Indoor Depth Recovery Based on Deep Unfolding with Non-Local Prior

Yuhui Dai[1,2†], Junkang Zhang[1,2†], Faming Fang[1,2*], Guixu Zhang[1,2]

[1]School of Computer Science and Technology, East China Normal University
[2]Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University
{51215901067, 52215901001}@stu.ecnu.edu.cn {fmfang, gxzhang}@cs.ecnu.edu.cn

Abstract

In recent years, depth recovery based on deep networks has achieved great success. However, the existing state-of-the-art network designs perform like black boxes in depth recovery tasks, lacking a clear mechanism. Utilizing the property that there is a large amount of non-local common characteristics in depth images, we propose a novel model-guided depth recovery method, namely the DC-NLAR model. A non-local auto-regressive regular term is also embedded into our model to capture more non-local depth information. To fully use the excellent performance of neural networks, we develop a deep image prior to better describe the characteristic of depth images. We also introduce an implicit data consistency term to tackle the degenerate operator with high heterogeneity. We then unfold the proposed model into networks by using the half-quadratic splitting algorithm. This proposed method is experimented on the NYU-Depth V2 and SUN RGB-D datasets, and the experimental results achieve comparable performance to that of deep learning methods.

1. Introduction

Dense depth recovery from sparse depth maps is crucial for various applications, including human-computer interaction [23], scene reconstruction [24], augmented realities [14] and autonomous driving. Therefore, depth recovery is currently a significant research area in the field of computer vision. Accurate depth maps have been shown to provide necessary 3D information for many computer vision tasks, including semantic labeling [19, 32], robot navigation [9], 3D reconstruction [27, 21], and so on. While high-quality texture information is easily captured by modern color cameras, the acquisition of depth information remains a challenging task under realistic conditions. Although some sensors can directly acquire depth information for indoor scenes, they often suffer from reflection and missing pixels on transparent surfaces, leading to inaccurate depth maps. Therefore, the estimation of the missing depth information for sparse depth maps has been extensively studied. In contrast to depth images, rgb images provide rich color and texture information. As a result, rgb images corresponding to depth maps are often utilized to guide depth recovery.

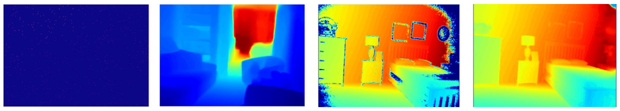
(a) depth samples *(b) ours* *(c) raw depth* *(d) ours*

Figure 1. Examples of indoor depth maps. (a) the sampled points of the original depth map collected by the NYU dataset; (b) the complete depth map by our model; (c) the pseudo depth map from the SUN RGB-D dataset [31]; (d) the corresponding depth recovery result of our model.

By the different starting points, they can be divided into two major categories: traditional model-based methods and data-driven methods. Among the early model-based depth recovery methods, Dong et al. [7] proposed a unified variational method that incorporates joint local and non-local regularization. Xue et al. [39] proposed a low gradient regularization to improve the excessive and spurious details in the restored region, which commonly arise from the low-rank method. This approach enables better characterization of depth images with sparse gradients. Yuan et al. [25] used non-local low rank to model the global similarity structure between depth blocks and combined it with Total variation (TV) [2] to capture the correlation between local depth pixels. These methods transform the depth recovery problem into a mathematical optimization problem that can make full use of the essential information of the image.

The advancements in deep learning, as evidenced by recent studies [4, 15, 18, 29, 41], have showcased the effec-

[†]These authors contributed equally to this work.
[*]Corresponding author.

tiveness of deep learning models in various tasks. Similarly, convolutional neural networks (CNNs) have emerged as powerful tools for depth recovery tasks. Early methods estimated dense depth directly from rgb images and sparse depth images. Ma [20] proposed an encoder-decoder structure to recover dense depth maps from sparse depth maps, guided by rgb images. This work highlights the significance of CNNs in the realm of depth restoration tasks. However, the dense depth maps predicted by previous methods often suffer from inaccuracies. To further generate finer and complete depth maps, a lot of work has emerged recently. Wang et al. [35] proposed an end-to-end GAN-based network that effectively integrates the original depth maps and rgb images to efficiently obtain accurate depth estimates.

Although CNNs have achieved excellent performance in depth recovery tasks, they often overlook the specific characteristics of depth images. On the other hand, traditional model-driven approaches attempt to transform the tasks into mathematically interpretable problems. However, they heavily depend on manually designed parameters during the solution process, which may not guarantee the discovery of an optimal solution. In our task, the challenge lies in recovering the complete depth map from sparse raw depth data using a random selection of a few hundred depth samples or the incomplete original depth map, as shown in Fig. 1. Considering its sparsity and similarity, we propose a deep unfolding model that combines traditional mathematical models with deep networks, and the contributions of our work can be summarized as follows:

- We first propose a deep unfolding model applied to the depth recovery tasks, which integrates the advantages of traditional mathematical models with CNNs to learn more generalized prior information about depth images. To enhance the global understanding of depth images, a non-local auto-regressive regularization term is introduced into our model. This term facilitates the recovery of the depth map by leveraging similarities among depth patches.

- We derive an alternate optimization algorithm for each variable to optimize this model, and then unfold the iterative algorithm into a deep network. Specifically, as the degenerate operator with highly global heterogeneity, we develop a convolutional network to build data consistency term and further integrate it with a gradient descent process.

- Our proposed method has experimented on the NYU-Depth V2 and SUN RGB-D datasets, and the experimental results achieve comparable performance with deep learning-based methods, demonstrating its effectiveness and availability in terms of performance for depth recovery tasks.

2. Related Work

The process of image recovery estimates an unknown image x from its degraded observation y, which can usually be represented by:

$$y = Ax + n, \tag{1}$$

where n denotes the additive noise and A denotes the degradation matrix associated with an image degradation system, which can express different image restoration problems due to different settings of A such as denoising problems with an identical matrix, super-resolution problems with a subsampling matrix/operator. Accordingly, model-based image restoration can be formulated into the following least squares optimization problem:

$$\min_{x} ||y - Ax||_2^2 + \lambda R(x), \tag{2}$$

where y is the initial input image, x is the recovered image, λ is the balance parameter. $R(x)$ is a regularization term, such as total variation (TV) [25], Gaussian mixture model (GMM) [10], K-SVD [1], and BM3D [6]. At the same time, implicit regularization terms are also very popular, which are often represented as implicit denoising prior or as prior information about the sparse depth in depth recovery. The choice of the implicit regularization term reflects different ways of combining prior knowledge about the unknown estimated image x.

To solve the above problem, the half-quadratic splitting (HQS) method [13] converts it into an equivalent bivariate problem:

$$\min_{x,z} ||y - Ax||_2^2 + \lambda R(z) + \frac{\beta}{2}||x - z||_2^2, \tag{3}$$

where z is an intermediate variable and β is a hyperparameter that will increase as the iterative process proceeds. This can be achieved by transforming Eq. (3) into an iteration that solves the following two steps:

$$z^{k+1} = \arg\min_{z} \frac{\beta}{2}||x^k - z||_2^2 + \lambda R(z), \tag{4}$$

$$x^{k+1} = \arg\min_{x} ||y - Ax||_2^2 + \frac{\beta}{2}||x - z^{k+1}||_2^2. \tag{5}$$

The implied prior term $R(\cdot)$ indicates the properties of the image, such as sharpness, completeness of restoration, etc. In some deep unfolding methods for image restoration, including image denoising [38, 33], image super-resolution [26], image deblurring [8], and video restoration [22], the deep CNN is used as a regularizer in each step, which implicitly learns the deep image prior as in Eq. (4). The fundamental concept behind deep unfolding networks is that the traditional iterative soft thresholding algorithm (ISTA)

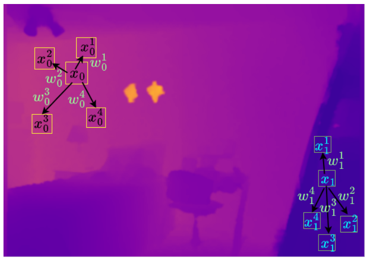

Figure 2. Illustration of the NLAR Module - a non-local extension of classic auto-regressive (AR) model for depth images.

used in sparse coding can be equivalently represented by a series of recurrent neural networks [37]. Building upon this idea, Gong et al. [11] introduced a learning-based approach to train a general gradient descent optimizer, constructing a recurrent gradient descent network (RGDN) for image denoising. Another related work by Fang et al. [8] introduced a kernel error term to validate the given blurring kernel, typically estimated from the observed image. They also incorporated a deep learning denoiser prior to preserve fine textures in the recovered image.

Although the image recovery model has a wide range of applications in natural image reconstruction, it has restrictions on accurately recovering depth information due to the high sparsity of the input. Note that these above models are designed to extract the local feature, which pays less attention to the area with similar information. As shown in Fig. 2, we can observe that there are many non-local regions with similar depth information in the depth image. Therefore, a reasonable global prior urgently needs to be introduced to recover more non-local information and structure.

3. Proposed Method

In this section, we present the DC-NLAR model, which incorporates a non-local auto-regressive regularization term into the traditional image degradation model. Additionally, we employ a deep unfolding strategy to effectively solve the formulated model problem for depth image recovery.

3.1. Non-Local Auto-Regressive Module

The basic idea of the non-local auto-regressive (NLAR) model is to extend the traditional auto-regressive (AR) model by redefining the neighborhoods. For a given patch x_i, the model seeks its sparse linear decomposition over a set of non-local (rather than local) neighborhoods. The following representation is available:

$$x_i \approx \sum_j w_i^j x_i^j, \tag{6}$$

where x_i^j denotes the j-th similar patch found in the non-local neighborhood, and these j similar patches together form all patches with similar structures to patch x_i.

The w_i^j in the above equation denotes the auto-regressive coefficient of the j-th similar patch in the non-local neighborhood of patch x_i, and we can represent the non-local auto-regressive model of image x in the following way:

$$x \approx Sx. \tag{7}$$

The matrix S in the non-local auto-regressive model is expressed as follows:

$$S_{i,j} = \begin{cases} w_i^j, & if\ x_i^j\ is\ a\ nonlocal\ neighbor\ of\ x_i; \\ 0, & otherwise. \end{cases} \tag{8}$$

Calculating similarity among the non-local neighbors in Eq. (7) can be implemented by NLAR module [36]. The output of NLAR (Sx) is expressed by:

$$Sx = \frac{\sum_{\forall j} f(x_i, x_j) g(x_j)}{\sum_{\forall j} f(x_i, x_j)}, \tag{9}$$

where $f(\cdot, \cdot)$ is the function to calculate the similarity between x_i and x_j. The following Gaussian function is used in [36] to define the similarity function f:

$$f(x_i.x_j) = e^{(\theta(x_i)^T \phi(x_j))}, \tag{10}$$

where $\phi(x_{i,p}^k) = W_\phi x_{i,p}^k$, $\theta(x_i^k) = W_\theta x_i^k$, $g(x_j) = W_g x_j$ and W_ϕ, W_θ, W_g are the weight matrices, which are learned by convolutional networks represented by θ, ϕ, and g in Fig. 4. Then the NLAR model can be written as follows:

$$E_{NLAR}(x) = \sum_i ||x_i - \sum_j S_{i,j} x_i^j||_2^2. \tag{11}$$

Since depth images usually contain rich repetitive depth information, non-local similarity has been shown to be effective in recovering hard-to-recover depth information lost in depth recovery models, such as black objects, car glass, etc.

3.2. Model Proposal and Optimization

Inspired by the NLAR model, we develop a novel depth recovery model by coupling the NLAR regularization term with the image degradation model, which is expressed as:

$$\min_x ||y - Ax||_2^2 + \lambda R(x) + \beta E_{NLAR}(x), \tag{12}$$

where A is the degenerate operator, the sparse depth map is the degradation (with non-uniform downsampling) of distance projection, which makes the problem of depth recovery unique. $R(x)$ is an implicit prior. To learn more general prior for x, it is better to set the implicit prior captured by the neural network parametrization. As has been mentioned above, the problem (12) can be optimized by the HQS method, which gives the following iteration solution:

$$z^{k+1} = \arg\min_z \frac{\gamma}{2} ||x^k - z||_2^2 + \lambda R(z), \tag{13}$$

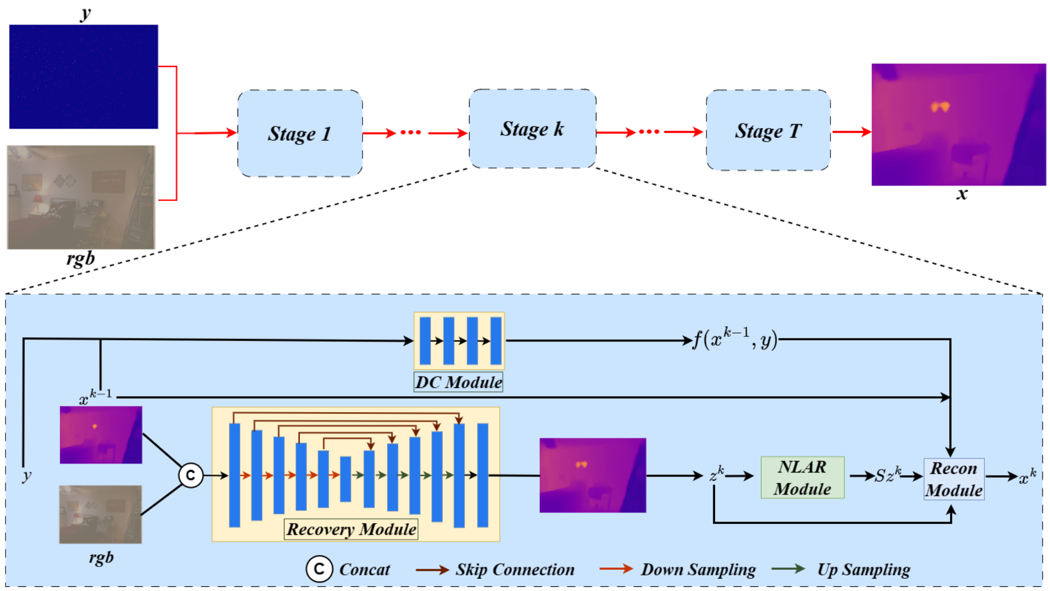

Figure 3. The overall architecture of the proposed network. The initial sparse depth x^{k-1} and the corresponding rgb images are input, and the model goes through four associated modules in each round of iteration: Recovery module, DC module, NLAR module, and Recon Module, and the recovered depth x^k of this round is input to the next round, so that the final predicted depth x is obtained after T rounds of iteration.

$$x^{k+1} = \arg\min_{x} ||y-Ax||_2^2+\beta E_{NLAR}(x)+\frac{\gamma}{2}||x-z^{k+1}||_2^2. \tag{14}$$

Inspired by the effectiveness of end-to-end networks in image restoration tasks, using the efficient learning ability of deep CNNs, a neural network module is used to solve the implicit regularization term in subproblem (13), which can be better guided by the corresponding rgb images to recover more accurate depth values. Thus, the solution of the z-subproblem can be expressed as:

$$z^{k+1} = Recovery\ \ Module(rgb, x^k). \tag{15}$$

While Eq. (14) can be viewed as the least squares optimization problem, it is solved in an approximate form. Let $f(x,y) = ||y-Ax||_2^2$, then the explicit solution of x^{k+1} can be performed directly using the single step gradient descent method:

$$\begin{aligned} x^{k+1} &= x^k - \delta[\bigtriangledown_x f + \beta(x^k - Sz^{k+1}) + \gamma(x^k - z^{k+1})] \\ &= \overline{A}x^k + \delta\beta S z^{k+1} + \delta\gamma z^{k+1} - \delta \bigtriangledown_x f(x^k, y), \end{aligned} \tag{16}$$

where $\overline{A} = (I - \delta\beta - \delta\gamma)$, and δ, β, γ are hyperparameters.

Since depth recovery is different from the general image restoration problem, the degenerate operator A has high complexity, whereas the super-resolution, image denoising, and other problem images are spatially uniformly sampled. Thus, with a high heterogeneity degenerate operator A, we can take full advantage of the deep networks to efficiently learn the data fidelity term. To better model the physical generation mechanism, we replace $\bigtriangledown_x f(x^k, y)$ by a nested network $F(x^k, y, \Theta_x)$. Therefore, Eq. (14) can be written as follows:

$$x^{k+1} = \overline{A}x^k + \delta\beta S z^{k+1} + \delta\gamma z^{k+1} - \delta F(x^k, y, \Theta_x). \tag{17}$$

where Θ_x is the parameter in the network. Note that without missing the interpretability, a data-driven strategy is adopted to predict the gradient of the data fidelity term to acquire better performance.

3.3. Designs of the Network

As mentioned earlier, the subproblems with implicit prior terms can be solved with deep unfolding networks. In general, we employ an altered encoder-decoder structure as the backbone of our module in this paper. As shown in the Recovery Module of the Fig. 3, a variant U-Net is used to solve Eq. (13) with the input initial sparse depth image and the corresponding rgb image to guide the recovery of the sparse depth, using a network with an encoder-decoder structure, where each layer is a convolution of 3×3 kernel. The channels at the encoder stage are increased from 64 to 2048, which are doubled while the size is downsampled to $\frac{1}{2}$ of the original size at the same time. Relatively at the decoder stage, the channels are restored from 2048 to 2 while upsampling to the original size, and the final output is the restored depth map.

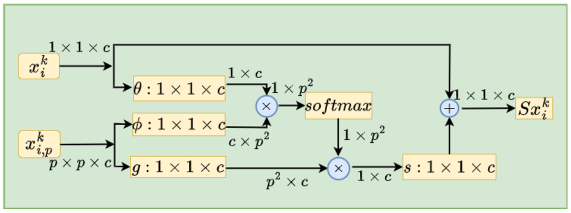

Figure 4. The architecture of the NLAR Module which designed for computing the similarity of a given image. c denotes the channel of the input feature map [36]. ”$\otimes$” denotes matrix multiplication. ”$\oplus$” denotes element-wise sum. We calculate the $p \times p$ block centered at position i. x_i^k denotes the k-th similar patch found in the non-local neighborhood, and $x_{i,p}^k$ is the center of the patch that calculates the similarity with x_i^k. θ, ϕ, and g represent convolutional networks to learn the weight matrices W_ϕ, W_θ, and W_g.

The NLAR module corresponds to the expansion of the NARM matrix S into a network implementation. Based on the observation that natural images usually contain rich repetitive structures, non-local similarity shows effectiveness for recovering missing low-frequency information in sparse depth maps. In model-based implementations, finding similar patches is often computationally problematic, as nearest neighbor search is an NP-hard problem [16]. In contrast, computing non-local relations between image patches can be efficiently implemented in parallel by non-local neural networks. Inspired by the design of non-local operations and their application in image restoration, we design a fast non-local operation module for computing the NARM matrix S. Fig. 4 shows the block diagram of the implementation of the non-local operation designed for computing the similarity of a given image.

It is worth noting that the data fidelity term we built in the previously introduced model $F(x^k, y, \Theta_x)$ using a simple 4-layers convolutional network module to compensate for the highly heterogeneous nature of the degenerate operator A, which corresponds to the DC module in Fig. 3. Input x^k into the network after concatenating it with the initial x^0. Through the four layers network, the change of channel is 2-32-32-1. Each layer of the network contains Convolution, BatchNorm2d, and ReLU nonlinearity.

The solution of Eq. (17) is represented by Recovery Module in the network structure. The z^{k+1}, Sk^{k+1} and $F(x^k, y, \Theta_x)$ obtained from the output of the recovery module, NLAR module, and DC module, respectively, are substituted into the Eq. (17) solved by gradient descent to obtain the recovered x^{k+1}. The final solution process for each subproblem is shown in Algorithm 1.

3.4. Loss Function

To ensure precise recovery of the dense depth map, we utilize L_1 and L_2 loss functions during the training process of our model. Additionally, in order to enhance the reconstruction of object edges within the image, we incorporate a gradient loss term, denoted as L_{grad}. These loss functions are applied to both the local depth map and the final prediction. The comprehensive loss function is defined as follows:

$$L_{overrall} = \lambda_1 L_1(d_{pre}, d_{gt}) + \lambda_2 L_2(d_{pre}, d_{gt}) + \lambda_{grad} L_{grad}(d_{pre}, d_{gt}), \tag{18}$$

where λ_1, λ_2 and λ_{grad} are the weight hyperparameters of the different terms in the loss function, all of which are set to 1. L_{grad} is designed to track the depth variation between adjacent pixels [12], which can be mathematically represented as:

$$L_{grad} = || \bigtriangledown_x d_{gt} - \bigtriangledown_x d_{pre} ||_1 + || \bigtriangledown_y d_{gt} - \bigtriangledown_y d_{pre} ||_1 + || \bigtriangledown_{diag} d_{gt} - \bigtriangledown_{diag} d_{pre} ||_1, \tag{19}$$

where $\bigtriangledown_x, \bigtriangledown_y, \bigtriangledown_{diag}$ denote the gradients along the horizontal (denoted by x), vertical (denoted by y) and diagonal (denoted by $diag$) directions [40], respectively. Unlike other works, we introduce the diagonal component into the gradient calculation. Considering the complexity of real world object shapes, L_{grad} is able to further penalize small structural errors and improve the fine details of the depth map. Experiments show that this loss is effective for obtaining clean edges.

Algorithm 1 **Deep Unfolding Network**

```
Input: the sparse depth map $y$, and the corresponding image $rgb$.
  for k = 1 to $T$ do
    $z^{k+1} = Recovery\ Module(rgb, x^k)$    Eq. (15)
    $Sz^{k+1} = NLAR\ Module(z^{k+1})$    Eq. (12)
    $F(x^k, y) = DC\ Module(x^k, y)$    Eq. (16)
    $x^{k+1} = (I - \delta\beta - \delta\gamma)x^k + \delta\beta Sz^{k+1}$
        $+\delta\gamma z^{k+1} - \delta F(x^k, y)$    Eq. (17)
  end for
Output: the recovered depth map $x^T$.
```

4. Experiments

4.1. Datasets and Metrics

We conducted experiments on two widely-used benchmarks: NYU-Depth V2 [30] and SUN RGB-D dataset [31]. **NYU-Depth V2**. The NYU-Depth V2 dataset [30] contains pairs of rgb and depth images collected from Microsoft Kinect in 464 indoor scenes. Following existing methods, we utilize the unlabeled 50K images for training and the labeled 654 images in the test set for evaluation. In the process of training, the initial input is a sparse depth map with 500 valid depth pixels which are randomly drawn from the

Method	RMSE↓	Rel↓	$\delta_{1.25}\uparrow$	$\delta_{1.25^2}\uparrow$	$\delta_{1.25^3}\uparrow$
CSPN [5]	0.117	0.016	99.2	99.9	100.0
GAENet [3]	0.114	0.018	99.3	99.9	100.0
DeepLidar [29]	0.115	0.022	99.3	99.9	100.0
NLSPN [28]	**0.092**	**0.012**	**99.6**	99.9	100.0
GuideNet [34]	0.101	0.015	99.5	99.9	100.0
ACMNet [41]	0.105	0.105	99.4	99.9	100.0
PRR [18]	0.104	0.014	99.4	99.9	100.0
RDF-GAN [35]	0.103	0.016	99.4	99.9	100.0
Ours	0.102	0.014	99.4	99.9	100.0

Table 1. Quantitative comparison with state-of-the-art results on NYU-Depth V2 dataset benchmark.

Method	RMSE↓	Rel↓	$\delta_{1.25}\uparrow$	$\delta_{1.25^2}\uparrow$	$\delta_{1.25^3}\uparrow$
CSPN [5]	0.153	0.079	97.5	99.0	99.5
NLSPN [28]	**0.101**	**0.023**	98.4	99.3	99.6
Ours	0.104	0.026	**98.5**	99.3	99.6

Table 2. Quantitative comparison results on SUN RGB-D dataset.

reconstructed depth map. The input images are resized to 320×240 and center-cropped cropped to 320×224.

SUN RGB-D. The SUN RGB-D dataset [31] contains 10,335 RGB-D images captured by four different sensors. We use 5,283 images for training and 5,049 for testing. We use the refined depth map based on multiple frames as the ground-truth for evaluation. The input images are centering cropped to 320 × 224.

Evaluation Metrics. Three widely used assessment metrics are used for the depth recovery evaluation: root mean squared error (RMSE), absolute relative error (Rel), and δ_i, which is the percentage of predicted pixels whose relative error is within a relative threshold [20].

Parameter Settings. Our training was implemented by Pytorch with 2 NVIDIA GTX3090 GPUs and set batch size to 12. In our current implementation, we used ADAM [17] as the optimization algorithm. The learning rate starts at 1×10^{-3} and the warm-up strategy is used for the first epoch. Starting from the 10th epoch, the learning rate from the 10th to the 15th epoch is reduced to 2×10^{-4}. Then the learning rate stays 4×10^{-5} after the start of the 16th epoch. The other parameters are all the same with (β_1, β_2) = (0.9, 0.999). In this paper, we set the unrolling stage T to 4.

4.2. Comparison with SOTA Methods

The performance comparison of our method with other state-of-the-art methods on NYU-Depth V2 is shown in Table 1. Among the listed methods, CSPN [5], GAENet [3], DeepLidar [29], NLSPN [28], and RDF-GAN [35] etc. are the classical network structures. In the comparison of RMSE and $\delta_{1.25}$, we are only worse than GuideNet [34] except for NLSPN [28], but we are better than it in Rel this metric. Other than that, all of our metrics are achieving highly competitive results.

The visualizations in Fig. 5 demonstrate the superior performance of our method. Our results excel not only in recovering large, smooth areas but also in capturing fine details compared to other methods. For instance, in the first row, our method accurately recovers the depth value of the chair and successfully captures the true contour of the vacant area, outperforming the CSPN [5] method. Similarly, in the second row, our results achieve the highest accuracy compared to the ground-truth, showcasing our ability to recover detailed information effectively. Even in complex scenarios, such as the stacked regions shown in the third row, our method closely approximates the true contours. Moreover, in images with challenging distant fields of view, like the hanging fan in the fourth row, our method significantly outperforms other techniques. These experiments demonstrate that our model not only recovers ground regions more comprehensively but also exhibits greater robustness and accuracy in capturing object contours, yielding satisfactory results.

Table 2 presents the performance comparison of our method on the SUN RGB-D dataset. While the NLSPN [28] method performs well in terms of RMSE and Rel, utilizing a coarse-to-fine approach with non-local spatial propagation and confidence-incorporated learnable affinity normalization, our proposed method outperforms in terms of δ_i. This indicates that our method achieves predictions that closely align with accurate values, exhibiting minimal overall deviations and reflecting the method's stability in the global region. The visualizations in Fig. 6 further showcase the accuracy and completeness of our method, not only in depth recovery but also in capturing edges effectively.

4.3. Ablation Study

In this section, we perform ablation studies on the NYU-Depth V2 dataset, which are divided into four main influencing factors: the patch size of the NLAR model, the effect of L_{grad}, the effect of NLAR term, and the DC module.

The Number of Patch Size. In order to verify the patch size of the NLAR model, we set different sizes as $\{7, 9, 11\}$ when other influencing factors were constant. The experimental results are shown in Table 4. The results show that there is no significant effect on δ_i, and comparing the results of RMSE and Rel metrics, it is clear that the optimal performance is achieved when the size is 7. To balance performance effects and computational complexity, the final size chosen for the experiment is 7.

The Number of Iterations. In order to verify the number of iterations, we set different sizes as $\{2, 4, 6\}$ when other influencing factors were constant. The experimental results are shown in Table 5. Note that the optimal performance is achieved when the number is 4 in terms of RMSE and Rel metrics. It can be seen that the metric performance

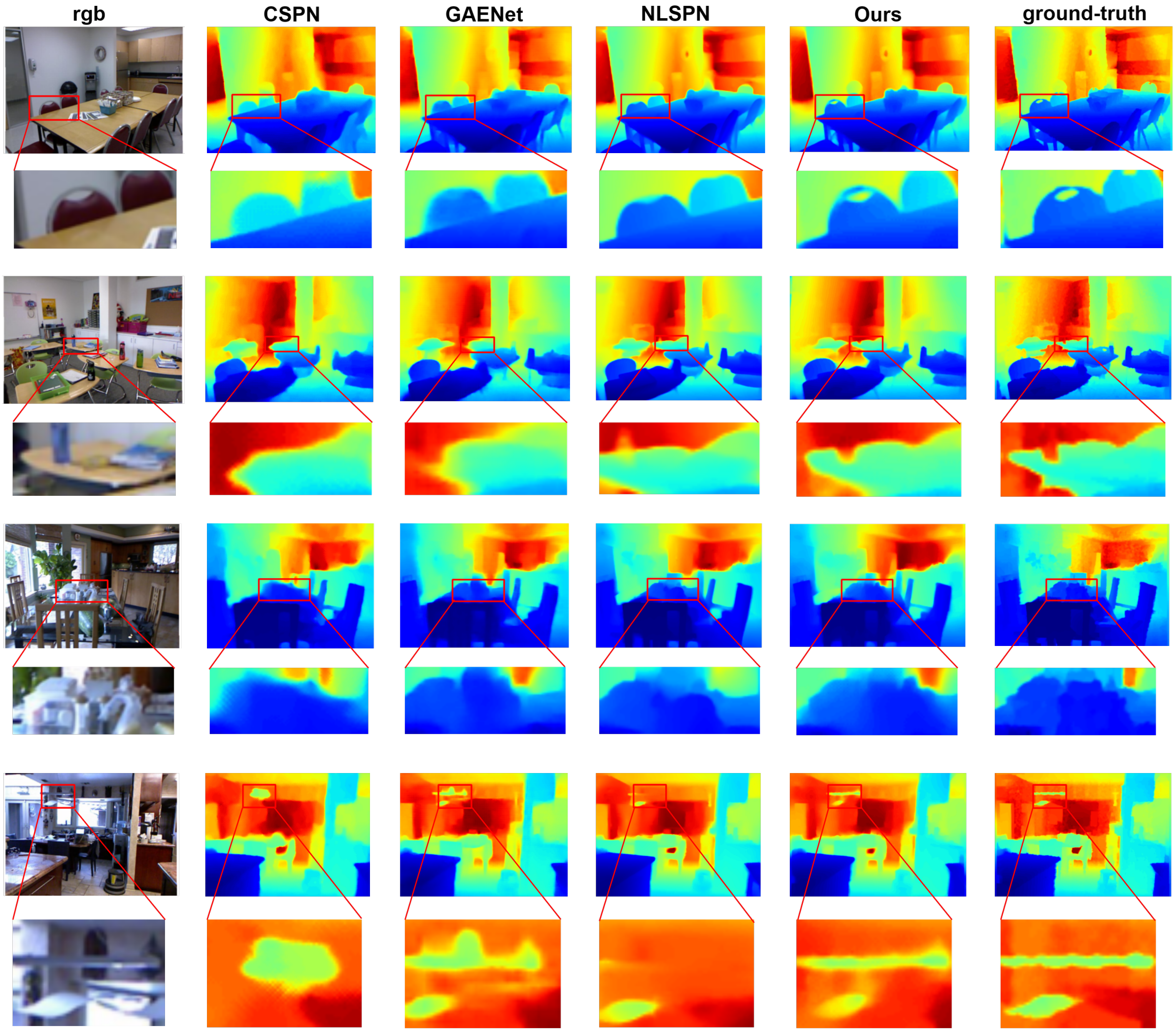

Figure 5. Visual quality comparison on NYU-Depth V2 test dataset. Left to right: the corresponding rgb image, CSPN [4], GAENet [3], NLSPN [28], Our model and ground-truth.

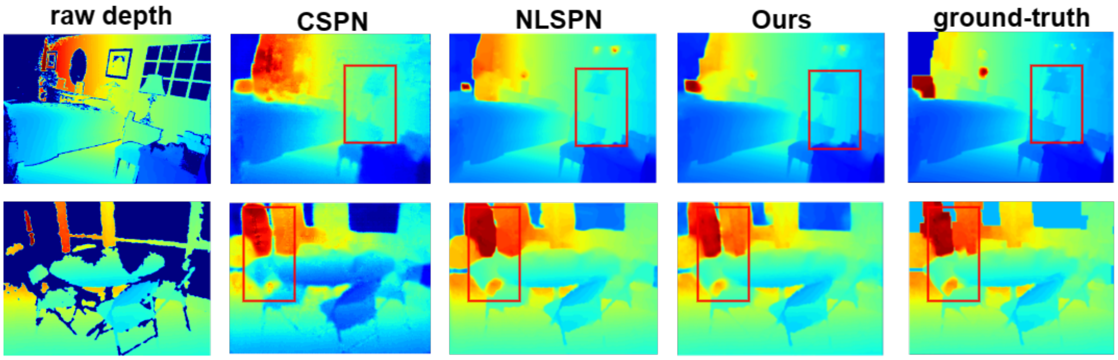

Figure 6. Visual quality comparison on SUN RGB-D test dataset. Left to right: the corresponding raw depth image, CSPN [4], NLSPN [28], Our model, and ground-truth.

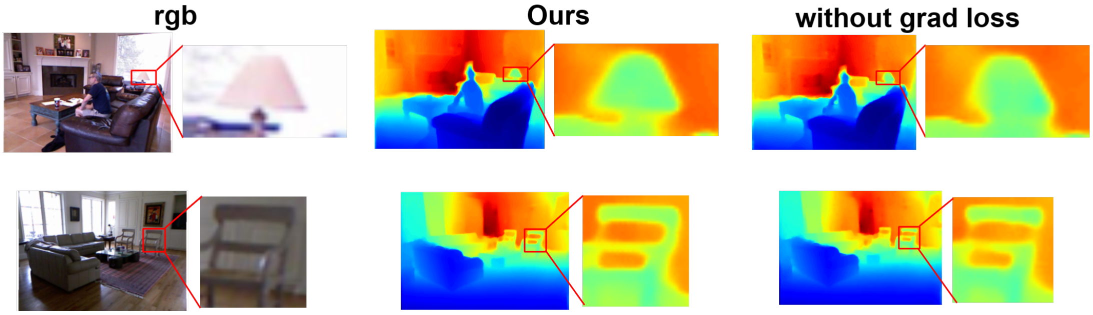

Figure 7. Visual quality comparison on Experiment C. Left to right: the corresponding rgb image, the results with L_{grad} and the results without L_{grad}.

Settings	Recovery	NLAR	L_{grad}	DC Module	RMSE↓	Rel↓	$\delta_{1.25}\uparrow$	$\delta_{1.25^2}\uparrow$	$\delta_{1.25^3}\uparrow$
A	√	×	×	-	0.115	0.016	99.3	99.9	100.0
B	√	×	√	√	0.108	0.015	99.4	99.9	100.0
C	√	√	×	√	0.103	0.015	99.4	99.9	100.0
D	√	√	√	×	0.103	0.014	99.4	99.9	100.0
E(complete)	√	√	√	√	0.102	0.014	99.4	99.9	100.0

Table 3. Ablation study of different influencing factors.

Size	RMSE↓	Rel↓	$\delta_{1.25}\uparrow$	$\delta_{1.25^2}\uparrow$	$\delta_{1.25^3}\uparrow$
7	0.102	0.014	99.4	99.9	100.0
9	0.102	0.015	99.4	99.9	100.0
11	0.103	0.015	99.4	99.9	100.0

Table 4. Ablation study of the patch size of the NLAR model.

Iterations	RMSE↓	Rel↓	$\delta_{1.25}\uparrow$	$\delta_{1.25^2}\uparrow$	$\delta_{1.25^3}\uparrow$
2	0.107	0.016	99.3	99.9	100.0
4	0.102	0.014	99.4	99.9	100.0
6	0.109	0.015	99.4	99.9	100.0

Table 5. Ablation study of the number of iterations.

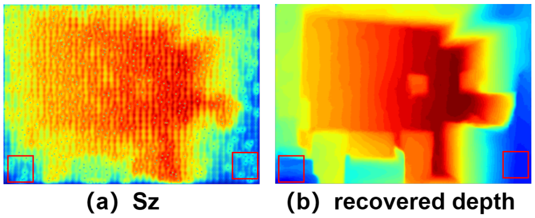
(a) Sz (b) recovered depth

Figure 8. (a) is the visualization of the output of Sz in the NLAR model, and (b) is the corresponding recovered depth image.

increases as the number of iterations increases, but the improvement of the RMSE and Rel values reaches saturation when T = 4. To balance performance effects and computational complexity, the final number of iterations chosen for the experiment is 4.

Different Settings of the Model. To verify the effect and performance of different model settings, experiments under four different settings of A, B, C, and D were designed, and the results are shown in Table 3.

Experiment A which only reserves the Recovery Module is the baseline of our ablation experiments. We can observe that the baseline is able to get improvement by adding additional modules.

Furthermore, experiment B is to verify the effect of the improved NLAR Module. Compared setting B with setting E, we can clearly see that the RMSE increases by 5.9% and the Rel value increases by 7.1% once the NLAR Module is removed, which also indicates the effectiveness. The output Sz from the NLAR model is visualized in Fig. 8. At the locations marked in the figure, it can be seen that this module is effective in finding regions that are not adjacent to each other with similar structures. We add Eq. (11) to the model as a regular term. The mathematical expression of x in the final recovered image is related to Sz. Since depth images usually contain rich similar depth information, Sz can use the NLAR module to get accurate results. The blue part we marked in Fig. 8 is considered to have similar depth information, and the points in the same part are the calculated pixel points with high similarity.

Experiment C not only shows the improvement in metrics by adding L_{grad} to the training process, but also demonstrates that the proposed model with L_{grad} loss provides better edge information and complete shape as shown in Fig. 7 (with L_{grad}), which also illustrates the effectiveness of gradient loss.

Experiment D is demonstrated to evaluate the influence of the DC Module. We replace the DC Module with an explicit fidelity term, which contains a 0 or 1 confidence matrix. In particular, the confidence matrix A is obtained by the initial depth. For example, its value is 1 when the input contains a depth value, and 0 otherwise. It can be seen from Table 3 that our model with implicit fidelity term modeling by convolutional networks has improved in terms of RMSE. These reveal that our model can weaken the effect of the degenerate operator A and get satisfactory results.

5. Conclusion

In this paper, we present the DC-NLAR model, a novel approach for depth unfolding that combines the image recovery model with the principle of non-local auto-regression. Our model addresses the challenges posed by the generalization limitations of deep networks and the complexity inherent in traditional mathematical models for achieving optimal solutions. By leveraging the strengths of both approaches, our model introduces fresh ideas and strategies to tackle the problem of depth recovery. Extensive experimental results are proven to show that our method has achieved comparable performance to existing deep network methods on the NYU-Depth V2 and SUN RGB-D datasets. Note that our method has only been applied to indoor scenes at the present time. Further improvements in the network and model can be considered in the future to achieve better results, and applied to large outdoor scenes such as the KITTI dataset.

Acknowledgements

This work was supported by the East China Normal University Graduate International Conference Special Fund, the National Key R&D Program of China (2022ZD0161800), the NSFC-RGC (61961160734), the National Natural Science Foundation of China (62271203), the Shanghai Rising-Star Program (21QA1402500), and the Science and Technology Commission of Shanghai Municipality (Grant No. 22DZ2229004).

References

[1] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

[2] Antonin Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical imaging and vision*, 20:89–97, 2004.

[3] Hu Chen, Hongyu Yang, Yi Zhang, et al. Depth completion using geometry-aware embedding. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8680–8686. IEEE, 2022.

[4] Xinjing Cheng, Peng Wang, Chenye Guan, and Ruigang Yang. Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In *2020 AAAI Conference on Artificial Intelligence*, volume 34, pages 10615–10622, 2020.

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

[6] Aram Danielyan, Vladimir Katkovnik, and Karen Egiazarian. Bm3d frames and variational image deblurring. *IEEE Transactions on Image Processing*, 21(4):1715–1728, 2011.

[7] Weisheng Dong, Guangming Shi, Xin Li, Kefan Peng, Jinjian Wu, and Zhenhua Guo. Color-guided depth recovery via joint local structural and nonlocal low-rank regularization. *IEEE Transactions on Multimedia*, 19(2):293–301, 2016.

[8] Yingying Fang, Hao Zhang, Hok Shing Wong, and Tieyong Zeng. A robust non-blind deblurring method using deep denoiser prior. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 735–744, 2022.

[9] Péter Fankhauser, Michael Bloesch, Diego Rodriguez, Ralf Kaestner, Marco Hutter, and Roland Siegwart. Kinect v2 for mobile robot navigation: Evaluation and modeling. In *2015 International Conference on Advanced Robotics (ICAR)*, pages 388–394. IEEE, 2015.

[10] Jianzhou Feng, Li Song, Xiaoming Huo, Xiaokang Yang, and Wenjun Zhang. Image restoration via efficient gaussian mixture model learning. In *2013 IEEE International Conference on Image Processing*, pages 1056–1060. IEEE, 2013.

[11] Dong Gong, Zhen Zhang, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, and Yanning Zhang. Learning deep gradient descent optimization for image deconvolution. *IEEE Transactions on Neural Networks and Learning Systems*, 31(12):5468–5482, 2020.

[12] Yong Guo, Qi Chen, Jian Chen, Junzhou Huang, Yanwu Xu, Jiezhang Cao, Peilin Zhao, and Mingkui Tan. Dual reconstruction nets for image super-resolution with gradient sensitive loss. *arXiv preprint arXiv:1809.07099*, 2018.

[13] Ran He, Wei-Shi Zheng, Tieniu Tan, and Zhenan Sun. Half-quadratic-based iterative minimization for robust sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 36(2):261–275, 2013.

[14] Aleksander Holynski and Johannes Kopf. Fast depth densification for occlusion-aware augmented reality. *ACM Transactions on Graphics (ToG)*, 37(6):1–11, 2018.

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

[16] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Annual ACM Symposium on Theory of Computing*, pages 604–613, 1998.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] Byeong-Uk Lee, Kyunghyun Lee, and In So Kweon. Depth completion using plane-residual representation. In *2021*

IEEE Conference on Computer Vision and Pattern Recognition, pages 13916–13925, 2021.

[19] Beyang Liu, Stephen Gould, and Daphne Koller. Single image depth estimation from predicted semantic labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1253–1260. IEEE, 2010.

[20] Fangchang Ma 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 4796–4803. IEEE, 2018.

[21] Qingwei Mi and Tianhan Gao. 3d reconstruction based on the depth image: A review. In *2022 International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)*, pages 172–183. Springer, 2022.

[22] Antoine Monod, Julie Delon, Matias Tassano, and Andrés Almansa. Video restoration with a deep plug-and-play prior. *arXiv preprint arXiv:2209.02854*, 2022.

[23] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136. IEEE, 2011.

[24] Trong-Nguyen Nguyen, Huu-Hung Huynh, and Jean Meunier. 3d reconstruction with time-of-flight depth camera and multiple mirrors. *IEEE Access*, 6:38106–38114, 2018.

[25] Uyen Nguyen, Truong Giang Tong, Tat Thang Hoa, Van Ha Tang, et al. A non-local low rank and total variation approach for depth image estimation. In *2021 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–6. IEEE, 2021.

[26] Qian Ning, Weisheng Dong, Guangming Shi, Leida Li, and Xin Li. Accurate and lightweight image super-resolution with model-guided deep unfolding network. *IEEE Journal of Selected Topics in Signal Processing*, 15(2):240–252, 2020.

[27] Hailong Pan, Tao Guan, Yawei Luo, Liya Duan, Yuan Tian, Liu Yi, Yizhu Zhao, and Junqing Yu. Dense 3d reconstruction combining depth and rgb information. *Neurocomputing*, 175:644–651, 2016.

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

[29] Jiaxiong Qiu, Zhaopeng Cui, Yinda Zhang, Xingdi Zhang, Shuaicheng Liu, Bing Zeng, and Marc Pollefeys. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In *2019 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3313–3322, 2019.

[30] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. *2012 European Conference on Computer Vision (ECCV)*, 7576:746–760, 2012.

[31] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, 2015.

[32] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1746–1754, 2017.

[33] Jian Sun and Marshall F Tappen. Learning non-local range markov random field for image restoration. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2745–2752. IEEE, 2011.

[34] Jie Tang, Fei-Peng Tian, Wei Feng, Jian Li, and Ping Tan. Learning guided convolutional network for depth completion. *IEEE Transactions on Image Processing*, 30:1116–1129, 2020.

[35] Haowen Wang, Mingyuan Wang, Zhengping Che, Zhiyuan Xu, Xiuquan Qiao, Mengshi Qi, Feifei Feng, and Jian Tang. Rgb-depth fusion gan for indoor depth completion. In *2022 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6209–6218, 2022.

[36] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, 2018.

[37] Scott Wisdom, Thomas Powers, James Pitton, and Les Atlas. Building recurrent networks by unfolding iterative thresholding for sequential sparse recovery. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4346–4350. IEEE, 2017.

[38] Jingzhao Xu, Mengke Yuan, Dong-Ming Yan, and Tieru Wu. Deep unfolding multi-scale regularizer network for image denoising. *Computational Visual Media*, 9(2):335–350, 2023.

[39] 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.

[40] Xin Zhang, Rabab Abdelfattah, Yuqi Song, Samuel A Dauchert, et al. Depth monocular estimation with attention-based encoder-decoder network from single image. *arXiv preprint arXiv:2210.13646*, 2022.

[41] Shanshan Zhao, Mingming Gong, Huan Fu, and Dacheng Tao. Adaptive context-aware multi-modal network for depth completion. *IEEE Transactions on Image Processing*, 30:5264–5276, 2021.