Depth Completion from Sparse LiDAR Data with Depth-Normal Constraints

Yan Xu[1,2,3] Xinge Zhu[2] Jianping Shi[1] Guofeng Zhang[3] Hujun Bao[3] Hongsheng Li[2]

[1]SenseTime Research [2]The Chinese University of Hong Kong
[3]State Key Lab of CAD&CG, Zhejiang University

Abstract

Depth completion aims to recover dense depth maps from sparse depth measurements. It is of increasing importance for autonomous driving and draws increasing attention from the vision community. Most of existing methods directly train a network to learn a mapping from sparse depth inputs to dense depth maps, which has difficulties in utilizing the 3D geometric constraints and handling the practical sensor noises. In this paper, to regularize the depth completion and improve the robustness against noise, we propose a unified CNN framework that 1) models the geometric constraints between depth and surface normal in a diffusion module and 2) predicts the confidence of sparse LiDAR measurements to mitigate the impact of noise. Specifically, our encoder-decoder backbone predicts surface normals, coarse depth and confidence of LiDAR inputs simultaneously, which are subsequently inputted into our diffusion refinement module to obtain the final completion results. Extensive experiments on KITTI depth completion dataset and NYU-Depth-V2 dataset demonstrate that our method achieves state-of-the-art performance. Further ablation study and analysis give more insights into the proposed method and demonstrate the generalization capability and stability of our model.

1. Introduction

The widely used depth sensors, such as LiDAR, RGB-D camera and TOF cameras, generally generate sparse depth measurements due to the limited sensing scope, interferences from environments and economic considerations. For example, the top-class LiDAR sensor, Velodyne HDL-64E, costs about $100,000, but can only provide sparse measurements with vertical resolution/angular resolution of $0.4°/0.08°$. On the other hand, dense depth maps are required in many high-level applications including semantic segmentation, 3D reconstruction, SLAM, *etc*. To mitigate the gap between sparse and dense depth maps, depth completion, *i.e*., generating dense depth maps from sparse depth measurements, has been widely adopted.

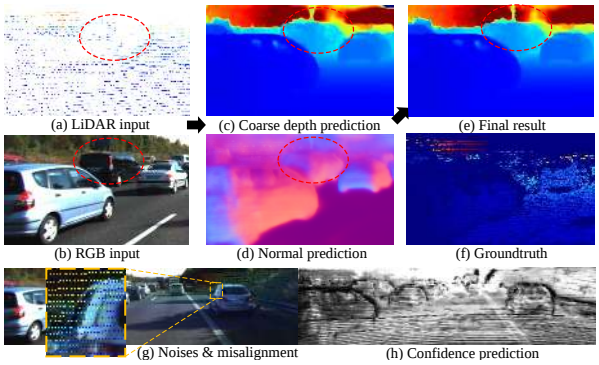


Figure 1: From sparse LiDAR measurements and color images (a-b), our model first infers the maps of coarse depth and normal (c-d), and then recurrently refines the initial depth estimation by enforcing the constraints between depth and normals. Moreover, to address the noises in practical LiDAR measurements (g), we employ a decoder branch to predict the confidences (h) of sparse inputs for better regularization. Best viewed on screen.

With the advances of deep learning methods, many depth completion approaches based on convolutional neural networks (CNNs) have been proposed. The mainstream of these methods is to directly input the sparse depth maps (with/without color images) into an encoder-decoder network and predict dense depth maps [26, 16, 36, 15, 10, 23, 2]. These black-box methods force the CNN to learn a mapping from sparse depth measurements to dense maps, which is generally a challenging task and leads to unsatisfactory completion results, as shown in Fig. 1 (c). We argue that proper geometric constraints should be incorporated into the end-to-end framework to regularize the completion process and make it more interpretable. Depth and surface normal are two strongly correlated factors in the 3D

This work was done when Yan Xu was an intern at SenseTime Research.

world and the locally linear orthogonality between them can be utilized in depth completion. Zhang *et al*. [46] takes the normal map (predicted by a CNN framework) as guidance and obtains the dense depth map by separately optimizing a linear system. Although their method performs better in post-processing the indoor RGB-D data compared with the methods neglecting 3D geometric constraints, it still suffers from huge running time-cost and limited generalization to driving scenarios. Moreover, their normal prediction training and the optimization of dense depth are isolated, which prohibits joint optimization in a data-driven manner.

In this paper, to regularize the depth completion results with 3D geometric constraints, we propose to model the locally linear orthogonality between depth and normal by associating them in the *plane-origin distance* space (the distance from the corresponding tangent plane to the origin, *i.e*., camera center in our case). We first adopt a CNN-based backbone to estimate the surface normal and depth (from sparse LiDAR measurements and color images). Then, we transform the predicted depth and normal to the plane-origin distance space, and conduct a refinement process in this space via a diffusion model to enforce the geometric constraints. Compared with previous works [21, 2] that model the depth variation in 2D space and assume piecewise constant depth, we model the geometric constraints in 3D space based on the assumption that 3D structures are constituted by piecewise planes and the plane-origin distances are therefore piecewise constant. The transformation to plane-origin distance enforces constraints between depth and normal during training, and improves the completion accuracy and stability in inference. Furthermore, to mitigate the effect of sensor noise which is inevitable on boundaries or moving objects as illustrated in Fig. 1 (g), a confidence branch is introduced in our framework to predict the uncertainties of sparse depth measurements from sensors.

Our contributions mainly lie in three aspects:

1. We reposition the focus of depth completion from 2D space to 3D space based on the assumption that a 3D scene is constituted by piecewise planes Specifically, we conjugate the depth and surface normal in the plane-origin distance space and refine it via a recurrent diffusion module, which enforces the constraints between depth and surface normal in depth completion process.

2. Based on this insight, we propose a unified two-stage CNN framework to achieve depth completion from very sparse inputs, *e.g*., LiDAR measurements. To improve the robustness to the noises in practical sensors, we further introduce a confidence prediction branch to impede the propagation of information associated with noises.

3. Our framework can be trained in an end-to-end manner, and extensive experimental results show that our model achieves state-of-the-art performance while keeping good generalization capability.

2. Related Work

Depth Completion. Depth completion has been intensively studied since the emergence of active depth sensors. Existing approaches mainly aim to handle the incomplete depth measurements from two types of sensors, *i.e*. structured-light scanners and LiDAR. The methods for structured-light scanners are widely used in 3D-reconstruction post-processing, while the methods for LiDAR usually require real-time responses in the scenarios of robotic navigation and autonomous driving.

The classic methods generally employ hand-crafted features or kernels to complete the missing values [13, 1, 8, 12, 27, 40, 19, 25, 17]. Most of these methods are task-specific and usually confronted with performance bottleneck due to the limited generalization ability. Recently, the learning-based methods have shown promising performance on depth completion. Some of these methods achieve depth completion based solely on the sparse depth measurements. Uhrig *et al*. [36] proposed a sparsity-invariant convolution layer to enhance the depth measurements from LiDAR. Besides, in the work of [11], they model the confidence propagation through layers and reduce the quantity of model parameters. However, the assistance from other modalities, *e.g*., color images, can significantly improve the completion accuracy. Ma *et al*. concatenated the sparse depth and color image as the inputs of an off-the-shelf network [26] and further explored the feasibility of self-supervised LiDAR completion [23]. Moreover, [14, 16, 33, 4] proposed different network architectures to better exploit the potential of the encoder-decoder framework. However, the encoder-decoder architecture tends to predict the depth maps comprehensively but fails to concentrate on the local areas. To mitigate this problem, Cheng *et al*. [2] proposed a convolutional spatial propagation refinement network (inspired by the work of [22]) to post process the depth completion results with neighboring depth values. They simply conduct the refinement in 2D depth space based on the assumption that the depth values are locally constant. However, different from the segmentation task[22], this assumption is sub-optimal for depth completion and their performance in outdoor scenes is still barely satisfactory. Furthermore, current approaches ignore the noises in LiDAR measurements, which are inevitable in practical scenarios.

Depth and Normal. In previous works, the relation between depth and surface normal has been exploited in various ways to improve the depth accuracy [45, 41, 28]. For the monocular depth estimation tasks, [41, 28] compute normal from depth and then recover the depth from normal inversely to enforce the constraints between them. Depth completion can also benefit from such geometric constraints. Zhang *et al*. [46] established a linear system based on the geometric constraints and solve it by Cholesky factorization. However, the optimization of a linear sys-

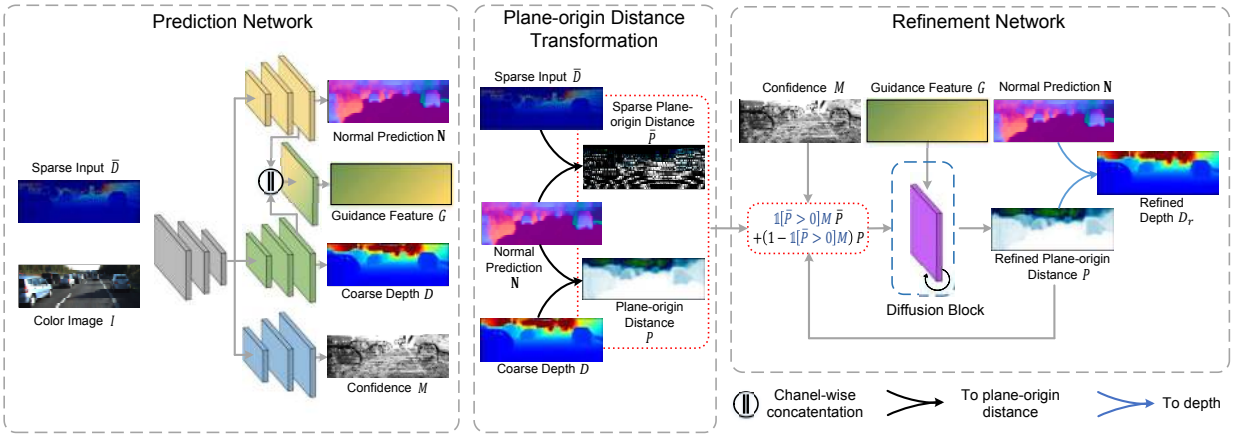


Figure 2: Overview of our proposed framework. The prediction network first predicts maps of surface normal $\mathbf{N}$, coarse depth D and confidence M of sparse depth input with a shared-weight encoder and independent decoders. Then, the sparse depth inputs $\bar{D}$ and coarse depth D are transformed to the plane-origin distance space as $\bar{P}$ and P, using Eq. (5). Next, the refinement network, an anisotropic diffusion module, refines the coarse depth map D in the plane-origin distance subspace to enforce the constraints between depth and normal and to incorporate information from the confident sparse depth inputs. During the refinement, the diffusion conductance depends on the similarity in guidance feature map G (See Eq. (7)). Finally, the refined P is inversely transformed back to obtain the refined depth map D_r when the diffusion is finished.

tem is hard to be employed in an end-to-end framework and achieve joint optimization. Moreover, although their method is suitable for post-processing the RGB-D camera data, but can hardly achieve real-time processing.

Anisotropic Diffusion Anisotropic diffusion originally models the physical process that equilibrates concentration differences without creating or destroying mass, *e.g.* heat diffusion. Anisotropic diffusion has been extensively used in image denoising [43, 42, 5], depth completion [21, 32, 2], segmentation [18, 22, 44, 37, 38, 31], *etc*. The previous classic methods define the diffusion conductance only based on the similarity in diffusion space or in the guidance map (*e.g.*, a color image), which limits the performance. In our work, we take advantages of feature extraction capability of CNN and use the high-dimension features to calculate the conductance.

3. Method

In this paper, we assume that a 3D scene is constituted by piecewise planes, and the distances between these planes and the origin (*plane-origin distance*) are therefore piecewise constant. Based on this assumption, we proposed a two-stage end-to-end deep learning framework, which regularizes the depth completion process using the constraints between depth and surface normal. As illustrated in Fig. 2, our framework mainly consists of two parts, *i.e.*, the prediction network and refinement network. The prediction network estimates the surface normal map, the coarse depth map and confidences of sparse depth inputs with a shared-weight encoder and independent decoders. Then, the sparse input and coarse depth maps are transfromed to the plane-origin distance subspace with normal estimation. Next, the refinement network, a diffusion model, recurrently refines plane-origin distance, which enforces the piecewise plane constraints and regularizes the depth completion. Compared with many previous works [21, 2] that assume *piecewise constant* depth, our method utilizes the geometric constraints between depth and surface normal, and performs better and more stably in the missing regions. Finally, the refined depth can be obtained via the inverse transformation without losing accuracy when the refinement is finished.

3.1. Prediction Network

The prediction network takes sparse depth $\bar{D}$ and the corresponding color image I as inputs, and predicts surface normal map $\mathbf{N}$, coarse depth completion D and confidence map M (of sparse input) via separate decoders. We adopt the widely used U-Net [29] architecture for prediction network, *i.e.*, using a ResNet-34 variant as the encoder and cascaded upsampling layers as the decoders. The specific architecture is included in the supplementary materials.

We apply $L2$ reconstruction loss for the coarse depth completion D, *i.e.*, $L_D = \frac{1}{n}\sum_{\mathbf{x}} ||D(\mathbf{x}) - D^*(\mathbf{x})||_2^2$, where n is the number of pixels. For normal prediction, we generate the normal target from depth groundtruth, *i.e.*, selecting a set of nearest 3D points for each location, and computing the normal direction based on them via principal component analysis (PCA) [30]. Then, a negative cosine loss proposed by [9] is used for normal prediction, that is

$$L_N = -\frac{1}{n}\sum_{\mathbf{x}} \mathbf{N}(\mathbf{x}) \cdot \mathbf{N}^*(\mathbf{x}), \tag{1}$$

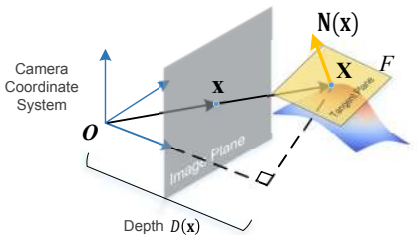


Figure 3: In camera coordinate system, the relation between depth and normal can be established via the tangent plane equation.

where $\mathbf{N}$ is the normal prediction while $\mathbf{N}^*$ denotes the computed normal target as stated above. The confidence map M is to mitigate the negative impact caused by noises in practical LiDAR measurements as illustrated in Fig. 1. Since there is no ground-truth for the confidence, we use a function to model it during training inspired by the probability density function of Laplace distribution, which is given by

$$M^* = \exp\left(-\frac{|\bar{D} - D^*|}{b}\right), \tag{2}$$

where $\bar{D}$ is the noisy sparse input, D^* denotes the depth ground-truth, and b is a parameter that controls the tolerance to the error when modeling the confidence. we apply an $L2$ loss denoted as L_C to draw the prediction close to M^*: $L_C = \frac{1}{n}\sum_{\mathbf{x}} ||M(\mathbf{x}) - M^*(\mathbf{x})||_2^2$, where n is the number of pixels. Meanwhile, the following refinement network can also affect the confidence prediction via backpropagation to achieve a better performance.

3.2. Recurrent Refinement Network

The afore mentioned prediction network estimates dense completion results from sparse depth inputs. The encoder-decoder architecture does not exploit the geometric constraints between depth and surface normal to regularize the estimated depth and has difficulties of taking full advantages of the sparse inputs. To address this problem, we propose to further refine the completion results in a novel *plane-origin distance* subspace via an anisotropic diffusion module [39] based on the assumption that the 3D surface of the scene is constituted by piece-wise planes and the plane-origin distance is piecewise constant.

3.2.1 Plane-origin Distance

As illustrated in Fig. 3, let $\mathbf{X}$ be a 3D point and $\mathbf{x}$ be its projected 2D point on the image plane. The surface normal $\mathbf{N}(\mathbf{x})$ at 3D-point $\mathbf{X}$ is defined as the vector starting from $\mathbf{X}$ and perpendicular to the tangent plane F. The point-normal equation of plane F can be written as

$$\mathbf{N}(\mathbf{x}) \cdot \mathbf{X} - P = 0 \tag{3}$$

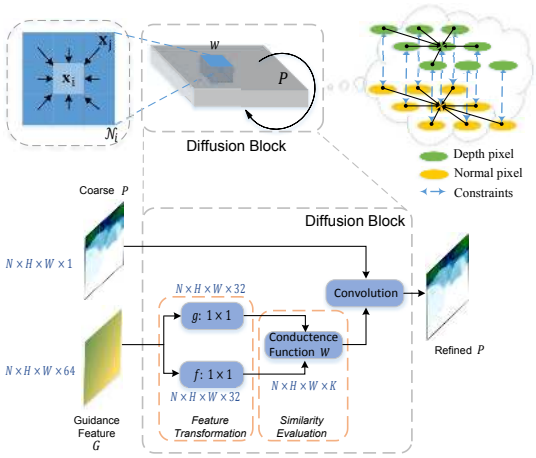


Figure 4: The proposed differentiable diffusion block. In each refinement iteration, high-dimensional feature vectors (*e.g.*, of dimension 64) in guidance feature map G are independently transformed via two different functions f and g (modeled as two convolution layers followed by normalization). Then, the conductances from each location $\mathbf{x}_i$ (in plane-origin distance map P) to its neighboring K pixels ($\mathbf{x}_j \in \mathcal{N}_i$) are calculated using Eq. (7). Finally, the diffusion is performed through a convolution operation with the kernels defined by the previous computed conductances. Through such diffusion, depth completion results are regularized by the constraint between depth and normal.

Hence, the value $P = \mathbf{N}(\mathbf{x}) \cdot \mathbf{X}$ should be constant for all 3D points on the same plane. As P is the distance between the plane and the origin (camera centre in our case), for simplicity, we refer to P as *plane-origin distance* in our paper.

By adopting the pinhole camera model, the 3D-point $\mathbf{X}$ can be reconstructed with its depth value $D(\mathbf{x})$ and 2D image location:

$$\mathbf{X} = D(\mathbf{x}) \cdot \mathbf{C}^{-1}\mathbf{x}, \tag{4}$$

where $\mathbf{C}$ denotes the camera intrinsic parameter matrix and 2D-point $\mathbf{x}$ is in homogeneous form. By further substituting Eq. (4) into Eq. (3), we have the relation between plane-origin distance P and depth $D(\mathbf{x})$:

$$P(\mathbf{x}) = D(\mathbf{x})\mathbf{N}(\mathbf{x})\mathbf{C}^{-1}\mathbf{x}. \tag{5}$$

Note that, with a slight abuse of notation, here we also use P to denote the map of plane-origin distances for all pixels. After the plane-origin distance map has been refined (to be discussed in the next subsection), the refined depth map $D(\mathbf{x})$ can be inversely obtained as $D(\mathbf{x}) = P(\mathbf{x})/(\mathbf{N}(\mathbf{x})\mathbf{C}^{-1}\mathbf{x})$.

3.2.2 Plane-origin Distance Diffusion for Depth Refinement

As stated before, for all the 3D points $\mathbf{X}_j$ on the same local plane with $\mathbf{X}_i$, we model that $P(\mathbf{x}_j) = P(\mathbf{x}_i)$, where $\mathbf{x}_j$ and $\mathbf{x}_i$ are the projected 2D locations for $\mathbf{X}_j$ and $\mathbf{X}_i$ respectively. To enforce this geometric constraint in depth completion, we conduct the anisotropic diffusion on the plane-origin distance map P:

$$\begin{aligned} P(\mathbf{x_i}) \leftarrow & (1 - \sum_{\mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)} w(\mathbf{x}_i, \mathbf{x}_j)) P(\mathbf{x_i}) \\ & + \sum_{\mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)} w(\mathbf{x}_i, \mathbf{x}_j) P(\mathbf{x_j}) \end{aligned} \tag{6}$$

During the diffusion process, pixel $\mathbf{x}_i$ receives information from surrounding pixels in neighborhood $\mathcal{N}(\mathbf{x}_i)$ while $w(\mathbf{x}_i, \mathbf{x}_j)$ measures how likely that $\mathbf{x}_i$ and $\mathbf{x}_j$ lie on the same plane.

Some classic methods, such as [21], define the diffusion conductance w based only on the similarity in the color image space. Thanks to the strong feature learning capability of CNN, we are able to measure the similarity in the high-dimension feature space. We take the geometrical feature map $\mathbf{G}$ generated by the prediction network (illustrated in Fig. 2) to model the diffusion conductance between $\mathbf{x}_i$ and $\mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)$. If two features at $\mathbf{x}_i$ and $\mathbf{x}_j$ are geometrically similar, they are likely to be on the same plane and should share similar $P(\mathbf{x}_i)$ and $P(\mathbf{x}_j)$ values. With this intuition, we model the conductance between $\mathbf{x}_i$ and $\mathbf{x}_j$ as

$$w(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{S(\mathbf{x}_i)} \exp\left(-\frac{(1 - f(\mathbf{G}(\mathbf{x}_i))^T g(\mathbf{G}(\mathbf{x}_j)))^2}{2\sigma^2} \right). \tag{7}$$

We adopt two different feature transformation functions f and g for $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively. Thus, the conductances from $\mathbf{x}_i$ to $\mathbf{x}_j$ and from $\mathbf{x}_j$ to $\mathbf{x}_i$ are asymmetric, *i.e.*, $w(\mathbf{x}_i, \mathbf{x}_j) \neq w(\mathbf{x}_j, \mathbf{x}_i)$. Such asymmetry provides more flexibility for the diffusion. For instance, the locations with confident sparse depth inputs may refuse the information from others, and the locations with unreliable values (*e.g.* sky) can stop their propagation to others. f and g are implemented as convolutional layers followed by a L2 normalization across channel dimmension as illustrated in Fig. 4. In addition, σ is a learnable parameter (initialized with 0.1 empirically) to control the diffusion strength globally and $S(\mathbf{x}_i) = \sum_{j \in N_i} \exp(-\frac{(1 - f(\mathbf{G}(\mathbf{x}_i))^T g(\mathbf{G}(\mathbf{x}_j)))^2}{2\sigma^2})$ is a normalization term.

3.2.3 Plane-origin Refinement and Depth Recovery

As demonstrated in Algorithm 1 and Fig. 2, our refinement framework first transforms the sparse depth inputs $\bar{D}$ and coarse depth map D (from previous prediction network) to plane-origin distances, obtaining $\bar{P}$ and P (Eq. (5)) respectively and then performs the diffusion refinement (Eq. (6)). During the diffusion, we take confident pixels in sparse plane-origin distance map $\bar{P}$ as seeds and refine the values in P with them at each iteration, which can be expressed as

$$\begin{aligned} P(\mathbf{x}) \leftarrow & \mathbb{1}[\bar{P}(\mathbf{x}) > 0] M(\mathbf{x}) \bar{P}(\mathbf{x}) \\ & + (1 - \mathbb{1}[\bar{P}(\mathbf{x}) > 0] M(\mathbf{x})) P(\mathbf{x}), \end{aligned} \tag{8}$$

where $\mathbb{1}[\bar{P}(\mathbf{x}) > 0]$ is an indicator for the availability of $\bar{P}$ (also sparse depth $\bar{D}$) at location $\mathbf{x}$ and M denotes the predicted confidences of sparse depth inputs. The confidence map M largely prevents the propagation of noises in sparse measurements while allowing the the confident sparse depth inputs and the predicted depth map from U-Net to complement each other. Moreover, this strategy couples the depth and normal during training, which enforces the normal-depth constraints and results in better accuracy.

Algorithm 1 The refinement procedure

for all $\mathbf{x}$ **do**
  $\bar{P}(\mathbf{x}) \leftarrow \bar{D}(\mathbf{x}) \mathbf{N}(\mathbf{x}) \mathbf{C}^{-1} \mathbf{x}$
  $P(\mathbf{x}) \leftarrow D(\mathbf{x}) \mathbf{N}(\mathbf{x}) \mathbf{C}^{-1} \mathbf{x}$
end for
$i \leftarrow 0$
while $i < max\_iteration$ **do**
  for all $\mathbf{x}$ **do**
    $P(\mathbf{x}) \leftarrow \mathbb{1}[\bar{P}(\mathbf{x}) > 0] M(\mathbf{x}) \bar{P}(\mathbf{x}) + (1 - \mathbb{1}[\bar{P}(\mathbf{x}) > 0] M(\mathbf{x})) P(\mathbf{x})$
  end for
  for all $\mathbf{x}$ **do**
    Conduct the refinment using Eq. (6)
  end for
  $i \leftarrow i + 1$
end while
for all $\mathbf{x}$ **do**
  $D(\mathbf{x}) \leftarrow P(\mathbf{x}) / (\mathbf{N}(\mathbf{x}) \mathbf{C}^{-1} \mathbf{x})$
end for

3.3. Loss Functions

Our proposed network is trained end-to-end. Besides the afore mentioned loss functions L_D, L_N, L_C in prediction network in Sec. 3.1. For the refinement network, we also apply a L2 loss to supervise the learning of refinement results D_r, *i.e.*, $L_{D_r} = \frac{1}{n} \sum_{\mathbf{x}} ||D_r(\mathbf{x}) - {D_r}^*(\mathbf{x})||_2^2$. Our overall loss function can be written as

$$L = L_D + \alpha L_{D_r} + \beta L_N + \gamma L_C, \tag{9}$$

where α, β and γ adjust the weights among different terms in the loss function. In our experiments, we empirically set $\alpha = 1, \beta = 1, \gamma = 0.1$.

4. Experiments

We perform extensive experiments to evaluate the effectiveness of our model. In this section, we will first briefly introduce the dataset and evaluation metrics adopted in our experiments and then discuss our experiments.

Table 1: The evaluation results on the test set of KITTI depth completion benchmark. The root mean square error (RMSE) and mean absolute error (MAE) are in millimeters, while inverse RMSE and inverse MAE are in 1/kilometer.

Method	RMSE	MAE	iRMSE	iMAE
Ours	777.05	235.17	2.42	1.13
Sparse-to-Dense [23]	814.73	249.95	2.80	1.21
NConv-CNN [10]	829.98	233.26	2.60	1.03
Spade-RGBsD [16]	917.64	234.81	2.17	0.95
HMS-Net [14]	937.48	258.48	2.93	1.14
CSPN [3]	1019.64	279.46	2.93	1.15
Morph-Net [7]	1045.45	310.49	3.84	1.57
DFuseNet [33]	1206.66	429.93	3.62	1.79

4.1. Dataset and Metrics

RGB-D data is available in many existing datasets, *e.g*. [6, 24, 36, 34]. We conduct extensive experiments on KITTI depth completion benchmark [36] to evaluate the performance with practical sparse LiDAR data. Moreover, to demonstrate the generalization ability, we also perform experiments on indoor dataset, *i.e*., NYU-Depth-v2 [34].

KITTI depth prediction dataset. KITTI depth completion dataset [36] contains over 93k annotated depth maps with aligned sparse LiDAR measurements and RGB images. We train our model on the training split, and evaluate it on the official validation set and test set.

NYU-Depth-v2 dataset. NYU-Depth-v2 dataset consists of paired RGB images and depth maps collected from 464 different indoor scenes with a Microsoft Kinect. We adopt the official data split strategy and sample about 43k synchronized RGB-depth pairs from the training data with the same experimental setup as [26]. Moreover, preprocessing is performed with the official toolbox. The origin images of size 640×480, are down-sampled to half and then center-cropped to the size of 304×224.

Evaluation metrics. For the evaluation on KITTI dataset, we adopt the same metrics used in the KITTI benchmark: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), root mean squared error of the inverse depth (iRMSE) and mean absolute error of the inverse depth(iMAE). For the experiments on NYU-Depth-v2 dataset, we adopt 1) RMSE, 2) mean relative error (rel): $\frac{1}{|D|}\sum_{\mathbf{x}}|D(\mathbf{x}) - D^*(\mathbf{x})|/D^*(\mathbf{x})$ and 3) δ_t: percentage of depth estimations that satisfy $max(\frac{D^*(\mathbf{x})}{D(\mathbf{x})}, \frac{D(\mathbf{x})}{D^*(\mathbf{x})}) < t$, where $t \in \{1.25, 1.25^2, 2.25^3\}$.

4.2. Experimental Setup

Our framework is implemented on PyTorch library and trained on an NVIDIA Tesla V100 GPU with 16GB of memory. The networks are trained for 30/20 epochs for KITTI/NYU with a batch size of 16 and an initial learning rate of 4×10^{-4}. Our models are trained with ADAM optimizer which decays the learning rate with the poly strategy.

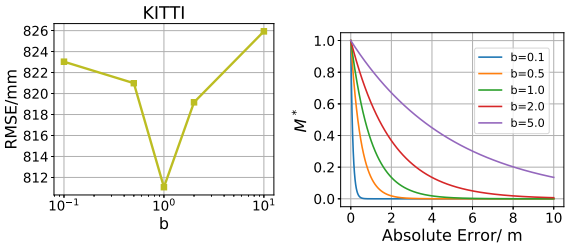


Figure 5: The effect of changing the tolerance parameter b in Eq. (2). The **left** figure exhibits the RMSE with different values of b and the **right** figure plots the curves of modeled confidence groundtruth M^* w.r.t the absolute difference between the sparse input and depth groundtruth.

Table 2: The performance comparison of different ablation variants on the validation set of KITTI benchmark.

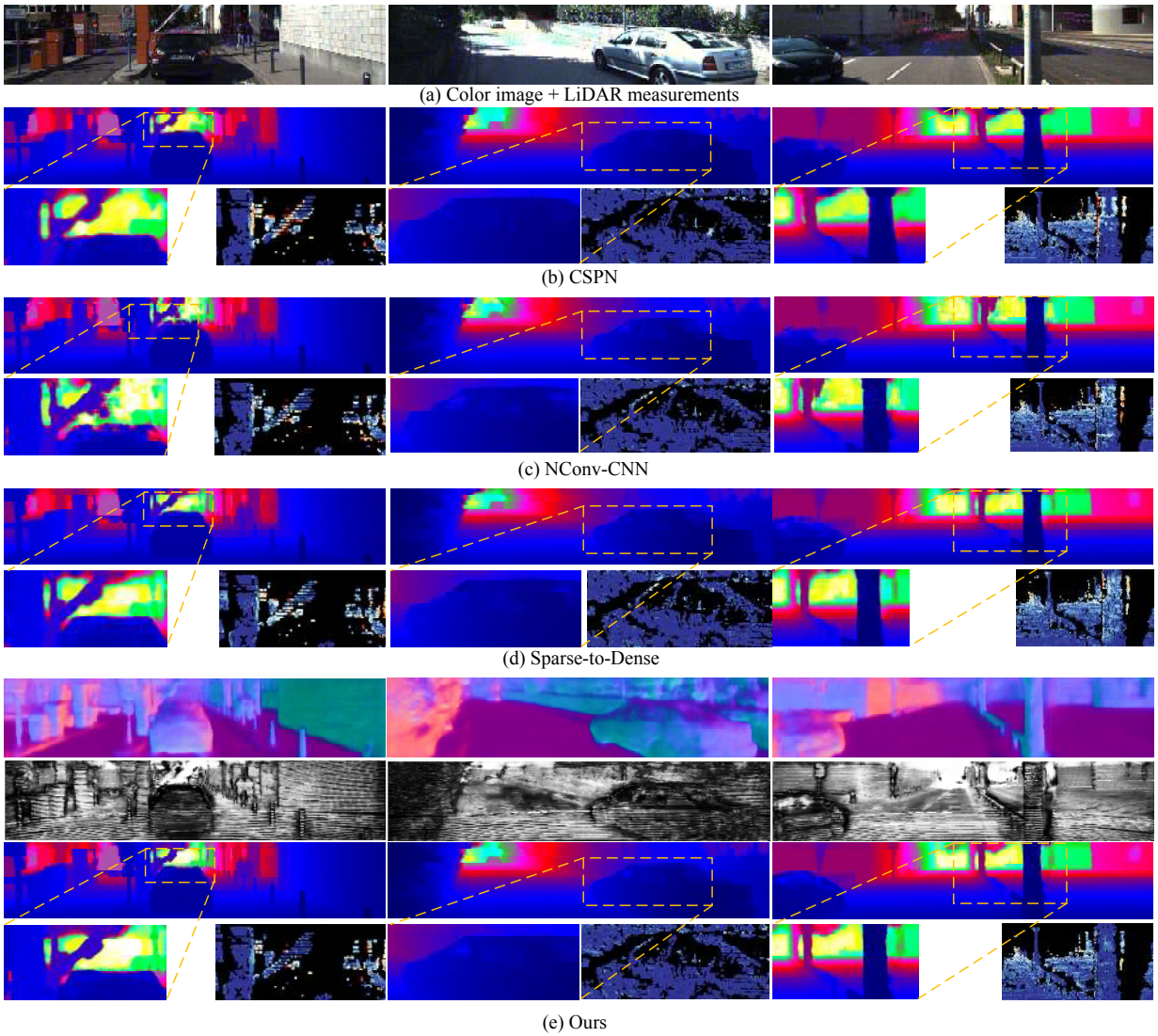
Method	RMSE	MAE	iRMSE	iMAE
w/o normal	846.51	256.71	3.07	1.35
w/o refinement	836.20	255.04	2.62	1.24
w/o replacement	825.85	258.5	2.56	1.26
w/o confidence	836.66	248.18	2.59	1.25
w/ same f,g	832.93	273.71	2.63	1.33
w/ Euclidean distance	843.34	238.55	2.89	1.57
w/ dot product	818.41	249.95	2.76	1.37
Full	**811.07**	**236.67**	**2.45**	**1.11**

4.3. Comparison with the State-of-the-Arts

We evaluate our model on the test set of KITTI depth completion benchmark and compare our method against other methods. Table 1 lists the comparison results with other high-ranking methods. Our method ranks 1st among these peer-reviewed methods according to the RMSE metric. We further conduct quantitative comparison with some competing approaches as demonstrated in Fig. 6. Our completion results benefit from the geometric constraints that the intermediate normal prediction and the depth estimation should be in consistency, which largely reduces the errors and recovers more details compared with these competing methods. For example, the outliers in the region of the telegraph pole (in last column of Fig. 6) are mostly eliminated via the geometry-aware diffusion refinement.

4.4. Ablation Study

To verify the effectiveness of each proposed component, we conduct extensive ablation studies by removing each component from our proposed framework. Apart from that, we also investigate the impact of different configurations of our proposed diffusion conductance function (Eq. (7)), *i.e*. with same feature transformation function (let $f = g$) or changing the embedded cosine similarity to Euclidean distance/dot product. The quantitative results are shown in Table 2, and the performances of all ablation variants degrade compared with our full model.



(a) Color image + LiDAR measurements

(b) CSPN

(c) NConv-CNN

(d) Sparse-to-Dense

(e) Ours

Figure 6: Quantitative comparison with other methods. For each method, we provide the whole completion results as well as the zoom-in views of details and error maps for better comparison. We also provide the normal prediction and confidence prediction of our method for better illustration.

Effectiveness of Geometric Constraints. To verify the effectiveness of the geometric constraints enforced by our plane-origin distance diffusion. We first evaluate our prediction network with only depth branch (w/o normal) and further remove our refinement network along with the confidence branch from the full model (w/o refinement) to see whether the encoder-decoder alone has the capability to exploit the geometric constraints (between depth and normal). Moreover, we also try to conduct the diffusion refinement without substituting the seeds $\bar{P}$ (w/o replacement) to see where the performance gain comes from. As exhibited in Table 2, the performance of two variants all degrades, but '*w/o replacement*' outperforms '*w/o refinement*', which demonstrates the effectiveness of our method in exploiting the geometric constraints.

Investigation of Diffusion Refinement Module. We investigate the configurations of our proposed diffusion module. First, we try to use same transformation function in Eq. (7) to calculate the similarity, *i.e.*, adopting a symmetric conductance function by letting $f = g$. As shown in Table 2, the performance with symmetric conductance (w/ same f, g) is inferior to the proposed asymmetric one (Full). Then, we also experiment on different similarity functions : $w(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{S(\mathbf{x}_i)} \exp(-\frac{||f(\mathbf{G}(\mathbf{x}_i)) - g(\mathbf{G}(\mathbf{x}_j))||_2^2}{2\sigma^2})$ (w/ Euclidean distance) and $w(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{S(\mathbf{x}_i)} \exp(f(\mathbf{G}(\mathbf{x}_i))^T g(\mathbf{G}(\mathbf{x}_j)))$ (w/ dot product). It can be found that the proposed conductance function performs better than these variants.

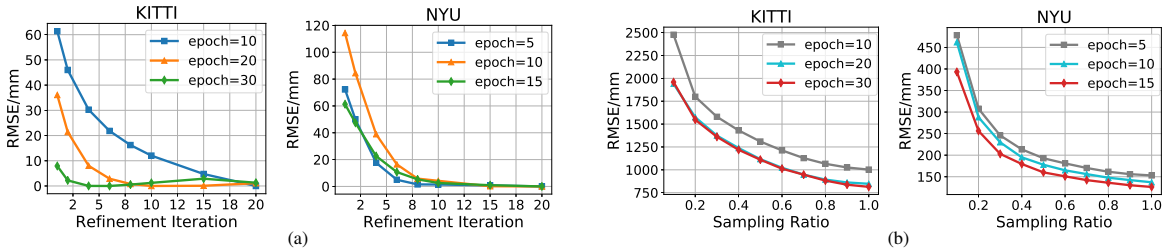


Figure 7: Stability analysis. (a) RMSE of our methods w.r.t. the number of refinement iterations on both KITTI and NYU validation sets. Here, we shift each curve by subtracting the minimum values for better demonstration. (b) RMSE of our model w.r.t. different sampling ratios of the sparse depth inputs.

Table 3: Evaluation on NYU-Depth-v2 dataset. The Root mean square errors (RMSE) are in millimeters and all the methods are evaluated with same sparsity of depth inputs (*i.e.*, 500 samples).

Method	RMSE	rel	$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
Diffusion [21]	1.231	0.202	89.1	91.2	94.3
Cross bilateral filter [35]	0.748	0.106	90.1	93.1	93.9
Colorization [20]	0.185	0.039	97.2	97.9	98.1
CSPN [3]	0.117	**0.016**	99.2	99.9	100.0
Ma *et al.* [26]	0.230	0.044	97.1	99.4	99.8
Ours (ResNet-34)	0.119	0.021	99.4	99.9	100.0
Ours (ResNet-50)	**0.112**	0.018	**99.5**	**99.9**	**100.0**

Effectiveness of the Confidence Prediction. We can see that the regions with lower confidence prediction (Fig. 6 (e)) are mainly concentrated in the areas of moving objects or objects boundaries, which is mostly consistent with the noise occurrence in Fig. 6 (a)). We further remove the confidence prediction scheme from our framework to verify the necessity of confidence map M in diffusion model. The performance ('w/o confidence') in Table 2 degrades as expected which is caused by the spreading of errors. Furthermore, we investigate the effects of different values of parameter b in the confidence model (Eq. (2)). As shown in Fig. 5, a too large or too small b will degrade the performance. This is because a too large b makes the model too tolerant to noises while a too small b makes the model too conservative to assign high confidence to valid measurements (the **right** plot in Fig. 5 shows a set of confidence curves with different b values).

4.5. Analysis of Generalization Ability and Stability

Generalization Ability to Indoor Scenes. Although we mainly focus on the outdoor application scenarios, we also train our model on indoor scenes, *i.e.*, NYU-Depth-v2 dataset. As NYU-Depth-v2 dataset provides relatively denser depth measurements by Microsoft Kinect, we uniformly sample the depth map to obtain the sparser version following previous works [26, 14]. We compare our results with latest CNN-based methods [26, 2] as well as the classic methods [21, 35, 20] as shown in Table 3, and our method achieves state-of-the-art performance as well. Moreover, our model with even a ResNet-34 encoder (denoted as 'Ours (ResNet-34)') achieves similar or even better performance compared with the previous methods with a ResNet-50 [26, 2], and the adoption of a ResNet-50 encoder (denoted as 'Ours(ResNet-50)') in our framework can further improve the performance.

Stability Analysis. To evaluate the refinement stability of our proposed recurrent refinement network, we select the model snapshots from different epochs that are all trained with a kernel size of 5 and refinement iteration of 8. But, for inference, we perform the refinement with different number of iterations. As shown in Fig. 7 (a), the error decreases and becomes steady as more refinement iterations are performed (even exceeding that in the training phase). Moreover, we also verify our model's robustness to different input sparsity levels by sub-sampling the raw LiDAR inputs in KITTI or the sampled depth maps in NYU. As shown in Fig. 7 (b), the performances drop when the sampling ratio decreases as expected, but the model can still provide reasonable results even with $1/10$ of the original sparse inputs.

5. Conclusion

In this paper, we propose a unified framework constituted by two modules, *i.e.*, prediction network and refinement network, to address the problem of depth completion from sparse inputs. We follow the 3D nature of depth to shift the focus from 2D space to 3D space and utilize the depth-normal constraints to regularize the depth completion via a diffusion model in plane-origin distance space. The proposed diffusion model adaptively adjusts the conductance between pairs of vertices according their similarities in the high-dimensional feature space. Moreover, we also handle the noises in LiDAR measurements by introducing a decoder branch to predict the confidences of sparse inputs, and impede the propagation of errors in refinement module. Extensive experiments demonstrate that our method achieves state-of-the-art performance on both outdoor and indoor datasets.

Acknowledgement This work is supported in part by SenseTime Group Limited, in part by General Research Fund through the Research Grants Council of Hong Kong under Grants CUHK14202217, CUHK14203118, CUHK14205615, CUHK14207814, CUHK14213616, CUHK14208417, CUHK14239816, in part by CUHK Direct Grant and in part by the Fundamental Research Funds for the Central Universities (No. 2018FZA5011).

References

[1] Marcelo Bertalmio, Andrea L Bertozzi, and Guillermo Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Conference on Computer Vision and Pattern Recognition.*, volume 1. IEEE, 2001.

[2] Xinjing Cheng, Peng Wang, and Ruigang Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–119, 2018.

[3] Xinjing Cheng, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. *arXiv preprint arXiv:1810.02695*, 2018.

[4] Nathaniel Chodosh, Chaoyang Wang, and Simon Lucey. Deep convolutional compressed sensing for lidar depth completion. In *Asian Conference on Computer Vision*, pages 499–513. Springer, 2018.

[5] Ulrich Clarenz, Udo Diewald, and Martin Rumpf. *Anisotropic geometric diffusion in surface processing*. IEEE, 2000.

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] Martin Dimitrievski, Peter Veelaert, and Wilfried Philips. Learning morphological operators for depth completion. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 450–461. Springer, 2018.

[8] David Doria and Richard J Radke. Filling large holes in lidar data by inpainting depth gradients. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 65–72. IEEE, 2012.

[9] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015.

[10] Abdelrahman Eldesokey, Michael Felsberg, and Fahad Shahbaz Khan. Confidence propagation through cnns for guided sparse depth regression. *arXiv preprint arXiv:1811.01791*, 2018.

[11] Abdelrahman Eldesokey, Michael Felsberg, and Fahad Shahbaz Khan. Propagating confidences through cnns for sparse data regression. *arXiv preprint arXiv:1805.11913*, 2018.

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

[13] Daniel Herrera, Juho Kannala, Janne Heikkilä, et al. Depth map inpainting under a second-order smoothness prior. In *Scandinavian Conference on Image Analysis*, pages 555–566. Springer, 2013.

[14] Zixuan Huang, Junming Fan, Shuai Yi, Xiaogang Wang, and Hongsheng Li. HMS-Net: Hierarchical multi-scale sparsity-invariant network for sparse depth completion. *arXiv preprint arXiv:1808.08685*, 2018.

[15] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. CCNet: Criss-cross attention for semantic segmentation. *arXiv preprint arXiv:1811.11721*, 2018.

[16] Maximilian Jaritz, Raoul de Charette, Emilie Wirbel, Xavier Perrotton, and Fawzi Nashashibi. Sparse and dense data with cnns: Depth completion and semantic segmentation. *arXiv preprint arXiv:1808.00769*, 2018.

[17] Martin Kiechle, Simon Hawe, and Martin Kleinsteuber. A joint intensity and depth co-sparse analysis model for depth map super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1545–1552, 2013.

[18] Gunhee Kim, Eric P Xing, Li Fei-Fei, and Takeo Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *International Conference on Computer Vision*, pages 169–176. IEEE, 2011.

[19] Jason Ku, Ali Harakeh, and Steven L Waslander. In defense of classical image processing: Fast depth completion on the cpu. In *15th Conference on Computer and Robot Vision (CRV)*, pages 16–22. IEEE, 2018.

[20] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. In *ACM transactions on graphics (tog)*, volume 23, pages 689–694. ACM, 2004.

[21] Junyi Liu and Xiaojin Gong. Guided depth enhancement via anisotropic diffusion. In *Pacific-Rim Conference on Multimedia*, pages 408–417. Springer, 2013.

[22] Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, and Jan Kautz. Learning affinity via spatial propagation networks. In *Advances in Neural Information Processing Systems*, pages 1520–1530, 2017.

[23] Fangchang Ma, Guilherme Venturelli Cavalheiro, and Sertac Karaman. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. *arXiv preprint arXiv:1807.00275*, 2018.

[24] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6120–6127, 2019.

[25] Oisin Mac Aodha, Neill DF Campbell, Arun Nair, and Gabriel J Brostow. Patch based synthesis for single depth image super-resolution. In *European conference on computer vision*, pages 71–84. Springer, 2012.

[26] Fangchang Mal and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2018.

[27] Kiyoshi Matsuo and Yoshimitsu Aoki. Depth image enhancement using local tangent plane approximations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3574–3583, 2015.

[28] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, pages 283–291, 2018.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[30] Radu Bogdan Rusu. *Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments*. PhD thesis, Computer Science department, Technische Universitaet Muenchen, Germany, October 2009.

[31] Sina Shamekhi, Miran Beygi, Mohammad Hossein, Bahareh Azarian, and Ali Gooya. A novel spot-enhancement anisotropic diffusion method for the improvement of segmentation in two-dimensional gel electrophoresis images, based on the watershed transform algorithm. *Iranian Journal of Medical Physics*, 12(3):209–222, 2015.

[32] Jianhong Shen and Tony F Chan. Mathematical models for local nontexture inpaintings. *SIAM Journal on Applied Mathematics*, 62(3):1019–1043, 2002.

[33] Shreyas S Shivakumar, Ty Nguyen, Steven W Chen, and Camillo J Taylor. Dfusenet: Deep fusion of rgb and sparse depth information for image guided dense depth completion. *arXiv preprint arXiv:1902.00761*, 2019.

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

[35] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Proceedings of the IEEE international conference on computer vision*, volume 98, page 2, 1998.

[36] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 International Conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017.

[37] Jingyue Wang and Weizhang Huang. Image segmentation with eigenfunctions of an anisotropic diffusion operator. *IEEE Transactions on Image Processing*, 25(5):2155–2167, 2016.

[38] Weiming Wang, Lei Zhu, Jing Qin, Yim-Pan Chui, Bing Nan Li, and Pheng-Ann Heng. Multiscale geodesic active contours for ultrasound image segmentation using speckle reducing anisotropic diffusion. *Optics and Lasers in Engineering*, 54:105–116, 2014.

[39] Joachim Weickert. *Anisotropic diffusion in image processing*, volume 1. Teubner Stuttgart, 1998.

[40] Hongyang Xue, Shengming Zhang, and Deng Cai. Depth image inpainting: Improving low rank matrix completion with low gradient regularization. *IEEE Transactions on Image Processing*, 26(9):4311–4320, 2017.

[41] Zhenheng Yang, Peng Wang, Wei Xu, Liang Zhao, and Ramakant Nevatia. Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[42] Yongjian Yu and Scott T Acton. Speckle reducing anisotropic diffusion. *IEEE Transactions on image processing*, 11(11):1260–1270, 2002.

[43] Jianjun Yuan and Lipei Liu. Anisotropic diffusion model based on a new diffusion coefficient and fractional order differential for image denoising. *International Journal of Image and Graphics*, 16(01):165, 2016.

[44] Juyong Zhang, Jianmin Zheng, and Jianfei Cai. A diffusion approach to seeded image segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2125–2132. IEEE, 2010.

[45] Shuangli Zhang, Weijian Xie, Guofeng Zhang, Hujun Bao, and Michael Kaess. Robust stereo matching with surface normal prediction. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2540–2547. IEEE, 2017.

[46] Yinda Zhang and Thomas Funkhouser. Deep depth completion of a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 175–185, 2018.