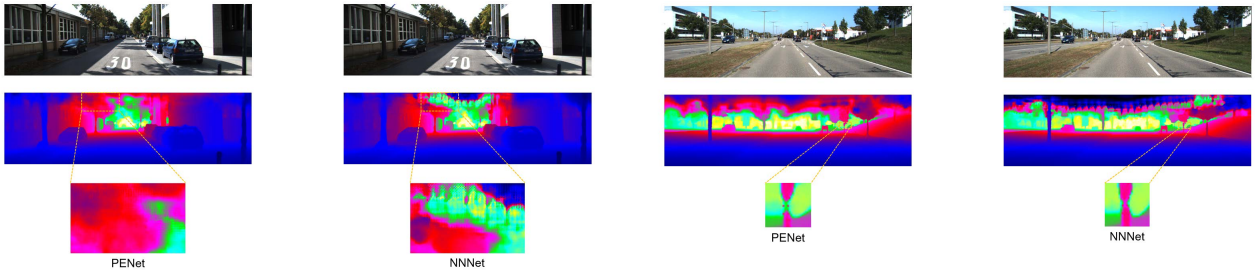


FIGURE 7. Visual comparison between PENet and NNNet. (a) shows that NNNet completes more accurate depth than PENet due to the influence of the new normal map. (b) shows that NNNet suppresses depth discontinuity more effectively than PENet.

evaluation,[1] and get the ranking. When we submit it (June 30, 2021), NNNet ranks 4th and now 14th among all. The RMSE value of NNNet on the KITTI test set is 724.14 and is higher than the RMSE value on the KITTI validation set (757.05), which indicates that NNNet has good generalization ability.

A. EXPERIMENTAL SETUP

1) DATASETS

To verify the effectiveness of NNNet, we generate the new normal maps according to the ground truth dense depth images in KITTI dataset. KITTI dataset contains color images and their corresponding sparse depth images with resolution 1216×352, in which the sparse depth image contains about 5% valid points and their dense depth images, i.e. ground truth, contain only about 16% of valid points. In the KITTI depth completion, there are 86898 images in the training set, 1000 images in the verification set and 1000 images in the test set. For comparison, we use a test set of 1000 images. In the ablation experiment, we use a validation set of 1000 images.

2) EVALUATION METRICS

According to the KITTI depth completion evaluation, we use four standard evaluation metrics: root mean squared error (RMSE, unit: mm), mean absolute error (MAE, unit: mm), root mean squared error of the inverse depth (iRMSE, unit: 1/km), and mean absolute error of the inverse depth (iMAE, unit: 1/km).

RMSE and MAE are defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x_i) - y_i)^2} \tag{4}$$

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}|h(x_i) - y_i| \tag{5}$$

Since RMSE in Eq. (4) reflects the error of outliers, we use it as an evaluation metric.

3) IMPLEMENTATION DETAILS

NNNet is implemented on Ubuntu 18.04 using Python 3.6.7 and PyTorch 1.4.0. We train NNNet on two NVIDIA Tesla V100 GPUs with batch size 8. During training, the input is cropped from the bottom to 320×1216. The whole training process is divided into two stages. In the first stage, we only train the first fusion network of two branches without the optimization module. We set the learning rate to 0.02 and halve the learning rate every 10 epochs. We perform total 30 epochs for training in the first stage. In the second stage, we implement end-to-end learning. When calculating the loss function, we only calculate the error between the estimated dense depth image and its ground truth. In the second stage, we halve the learning rate every 10 epochs. In the second stage, we perform total 45 epochs for training. Thus, the total number of epochs in training is 75. There is no significant change in performance after 75 epochs.

[1] http://www.cvlibs.net/datasets/kitti/eval_depth.php?benchmark=depth_completion

TABLE 1. Quantitative measurements and runtime comparison among different methods on the KITTI test set. The best performance is marked in bold. Although NNNet achieves better performance than the others, it also runs faster than most methods.

Method	RMSE	MAE	iRMSE	iMAE	Runtime
NNNet	**724.14**	**205.57**	**1.99**	**0.88**	0.034s
PENet[39]	730.08	210.55	2.17	0.94	**0.032s**
ACMNet[40]	732.99	206.80	2.08	0.90	0.080s
FCFR-Net[41]	735.81	217.15	2.20	0.98	0.130s
GuideNet[42]	736.24	218.83	2.25	0.99	0.140s
NLSPN[45]	741.68	199.59	**1.99**	0.84	0.220s
CSPN++[43]	743.69	209.28	2.07	0.90	0.200s
UberATG-FuseNet[46]	752.88	221.19	2.34	1.14	0.090s
DenseLiDAR[47]	755.41	214.13	2.25	0.96	0.020s
DeepLiDAR[10]	758.38	226.50	2.56	1.15	0.070s
DANConv[48]	759.65	213.68	2.17	0.92	0.050s

TABLE 2. Performance comparison among different methods on the KITTI validation set. NNNet is superior to the others in four metrics.

Method	RMSE	MAE	iRMSE	iMAE
Fast[24]	3548.87	1767.80	26.48	9.13
TGV[25]	2761.29	1608.69	15.02	6.28
DFusenet[49]	1240.00	429.00	-	-
Ma et al.[35]	858.00	311.00	-	-
PENet[39]	757.20	209.00	2.22	0.92
NNNet	**757.05**	**205.18**	**2.05**	**0.88**

B. COMPARISON WITH THE STATE-OF-THE-ART METHODS

We provide visual comparison in Fig. 6. Compared with the latest methods with similar indicators, the results of NNNet are not very different from them. The depth completion results by NNNet are affected by the relative relationship

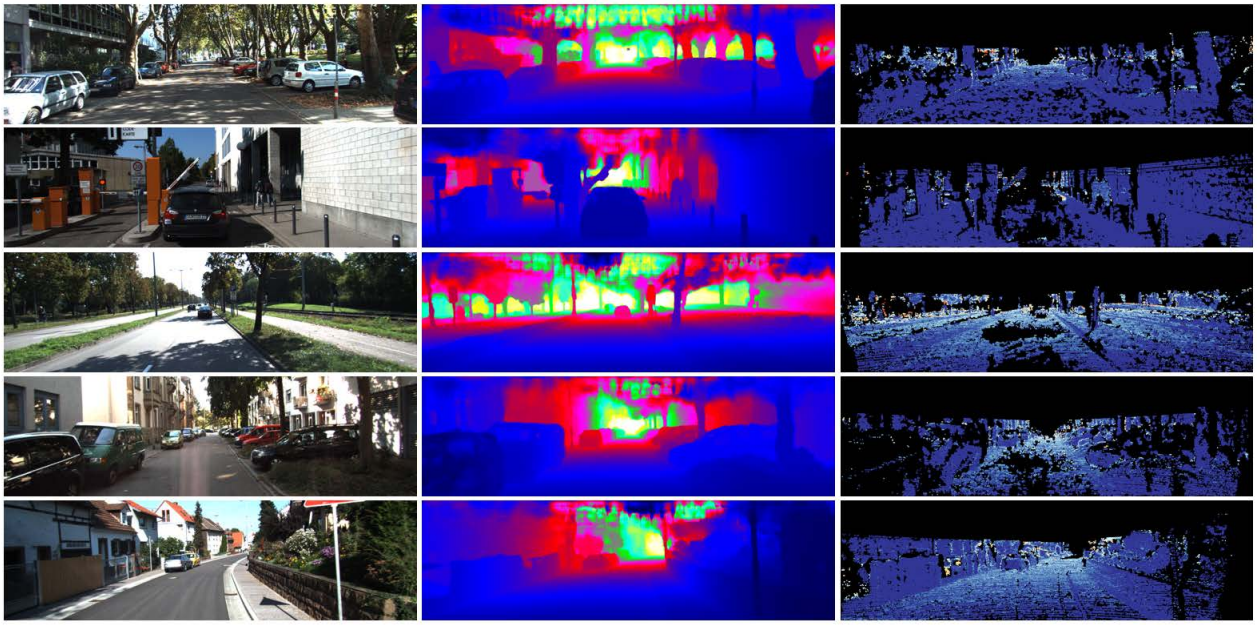


FIGURE 8. Error maps for five test images in the KITTI test set by NNNet. We obtain them from the KITTI depth completion evaluation.

between the pixels in the new normal image. The edge in the depth image is more sharp and tends to expand outward. This feature is beneficial for depth prediction of small objects at the distant area. As shown in Fig. 7, NNNet suppresses the discrete point phenomenon to some extent. When there are discrete depth points on the long-distance trunk, NNNet can learn the relative relationship between each depth point from the corresponding new normal map to complete the depth on the trunk. NNNet ranks the fourth in the KITTI depth completion at the time of submission (now the seventh among all methods) and ranks the first among all publications. In Table 1, we provide the quantitative measurements among different methods. We compare the latest and best methods in 10 publications. As shown in Table 1, NNNet has a greater improvement in RMSE than the others. Moreover, NNNet takes less runtime (0.034s) and performs better than most methods. As shown in Table 2, we further compare NNNet with state-of-the-art methods on KITTI validation set. The data in Table 2 are partly from our experimental results and partly from literature. To make comparison comprehensive, we also select some methods of not using deep learning. From the table, it can be observed that the RMSE value of NNNet on the test set is higher than that the RMSE value on the validation set. In terms of RMSE, NNNet is close to PENet, but it performs better than PENet in the other metrics. NNNet is superior to the others in four metrics. Fig. 8 shows error maps for five test images in the KITTI test set by NNNet (Test Image 0 to Test Image 4). We obtain them from the KITTI depth completion evaluation.[2]

[2]http://www.cvlibs.net/datasets/kitti/eval_depth.php?benchmark=depth_completion

TABLE 3. Quantitative measurements on different constraints on the KITTI validation set. Original: Baseline method without the refinement module. Normal: Normal map constraint. Sparse: Sparse depth constraint. NNNet: New normal map constraint.

Method	RMSE	MAE	iRMSE	iMAE
Original	793.16	217.98	-	-
Normal	781.97	219.25	2.36	0.98
Sparse	774.17	213.46	2.19	0.92
NNNet	**757.05**	**205.18**	**2.05**	**0.88**

C. ABLATION STUDIES

As shown in Fig. 9, we perform ablation experiments on each module of NNNet. First, we remove the optimization module. It can be observed that the NNNet performance is greatly reduced. Second, we use normal map as the intermediate constraint, and its performance is also reduced a lot. The ablation experiments show that without accurate depth guidance, the use of the normal map as the intermediate constraint can not fuse the features of RGB and sparse depth image well. We also take the depth image as the intermediate constraint, and the performance is also reduced. The results show that the depth image lacks the relationship between depth points and also performs poorly in feature fusion. Finally, the proposed new normal map not only effectively captures the relative relationship between depth points, but also successfully estimates accurate depth values. As the intermediate constraint, the new normal map enables NNNet to successfully guide feature fusion. As shown in Table 3, we perform a series of ablation experiments on KITTI validation set to verify the effectiveness of NNNet. First, we remove the optimization module and the performance of NNNet decreases significantly. Second, we verify that the new normal map plays an effective role in constraint. Since the new normal map is our main contribution to depth completion, we feed the

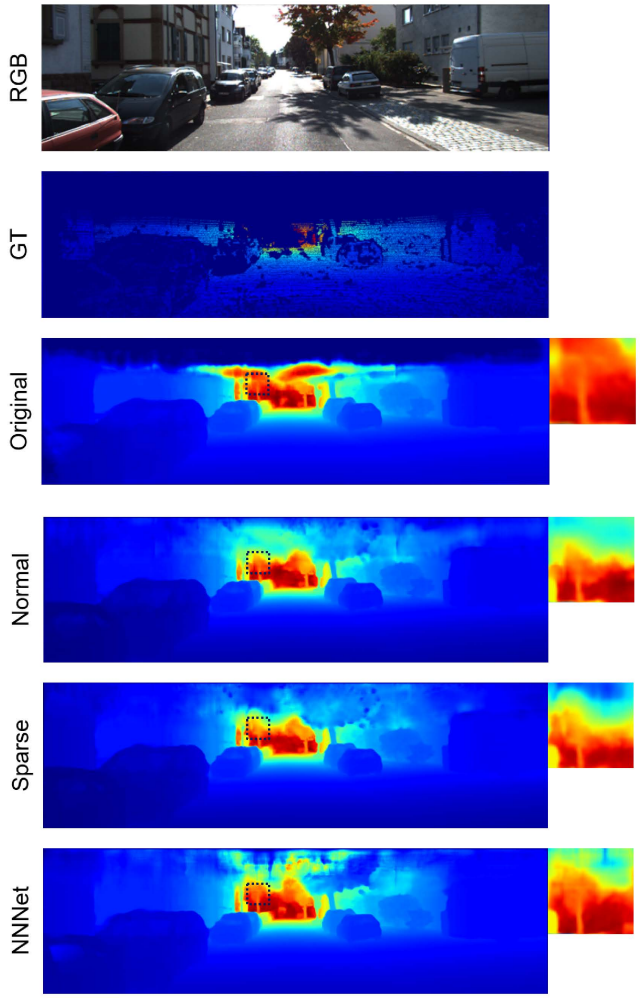


FIGURE 9. Visual comparison among different constraints on the KITTI validation set. GT: Ground truth. Original: Baseline method without the refinement module. Normal: Normal map constraint. Sparse: Sparse depth constraint. NNNet: New normal map constraint. The result of our method has sharper edges on the tree crown.

sparse depth image and the normal map from the depth image to NNNet as the constraint separately. Table 3 shows that sparse depth image is a more effective constraint than the normal map and the new normal map performs better than both sparse depth image and normal map. Note that NNNet needs adjustment to be applied to indoor scenes due to much difference between indoor and outdoor environments.

V. CONCLUSION

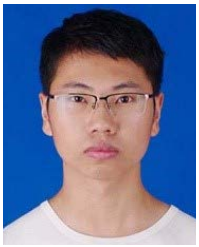
In this paper, we have proposed NNNet for sparse depth completion. We have generated the new normal map and use it as a constraint for depth completion. The new normal map contains not only accurate depth values at each depth point, but also the relative relationship between depth points, thus providing a stronger constraint for network training. NNNet consists of two branch-CNNs and one optimization module. The first branch is to predict new normal map from the input RGB and sparse depth image for constraint, while the second branch is to predict dense depth image from the sparse depth image and new normal map. We have used skip connections to fuse the features from two branches. The optimization module generates the final dense depth image by optimizing the predicted dense depth image based on a spatial propagation network. Various experiments demonstrate that the new normal map effectively fuses features of sparse depth and RGB images. Thus, NNNet achieves good performance in KITTI validation and test sets in terms of both visual comparison and quantitative measurements. Moreover, the average runtime of NNNet is only 0.034s, which has low computational complexity.

In the future work, we would like to extend NNNet to detection and identification of objects in road scenes such as pedestrians and cars. Also, we will explore semantic segmentation of LiDAR depth images and consider NNNet to improve the semantic segmentation accuracy.

REFERENCES

[1] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2007.
[2] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5695–5703.
[3] M. P. Muresan, I. Giosan, and S. Nedevschi, "Stabilization and validation of 3D object position using multimodal sensor fusion and semantic segmentation," *Sensors*, vol. 20, no. 4, p. 1110, Feb. 2020.
[4] H. Kim, J. Cho, D. Kim, and K. Huh, "Intervention minimized semi-autonomous control using decoupled model predictive control," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017, pp. 618–623.
[5] A. Arikan, A. Kayaduman, S. Polat, Y. Simsek, I. C. Dikmen, H. G. Bakir, T. Karadag, and T. Abbasov, "Control method simulation and application for autonomous vehicles," in *Proc. Int. Conf. Artif. Intell. Data Process. (IDAP)*, Sep. 2018, pp. 1–4.
[6] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant CNNs," in *Proc. Int. Conf. 3D Vis. (DV)*, Oct. 2017, pp. 11–20.
[7] F. Ma, G. V. Cavalheiro, and S. Karaman, "Self-supervised sparse-to-dense: Self-supervised depth completion from LiDAR and monocular camera," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 3288–3295.
[8] A. Eldesokey, M. Felsberg, and F. S. Khan, "Confidence propagation through CNNs for guided sparse depth regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2423–2436, Oct. 2019.
[9] Y. Xu, X. Zhu, J. Shi, G. Zhang, H. Bao, and H. Li, "Depth completion from sparse LiDAR data with depth-normal constraints," in *Proc. IEEE Conf. Comput. Vis.*, Oct. 2019, pp. 2811–2820.
[10] J. Qiu, Z. Cui, Y. Zhang, X. Zhang, S. Liu, B. Zeng, and M. Pollefeys, "DeepLiDAR: Deep surface normal guided depth prediction for outdoor scene from sparse LiDAR data and single color image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3313–3322.
[11] X. Cheng, P. Wang, and R. Yang, "Depth estimation via affinity learned with convolutional spatial propagation network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 103–119.
[12] Y. Zhang and T. Funkhouser, "Deep depth completion of a single RGB-D image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 175–185.
[13] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1–8.
[14] K. Karsch, C. Liu, and S. B. Kang, "Depth extraction from video using non-parametric sampling," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 775–788.
[15] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," 2014, *arXiv:1406.2283*.
[16] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 4th Int. Conf. 3D Vis. (DV)*, Oct. 2016, pp. 239–248.
[17] L. Ladicky, J. Shi, and M. Pollefeys, "Pulling things out of perspective," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 89–96.

[18] W. Chen, D. Xiang, and J. Deng, "Surface normals in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1557–1566.
[19] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2015.
[20] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, "GeoNet: Geometric neural network for joint depth and surface normal estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 283–291.
[21] N. Chodosh, C. Wang, and S. Lucey, "Deep convolutional compressed sensing for lidar depth completion," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, Dec. 2018, pp. 499–513.
[22] S. Hawe, M. Kleinsteuber, and K. Diepold, "Dense disparity maps from sparse disparity measurements," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2126–2133.
[23] L.-K. Liu, S. H. Chan, and T. Q. Nguyen, "Depth reconstruction from sparse samples: Representation, algorithm, and sampling," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1983–1996, Jun. 2015.
[24] J. T. Barron and B. Poole, "The fast bilateral solver," in *Computer Vision—ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, Sep. 2016, pp. 617–632.
[25] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 993–1000.
[26] J. Ku, A. Harakeh, and S. L. Waslander, "In defense of classical image processing: Fast depth completion on the CPU," in *Proc. 15th Conf. Comput. Robot Vis. (CRV)*, May 2018, pp. 16–22.
[27] A. Eldesokey, M. Felsberg, and F. S. Khan, "Propagating confidences through CNNs for sparse data regression," 2018, *arXiv:1805.11913*.
[28] M. Jaritz, R. D. Charette, E. Wirbel, X. Perrotton, and F. Nashashibi, "Sparse and dense data with CNNs: Depth completion and semantic segmentation," in *Proc. Int. Conf. 3D Vis. (DV)*, Sep. 2018, pp. 52–60.
[29] O. M. Aodha, N. D. Campbell, A. Nair, and G. J. Brostow, "Patch based synthesis for single depth image super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, Oct. 2012, pp. 71–84.
[30] M. Hornacek, C. Rhemann, M. Gelautz, and C. Rother, "Depth super resolution by rigid body self-similarity in 3D," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1123–1130.
[31] G. Riegler, M. Rüther, and H. Bischof, "ATGV-NET: Accurate depth super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 268–284.
[32] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[33] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. Kweon, "High quality depth map upsampling for 3D-TOF cameras," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1623–1630.
[34] J. Xie, R. S. Feris, and M.-T. Sun, "Edge-guided single depth image super resolution," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 428–438, Jan. 2015.
[35] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2650–2658.
[36] F. Ma and S. Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 4796–4803.
[37] Z. Huang, J. Fan, S. Cheng, S. Yi, X. Wang, and H. Li, "HMS-Net: Hierarchical multi-scale sparsity-invariant network for sparse depth completion," *IEEE Trans. Image Process.*, vol. 29, pp. 3429–3441, 2020.
[38] W. Van Gansbeke, D. Neven, B. De Brabandere, and L. Van Gool, "Sparse and noisy LiDAR completion with RGB guidance and uncertainty," in *Proc. 16th Int. Conf. Mach. Vis. Appl. (MVA)*, May 2019, pp. 1–6.
[39] M. Hu, S. Wang, B. Li, S. Ning, L. Fan, and X. Gong, "PENet: Towards precise and efficient image guided depth completion," 2021, *arXiv:2103.00783*.
[40] S. Zhao, M. Gong, H. Fu, and D. Tao, "Adaptive context-aware multi-modal network for depth completion," *IEEE Trans. Image Process.*, vol. 30, pp. 5264–5276, 2021.
[41] L. Liu, X. Song, X. Lyu, J. Diao, M. Wang, Y. Liu, and L. Zhang, "FCFR-Net: Feature fusion based coarse-to-fine residual learning for depth completion," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 3, pp. 2136–2144.
[42] J. Tang, F.-P. Tian, W. Feng, J. Li, and P. Tan, "Learning guided convolutional network for depth completion," *IEEE Trans. Image Process.*, vol. 30, pp. 1116–1129, 2021.
[43] X. Cheng, P. Wang, C. Guan, and R. Yang, "CSPN++: Learning context and resource aware convolutional spatial propagation networks for depth completion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 10615–10622.
[44] S. Liu, S. De Mello, J. Gu, G. Zhong, M.-H. Yang, and J. Kautz, "Learning affinity via spatial propagation networks," 2017, *arXiv:1710.01020*.
[45] J. Park, K. Joo, Z. Hu, C.-K. Liu, and I. S. Kweon, "Non-local spatial propagation network for depth completion," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, May 2020, pp. 120–136.
[46] Y. Chen, B. Yang, M. Liang, and R. Urtasun, "Learning joint 2D-3D representations for depth completion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10023–10032.
[47] J. Gu, Z. Xiang, Y. Ye, and L. Wang, "DenseLiDAR: A real-time pseudo dense depth guided depth completion network," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 1808–1815, Apr. 2021.
[48] L. Yan, K. Liu, and L. Gao, "DAN-Conv: Depth aware non-local convolution for LiDAR depth completion," *Electron. Lett.*, vol. 57, no. 20, pp. 754–757, Sep. 2021.
[49] S. S. Shivakumar, T. Nguyen, I. D. Miller, S. W. Chen, V. Kumar, and C. J. Taylor, "DFuseNet: Deep fusion of RGB and sparse depth information for image guided dense depth completion," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 13–20.



JIADE LIU received the B.S. degree in mechatronic engineering from Xidian University, China, in 2019, where he is currently pursuing the M.S. degree in electronic engineering. His research interests include computer vision and deep learning.

CHEOLKON JUNG (Member, IEEE) is a born again Christian. He received the B.S., M.S., and Ph.D. degrees in electronic engineering from Sungkyunkwan University, Republic of Korea, in 1995, 1997, and 2002, respectively. He was a Research Staff Member with the Samsung Advanced Institute of Technology, Samsung Electronics, Republic of Korea, from 2002 to 2007. He was also a Research Professor with the School of Information and Communication Engineering, Sungkyunkwan University, from 2007 to 2009. Since 2009, he has been with the School of Electronic Engineering, Xidian University, China, where he is currently a Full Professor and the Director of the Xidian Media Laboratory. His research interests include image and video processing, computer vision, pattern recognition, machine learning, computational photography, video coding, virtual reality, information fusion, multimedia content analysis and management, and 3DTV.