Super-Resolution of Multi-Observed RGB-D Images Based on Nonlocal Regression and Total Variation

Qingzheng Wang, Shuai Li, Hong Qin, *Senior Member, IEEE*, and Aimin Hao

***Abstract*— There is growing demand for accuracy in image processing and visualization, and the super-resolution (SR) technique for multi-observed RGB-D images has become popular, because it provides space-redundant information and produces a detailed reconstruction even with a large magnification factor. This technique has been thoroughly investigated in recent years. Nevertheless, technical challenges remain, such as finding subpixel correspondences with low-resolution (LR) observations, exploiting space-redundant information, formulating space homogeneity constraints, and leveraging cross-image similarities in structures. To address these challenges, this paper proposes a unified optimization framework to estimate both the super-resolved RGB image and the super-resolved depth image from the multi-observed LR RGB-D images using their correlations. Using depth-assisted cross-image correspondences, the RGB image SR problem is formulated as an effective regularization function by incorporating the normalized bilateral total variation regularizer, and it is efficiently solved by a first-order primal-dual algorithm. The depth image SR estimate can be obtained by minimizing a nonlocal regression-based energy, which integrates the structural cues of the super-resolved RGB image in a detail-preserving fashion. Essentially, our unified optimization framework uses the RGB image and depth image as a priori knowledge that the SR process uses for better accuracy. Our extensive experiments on public RGB-D benchmarks and real data and our quantitative comparison with several state-of-the-art methods demonstrate the superiority of our method in terms of accuracy, versatility, and reliability of details and sharp feature preservation.**

***Index Terms*— Super-resolution, depth image recovery, RGB-D image, nonlocal regression, normalized bilateral total variation.**

I. Introduction and Motivation

CURRENTLY, the super-resolution (SR) technique is considered a second-generation technique for image restoration. It plays a vital role in image refining for downstream image- or geometry-related applications, such as visualization and 3D printing. Unlike traditional restoration techniques, SR aims to produce high-resolution (HR) results using the intrinsic information within a single image or a sequence of images, and it solves the ill-posed inverse problem caused by several types of degradation (e.g., noise, blur, and LR) [1], [2].

Existing techniques are unable to produce convincing SR results for RGB-D images. For the SR of RGB images, the interpolation-based methods are simple but tend to introduce staircase artifacts and overly blurred edges. The reconstruction-based approaches require multiple observations and resort to registration transformation [3], [4] for their correspondences. However, the obtained correspondences are often inaccurate due to the limited resolution of the observations. Learning-based techniques [8], [36], [37] use the high-frequency information from a training set of HR/LR image pairs, and the selection of the training set and the parameters used for tuning (e.g., number of HR/LR image pairs) are difficult to determine but crucial for the results. For SR of a depth image, existing color-guided methods [11]–[13] require HR RGB guidance, which is not possible for popular devices such as Microsoft Kinect (whose resolution is fixed at 640×480 for RGB-D images). We list the existing challenges for high-quality image SR here:

Firstly, to leverage redundant information, SR approaches based on single images are suitable for small magnification factors but deteriorate quickly as the magnification factor increases. This is because the limited information provided by a single image is not a sufficiently powerful basis for image restoration. Thus, researchers endeavor to use image sequences/external datasets to refine the super-resolved texture details. Such methods take multi-observed images into account, but the registration problem is tough, and there is insufficient 3D redundant information, which limits the accuracy. Hence, we propose a smarter way to leverage multi-observed redundant information to streamline the mapping between LR images and HR results.

Secondly, current methods can be roughly classified into two categories for the prior knowledge exploration: those based on a statistic prior (e.g., gradient profile [6]), and those that use an empirical distribution (e.g., heavy tailed distribution, sparse distribution, total variation.). The static prior can be quickly obtained from the input image and is data-specific, but it is sensitive to noise because of the explicit dependence on pixel value. An empirical distribution can be obtained from natural images and is more universal, but it allows for only slight adjustments in images from different observations, and it introduces blurring and edge artifacts. Therefore, a more effective global prior is needed to improve SR outcomes.

Manuscript received June 26, 2015; revised November 14, 2015 and January 9, 2016; accepted January 20, 2016. Date of publication January 22, 2016; date of current version February 12, 2016. This work was supported in part by the National Science Foundation of USA under Grant IIS-0949467, Grant IIS-1047715, and Grant IIS-1049448 and in part by the National Natural Science Foundation of China under Grant 61190120, Grant 61190121, Grant 61190125, Grant 61532002, and Grant 61300067. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Javier Mateos. *(Corresponding authors: Shuai Li, lishuai@buaa.edu.cn; Aimin Hao, ham_buaa@163.com.)*

Q. Wang, S. Li, and A. Hao are with the State Key Laboratory of Virtual Technology and Systems, Beihang University, Beijing 100191, China.

H. Qin is with Stony Brook University, Stony Brook, NY 11794 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2016.2521180

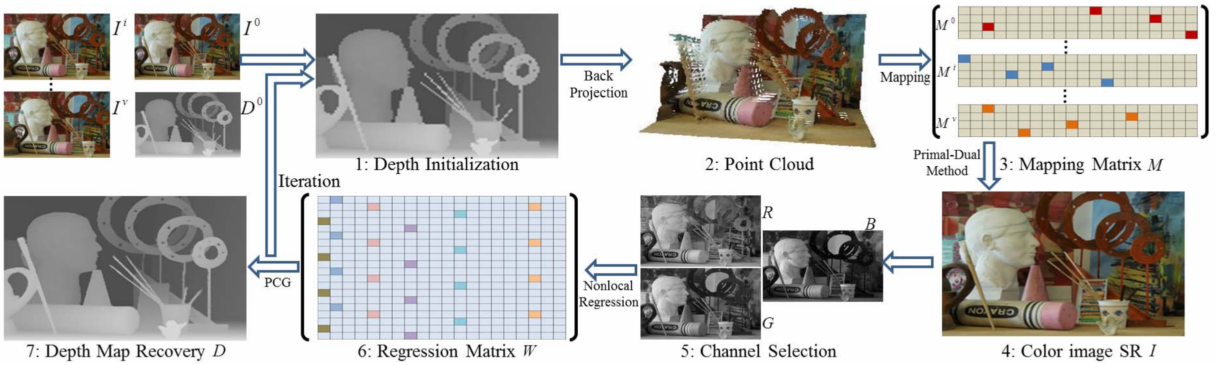


Fig. 1. The pipeline of our RGB-D image super-resolution algorithm, which iteratively estimates HR RGB and depth images. *RGB image SR:* Steps 1-3: the preprocessing steps to establish the mapping matrices between the LR RGB images and the estimated HR image (Section III-B and III-C), of which, step1 initializes the depth image with the same size as the SR depth result; step2 constructs the point cloud by back-projecting the RGB image with the corresponding depth image; and step3 constructs the mapping matrix between shifting viewpoints using the pin-hole model. Step 4: RGB image SR based on the normalized bilateral total variation (NBTV) (Section IV-B and Section IV-D.1); *Depth image SR*: Step 5: select channel based on the variance to adaptively acquire local structure support; Step 6: construct nonlocal regression matrix using depth-color pair (Section IV-C); Step 7: recover color-guided depth image based on nonlocal regression (Section IV-D.2), which is then used to initialize the depth in the next iteration. In step 3 and step 6, the grids are used to index the rows and columns of the matrix. The colors represent non-zero values in the matrix.

Thirdly, concerning the complementary cue formulation of RGB-D image pairs, SR's prospects have improved greatly with the rapid deployment of portable RGB-D capture devices. SR-related research has reached a new phase wherein researchers are simultaneously pursuing super-resolved texture images and depth images because both types of images are indispensable for 3D reconstruction applications that demand high accuracy. Although LR color-depth pair approaches [17], [18] solve the SR problem for RGB-D images by exploring mutual information and incorporating segmentation data to avoid over-smoothed boundaries [14], [15], the accuracy is dependent on the quality of the high-level feature extraction. Additionally, it remains difficult to rigorously formulate the underlying complementary color-depth correlations to boost SR performance on RGB-D images.

To combat the aforementioned problems, we use the correlation of depth-color pairs in 3D to effectively guide the SR of an RGB-D image, as highlighted in Fig. 1. Given multi-observed RGB-D images, we first back-project the RGB image to form 3D point clouds using the corresponding depth image, and establish the mapping matrices between the LR RGB images and the estimated HR image (Steps 1-3). We also define a novel image prior (Normalized Bilateral Total Variation) to prevent the ringing artifacts and the over-blurring problem of TV-based methods (Step 4). Afterwards, the color-guided nonlocal regression is used to enhance the stability and accuracy of HR depth image estimation (Step 7). The adaptive RGB-channel selection strategy helps further enhance the sharpness of the edges (Steps 5-6). In this way, RGB image SR and depth recovery can be elegantly integrated into a unified framework to ensure robustness and accuracy. Specifically, the salient contributions of this paper can be summarized as follows:

- We propose a unified optimization framework to simultaneously estimate the super-resolved RGB and depth images by exploiting the geometrical-structure correlations of multi-observed LR depth-color pairs, which helps boost accuracy.
- We formulate an effective image regularization prior (normalized bilateral total variation) to super-solve the RGB image. The method respects sharp edges and preserves details well, and it can be efficiently solved via a first-order primal-dual optimization algorithm.
- We design an adaptive nonlocal regression function to estimate the HR depth image in an RGB-guided way, which can be naturally integrated into the optimization function to guarantee depth-color consistency and select for stable and accurate structure details.

Note that in previous studies, the input image configuration can be classified into two categories: multiple RGB images + single depth map, and multiple RGB images + multiple depth maps. In the first case, the RGB and depth images boost each other's reconstruction accuracy. Multiple RGB images also provide more redundant information to guarantee a satisfactory RGB SR result. Representative RGB-D super-resolution approaches have been reported previously [19], [20], and our method also falls into this category. In the second case, multiple RGB and depth images are adopted to enhance the spatial and temporal consistency. Several existing methods [15], [16] incorporate multiple depth maps SR by employing the corresponding color images to guide the propagation of the local structures. However, for RGB image SR, incorporating multiple depth maps received much less attentions.

II. Related Work

We briefly review previous works related to three central themes of this paper.

A. RGB Image Super-Resolution

RGB image SR includes a wide spectrum of methods. Interpolation-based method is the simplest and fastest, such as the bicubic method. However, those methods usually give rise to blurred results. Learning-based approach is also important. For example, Yang *et al.* [7], [21] took the LR/HR pairs as a training set to learn LR and HR dictionaries. By taking into account the similarities between the training set and the test set and sharing the sparse reconstruction coefficients, the fine details can be recreated in the HR image. However, the sparse coding requires extensive computation time to find a set of satisfying sparse coefficients. To reduce the time cost, Yang and Yang [23] split the feature space into numerous subspaces and learned a set of linear regression functions to infer the details in an HR image. Zhu *et al.* [22] proposed the concept of a deformable patch, with which the dictionary becomes more compressive and can cover more patterns than previous approaches. Dong *et al.* [24] used convolutional neural networks to learn an end-to-end mapping between the LR and HR patches. Timofte *et al.* [9] introduced a neighbor-embedding method into dictionary learning to pre-compute the projection matrix for each atom; they then mapped the LR patches to HR space using the obtained projection matrix. Dai *et al.* [10] learned a collection of regressors to guarantee the minimal super-resolved error for the training set. By searching for a similar patch in the training set, the testing patch can adaptively leverage the most appropriate regressor to produce the HR patch. The learning-based methods, however, commonly have to address two main questions: how many training examples are sufficient for the generic images, and how many subspaces are best for dictionary/regressor training?.

Unlike learning-based methods that use an external dataset, reconstruction-based methods directly employ the prior to constrain the image SR; they rely on details such as total variation (TV) regularization [5]. Aly and Dubois [25] presented a theoretical analysis of TV in the frequency domain. Farsiu *et al.* [26] proposed the bilateral TV prior (BTV) and formulated the super-resolution problem as an L_1-norm minimization problem using the BTV prior to enhance the robustness. Mitzel *et al.* [27] used TV regularization to reconstruct the high-resolution image with optic flow for the estimation of the motion field. Unger *et al.* [28] extended the work of Mitzel et al. and introduced the Huber norm, a replacement for the TV prior, to preserve the sharp edges. However, TV-based methods also have some well-recognized disadvantages: staircase artifacts may appear around the edge regions; some small-scale details tend to be lost; and it cannot perfectly preserve the sharp edges. To overcome these difficulties, many researchers have proposed improved methods, such as bilateral TV [26], beyond digital TV (BDTV) [29], higher degree total variation [30], and locally digital bilateral TV [31]. These works all rely on a similar idea of higher degree derivatives to improve the contour regularity and reduce artifacts. Moreover, Shan *et al.* [32] used the heavy-tailed distribution of the natural image gradients to constrain the unique solution. In this paper, we propose normalized bilateral TV by integrating bilateral filter and normalized sparsity measure, which is simple and robust and can better respect the sharpness and consistency of the edges and details.

B. Color-Guided Depth Super-Resolution

Color-guided depth SR can take advantage of the fact that the additional color information can provide complementary clues for depth SR. One way to implement this idea is to resort to context-specific filters. For example, Yang *et al.* [33] used bilateral filtering to produce a soft color segmentation, and then used the quadratic polynomial interpolation to refine the HR depth image estimation. Liu *et al.* [12] proposed a joint filter using the geodesic distances for depth SR, which introduces some artifacts in the regions with rich textures. Li *et al.* [34] proposed a Bayesian approach by combining the geometrical structures of the color image. Park *et al.* [14], [15] proposed a more complex framework using the local and nonlocal regularization terms. Despite using high level features, it tends to introduce jag artifacts around the edges. In sharp contrast, this paper involves using only low level information (intensity) to constrain the depth recovery model. Ferstl *et al.* [11] formulated a generalized second order TV as the regularization term, and guided the depth SR with an anisotropic kernel calculated from the corresponding HR color image. Nevertheless, the generated edge is blurred in large magnification case. Yang *et al.* [13] introduced a local autoregressive model by taking into account both the depth and color information to adaptively calculate the pixelwise weights. However, this method fails to maintain the sharp edges because the average of RGB channels should affect the structure awareness. In contrast, this paper uses a nonlocal regression strategy with an adaptive channel selection scheme, which should work better than a local autoregressive model.

C. Super-Resolution by 3D Reconstruction

Some works have pointed out the relationship between 3D reconstruction and SR. For example, Mudenagudi *et al.* [17] formulated the image SR problem using a MRF-MAP framework with the help of a calibrated 3D geometry, and solved it via graph cut optimization. However, their method has little contribution to depth map SR. Silva *et al.* [18] proposed a real-time SR method by improving the 3D reconstruction pipeline, wherein the spatial-resolution of depth is increased by exploiting the coincidence of color and depth edges. However, their framework ignores color image SR. Bhavsar and Rajagopalan [19] integrated the color image SR and HR depth recovery in a unified framework to correct and improve each other's accuracy. The graph cut and iterative conditional model are used to estimate RGB and depth images in each iteration. Recently, Lee and Lee [20] also proposed to simultaneously super-resolve RGB and depth images; they used the Huber norm as a regularization term to force the solution to be unique. Unlike these previous studies [19], [20], we define a more elegant energy function with different priors to intrinsically guide the correspondence construction and to replace simultaneous HR RGB-D image estimation with sequential SR.

III. Preprocessing of Multi-Observed RGB-D Images

Given the reference RGB-D image $\mathbf{I}^0$, we begin with an assumption that the images captured from multiple viewpoints are roughly located on the same plane. Therefore, for the LR images (denoted by $\mathbf{I}^{\upsilon}$ $(\upsilon = 0, \cdots, \nu)$) and the HR result $\mathbf{I}$,

TABLE I

LIST OF THE KEY MATHEMATICAL SYMBOLS USED IN THIS PAPER

$\imath$ ($\imath'$), $\jmath$ ($\jmath'$)	The pixel coordinates in RGB image and Depth image
κ, ι, o	The vertex coordinates in 3D space
υ, β, N	The index of viewpoint, the magnification factor, and the number of pixel in RGB image or Depth image
$p(\imath, \jmath)$	The vertex location of 3D space back-projected from the pixel coordinate ($\imath$, $\jmath$) in RGB image
$\mathbf{I}, \mathbf{D} \in \mathbb{R}^{\beta^2 N\times 1}$, and $\mathbf{I}^\upsilon, \mathbf{D}^\upsilon \in \mathbb{R}^{N\times 1}$	The estimated HR RGB and Depth images, and the LR RGB and Depth images under viewpoint υ
$\mathbf{I}_i$, $\mathbf{D}_i$	The intensity value and depth value at pixel i in RGB and Depth images
$\mathbf{A}^\upsilon$, $\mathbf{R}^\upsilon$, $\mathbf{t}^\upsilon$	The internal parameter, rotation, and translation matrices of the camera under viewpoint υ
$\mathbf{A}_H^0$	The internal parameter matrix corresponding to HR camera under viewpoint 0
$\mathbf{M}^\upsilon$,$\mathbf{B}$	The mapping matrix transforming HR image $\mathbf{I}$ to LR image $\mathbf{I}^\upsilon$, and the matrix form of the blur kernel
$\hat{\mathbf{D}}$, $\mathbf{C}$	The up-sampling depth image of $\mathbf{D}^0$, the selected channel with the biggest variance in three monochrome channels
$\nabla_{x1,m}$ ($\nabla_{y1,m}$), $\nabla_{x2,m}$ ($\nabla_{y2,m}$)	The forward and backward difference operators with the corresponding m-ring adjacent pixel in the x (y) direction
$\mathbf{H}_{x1,1}$	The matrix form of the discrete difference operator $\nabla_{x1,1}$
$\Psi()$, $\Phi(\mathbf{I}, m)$	The spatial weight function, the convolution function between $\mathbf{I}$ and m-ring difference operator
w_{ij}, $w_{ij}^{\hat{\mathbf{D}}}$, $w_{ij}^{\mathbf{I}}$	The regression coefficient between pixel i and pixel j, the depth similarity weight and the color similarity weight
$\mathbf{F}_i$, q	The bilateral filter kernel of the extracted patch centered at pixel i, the width of the extracted patch
$\mathbf{L}$, $\mathbf{J}$, $\mathbf{S}$, $\mathbf{Z}$	The simplified notations of large size matrices
$\mathbf{W}$, $\mathbf{S}^\star$	The regression matrix in Depth image SR, the convex conjugate of operator $\mathbf{S}$
$\tilde{\mathbf{I}}$, $\tilde{\mathbf{Z}}$, $\bar{\mathbf{I}}$	The intermediate result in the First-order primal-dual algorithm

a mapping can be established based on the imaging principle when the calibrated camera parameters are given.

To make our mathematical formulations clear, Table I summarizes the key symbols used in the following sections. Here, normal-case letters denote scalars, and bold upper-case letters denote matrices.

A. Camera Parameter Estimation

Before conducting RGB-D image super-resolution, we need to use the camera localization algorithm to estimate camera parameters. We use the multi-observed RGB-D image sequence to provide enough feature information. Then, we resort to feature-based MonoSLAM for camera localization [35]; this is a real-time algorithm used to recover the 3D trajectory of a monocular camera. Based on the estimated camera parameters, we build a mapping matrix under with the same viewpoint or with shifting viewpoints, which are detailed in the following sections.

B. LR-HR Image Mapping Under the Same Viewpoint

The well-known pin-hole model uses the internal parameter $\mathbf{A} \in \mathbb{R}^{3\times 3}$ (including the principal point c and focal length f), rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times 3}$ and translation matrix $\mathbf{t} \in \mathbb{R}^{3\times 1}$ of the camera to describe the imaging process. Under the reference viewpoint 0, the corresponding HR RGB-D images can be captured with the same extrinsic parameters $\mathbf{R}^0$ and $\mathbf{t}^0$ and a different internal matrix $\mathbf{A}_H^0$. As shown in Fig. 2, for a given magnification factor β, c_H and f_H of the corresponding HR camera are, respectively defined as $c_H^0 = \beta c^0$ and $f_H^0 = \beta f^0$. Please refer to a previous study [34] for more details. The mapping matrix $\mathbf{M}^0 \in \mathbb{R}^{N\times\beta^2 N}$ can be determined by the internal parameters $\mathbf{A}^0$ and $\mathbf{A}_H^0$. Therefore, matrix $\mathbf{M}^0$ should satisfy $\mathbf{M}^0\mathbf{I} = \mathbf{I}^0$, which is a down-sampling operator.

C. Image Mapping Between Shifting Viewpoints

Given the camera parameters, the mapping matrix between shifting viewpoints is more complex. Here, the extrinsic and internal parameters under other viewpoints are different.

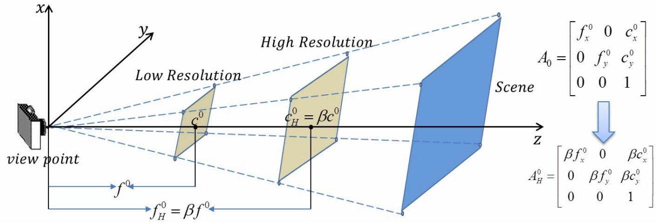

x
y
z
High Resolution
Low Resolution
Scene
view point
c^0
$c_H^0=\beta c^0$
f^0
$f_H^0=\beta f^0$
$A_0 = \begin{bmatrix} f_x^0 & 0 & c_x^0 \\ 0 & f_y^0 & c_y^0 \\ 0 & 0 & 1 \end{bmatrix}$
$A_H^0 = \begin{bmatrix} \beta f_x^0 & 0 & \beta c_x^0 \\ 0 & \beta f_y^0 & \beta c_y^0 \\ 0 & 0 & 1 \end{bmatrix}$

Fig. 2. Illustration of the pin-hole imaging model. For the same viewpoint, the relationship of intrinsic parameters between HR imaging and LR imaging directly depends on the magnification factor.

For a point $\mathbf{I}(\imath, \jmath)$ of HR image obtained under the reference viewpoint 0, the coordinate of $\mathbf{I}(\imath, \jmath)$ in 3D space is $(\imath, \jmath, 1)^T$ [11]. Given its associated inverse depth value $d_{\imath\jmath}$, we can backproject it into 3D space via

$$p(\imath, \jmath) = (\mathbf{R}^0)^{-1}(\frac{(\mathbf{A}_H^0)^{-1}\mathbf{I}(\imath, \jmath)}{d_{\imath\jmath}} - \mathbf{t}^0) = \frac{1}{d_{\imath\jmath}}(\mathbf{A}_H^0)^{-1}\mathbf{I}(\imath, \jmath), \tag{1}$$

where $\mathbf{R}^0$ is an identity matrix, $\mathbf{t}^0$ is a zero vector for the viewpoint 0. And $p(\imath, \jmath)$'s projection to the adjacent shifting viewpoint (with $\mathbf{A}^\upsilon$, $\mathbf{R}^\upsilon$ and $\mathbf{t}^\upsilon$) can be formulated as

$$\mathbf{I}^\upsilon(\imath', \jmath') = \omega(\frac{1}{d_{\imath\jmath}}\mathbf{A}^\upsilon\mathbf{R}^\upsilon(\mathbf{A}_H^0)^{-1}\mathbf{I}(\imath, \jmath) + \mathbf{A}^\upsilon\mathbf{t}^\upsilon), \tag{2}$$

where $\omega((\kappa, \iota, o)^T) = (\kappa/o, \iota/o)^T$ is the dehomogenization function. Therefore, given the pixel position $(\imath, \jmath)^T$ in HR image, we can calculate its associated pixel coordinates $(\imath', \jmath')^T$ under the adjacent viewpoint υ. The mapping matrix $\mathbf{M}^\upsilon \in \mathbb{R}^{N\times\beta^2 N}$ thus can be defined by Eq. 2, which transforms the image $\mathbf{I}$ to its adjacent viewpoint image $\mathbf{I}^\upsilon$. This process helps avoid the explicit registration among multi-view images by taking into account the depth image of 3D scene.

IV. SR OPTIMIZATION FRAMEWORK FOR RGB-D IMAGES

This section describes our optimization framework used for multi-observed RGB-D image SR. Our framework uses the geometrical-structure correlation of depth-color pair to improve the robustness and accuracy of SR processing. The technical elements are detailed in the following subsections.

A. Objective Function Definition

To leverage the correlation between depth-color pair, we integrate RGB image SR and depth recovery into a unified framework that guarantees the accuracy and robustness of both. The underlying motivation is that depth information can improve the correspondence accuracy between LR observations and HR results, and the color information can boost the depth recovery accuracy for high-magnification cases. The objective optimization function is formulated as

$$\underset{\mathbf{I},\mathbf{D}}{\operatorname{argmin}}\, E_d(\mathbf{I}) + E_d(\mathbf{D}) + \lambda_1 E_{NB}(\mathbf{I}) + \lambda_2 E_{NL}(\mathbf{D}). \tag{3}$$

Here, $\mathbf{I}$ and $\mathbf{D}$ denote the estimated HR color and depth images, respectively. $E_d(\mathbf{I})$ and $E_d(\mathbf{D})$ are fidelity terms that make the super-resolved RGB-D image consistent with the multi-view LR observations. $E_{NB}(\mathbf{I})$ is the normalized bilateral total variation regularization term defined over the super-resolved RGB image, and $E_{NL}(\mathbf{D})$ is the nonlocal regression term defined over the recovered depth image. λ_1 and λ_2 are the weight parameters that are used to balance the fidelity term and the regularization term. The fidelity terms are formulated as

$$E_d(\mathbf{I}) = \sum_{\upsilon=0}^{\upsilon} \|\mathbf{M}^{\upsilon}\mathbf{B}\mathbf{I} - \mathbf{I}^{\upsilon}\|_2^2, \tag{4}$$

$$E_d(\mathbf{D}) = \|\mathbf{M}^0\mathbf{D} - \mathbf{D}^0\|_2^2, \tag{5}$$

where $\mathbf{I}^{\upsilon}$ represents the captured RGB image at the υ-th viewpoint. $\mathbf{M}^{\upsilon}$ is the mapping matrix. $\mathbf{I}^0$ and $\mathbf{D}^0$ are the reference RGB-D images obtained from the 0-th viewpoint. $\mathbf{B}$ is the convolution matrix of the blur kernel.

B. Regularization Prior Formulation for RGB Image SR

The regularization term $E_{NB}(\mathbf{I})$ has to preserve the sharp edges well and remove artifacts. The widely used regularization term is TV regularization (formulated as $\|\nabla\mathbf{I}\|_1$) [5], whose L_1 norm can be regarded as a sparsity metric on the gradient domain. Although the TV model has achieved great success, there is still room for further improvement in maintaining the clarity of important geometric structures (edges and details). Here we define a new prior, which we call the normalized bilateral TV (NBTV). This prior can effectively maintain the sharpness and consistency of true edges and details. NBTV intrinsically couples the bilateral filter with a normalized sparsity metric, which is formulated as

$$E_{NB}(\mathbf{I}) = \frac{\sum_{m=1}^{\gamma}\|\Psi(m)\Phi(\mathbf{I},m)\|_1}{\sum_{m=1}^{\gamma}\|\Psi(m)\Phi(\mathbf{I},m)\|_2}, m = 1,\cdots,\gamma, \tag{6}$$

where $\Psi(m) = \mu^m$ is the spatial weight ($\mu = 0.7$). $\Phi(\mathbf{I}, m)$ generalizes the BDTV by exploring a larger neighborhood during reconstruction. It has four generalized discrete filters, which are formulated as

$$\nabla_{x1,m} = [\overbrace{0,\cdots,0}^{m}, -1, \overbrace{0,\cdots,0}^{m-1}, 1],\ \nabla_{y1,m} = \nabla_{x1,m}^T,$$
$$\nabla_{x2,m} = [1, \overbrace{0,\cdots,0}^{m-1}, -1, \overbrace{0,\cdots,0}^{m}],\ \nabla_{y2,m} = \nabla_{x2,m}^T. \tag{7}$$

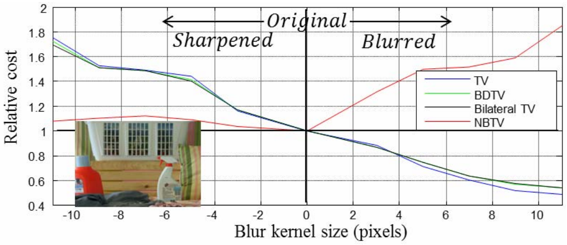


Fig. 3. A comparison of priors. The y-axis represents the relative cost, which measures the possibility of generating a test image using the prior. The relative cost is defined as the ratio of the cost (L_1 norm of the prior) of the test image to that of the original image. The lower the relative cost is, the easier it is to generate the testing image based on the prior. The x-axis represents the size of the blur kernel. A negative size corresponds to a sharpened image, and a positive size corresponds to a smoothed image. Existing TV priors tend to generate blurred results. In contrast, the NBTV prior (red curve) performs better; its cost for the original image is lowest. This demonstrates that the NBTV prior can be consistent with the original image.

$\Phi(\mathbf{I}, m) = \left[\nabla_{x1,m}, \nabla_{y1,m}, \nabla_{x2,m}, \nabla_{y2,m}\right]^T \otimes \mathbf{I}$ is defined as the convolution of $\mathbf{I}$ with Eq. 7 as kernel. It reduces to the original BDTV model when $m = 1$. The spatial weight function $\Psi(m)$ weakens the influence of the distant neighborhoods and can be regarded as the spatial kernel. $\Phi(\mathbf{I}, m)$ increases the penalty for the pixels that have high inconsistency with respect to the true structures, and it is used as the range kernel. $\sum_{m=1}^{\gamma}\|\Psi(m)\Phi(\mathbf{I},m)\|_2$ is used to normalize the bilateral TV term $\sum_{m=1}^{\gamma}\|\Psi(m)\Phi(\mathbf{I},m)\|_1$. Therefore, the proposed regularization term is named Normalized Bilateral TV. It should be noted that we empirically set $\gamma = 2$. Fig. 3 presents a comparison between NBTV and existing TV priors, which indicates that our prior has better performance in maintaining sharp edges.

The newly proposed Normalized Bilateral TV (NBTV) has the form $\frac{L_1}{L_2}$, and it can be regarded as a normalized version of BTV function (L_1). The BTV prior typically penalizes the high frequency bands in image reconstruction. Thus, the BTV regularization term generally tends to produce a blurry result for image super-resolution because the blur operator attenuates the high-frequency bands and gives rise to a lower BTV norm. This situation has been studied [45], and Fig. 3 also illustrates such a case. Nevertheless, benefiting from the fact that NBTV is equivalent to the BTV norm rescaled by their total energy (L_2 norm), the blur can be effectively controlled. Given a blurry result, although blur can decrease both the BTV norm and its total energy (L_2 norm), it is vital that the latter is reduced faster. Thus, the value of $\frac{L_1}{L_2}$ function will be increased for a blurry result, which had been demonstrated in Fig. 3. For the noise and sharpened cases, as shown in Fig. 3, we can see that NBTV increases the L_1/L_2 value. Therefore, the prior of NBTV can be consistent with the original image for the three cases, and it generates a result with better visual quality.

C. Nonlocal Regression Function for Depth Image SR

First, the depth image generally has a high amount of self-similarity (e.g., flat regions), which encourages us to further introduce the regression model to improve the performance of depth image recovery. Nonetheless, the local self-similarity may not provide enough information for faithful depth reconstruction. Ringing and jag artifacts often appear around the edges of the recovered depth image. Note that the depth map mainly contains smooth regions separated by the edges and structures corresponding to the depth discontinuities. Numerous nonlocal self-similarities can also be found in the depth image, which can enhance the accuracy of the constructed depth within a larger neighborhood. Considering the computationally demands of L_1 regularization term, we can formulate the depth image SR as a problem of squares regression. Thus, we further introduce a nonlocal regression (NLR) regularization term into the optimization framework.

To further improve the quality of the recovered depth, we use the cues in HR RGB image to guide the whole reconstruction process. Then, the color-guided NLR model is incorporated into the optimization function. We use a pixel-wise strategy to define newly proposed NLR model, which can achieve better reconstruction quality than the patch-wise model. Please refer to a previous study [13] for more details. Therefore, our color-guided NLR model can be formulated as

$$E_{NL}(\mathbf{D}) = \|\mathbf{D} - \mathbf{W}\mathbf{D}\|_2^2 = \|(\mathbf{1} - \mathbf{W})\mathbf{D}\|_2^2. \tag{8}$$

Here, $\mathbf{1}$ is the identity matrix, and $\mathbf{W} \in \mathbb{R}^{\beta^2 N \times \beta^2 N}$ is the regression coefficient matrix. The regression coefficient w_{ij} is defined as

$$w_{ij} = \frac{1}{S_i} w_{ij}^{\hat{\mathbf{D}}} w_{ij}^{\mathbf{I}}, \tag{9}$$

where S_i is the normalization factor. $\hat{\mathbf{D}}$ is the up-sampled depth image via Bicubic interpolation of $\mathbf{D}^0$. The color term $w_{ij}^{\mathbf{I}}$ is introduced to take advantage of the correlations embedded in the depth-color pair. Due to the co-occurrence of edges in the RGB-D pair, the color term can help avoid depth discontinuities. The depth term $w_{ij}^{\hat{\mathbf{D}}}$ is incorporated to prevent the NLR model from recovering incorrect depths due to depth-color inconsistency: different-colored pixels could have the same depth, and pixels with similar color could be at different depth levels.

The depth term $w_{ij}^{\hat{\mathbf{D}}}$ can be computed by using the Gaussian function on the initial depth $\hat{\mathbf{D}}$:

$$w_{ij}^{\hat{\mathbf{D}}} = exp(-\frac{(\hat{D}_i - \hat{D}_j)^2}{2\sigma_1^2}), \tag{10}$$

where $\hat{D}_i$ is a scalar, denoting the depth value at pixel i. The standard deviation σ_1 is used to control the similarity of depth. The definition of the color term $w_{ij}^{\mathbf{I}}$ also uses the Gaussian function to calculate the weight. However, we calculate it over a pixel window rather than a single pixel. Based on the pixel window, we introduce the bilateral kernel to weight the distances between local patches. Using the structural awareness of the bilateral kernel, we can use the structure information in the local patch to produce more candidate patches and form a more well-determined nonlocal regression model.

The bilateral kernel can be calculated over three RGB channels, as described previously [13]. As shown on the right side of Fig. 4, this approach generally cannot produce satisfactory results with sharp edges. The blurred edges appear in the results of the adaptive auto-regressive model method (AAR14) [13]. That is because the average operator on three channels may impair structural awareness. We propose selecting the channel so that edges preserved in one channel can be used to recover sharp edges in the depth image. Our strategy is based on the assumption that the larger-variance channel has better structure-awareness (see the left figure of Fig. 4). Therefore, the color term is formulated as

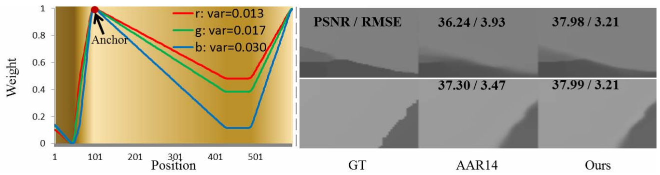


Fig. 4. Illustration of channel selection. Left: the bilateral filter kernels defined over each RGB channel. “var” denotes the variance of the curves of the single channels, “Position” indexes the pixel location, and “Weight” means the similarity between the adjacent pixel and the anchor pixel (red point). Channel “b” achieves the highest variance, and its weight curve (indicated by blue color) decreases rapidly for the pixels on the edge, which shows that channel “b” is more structure-aware. Thus, the larger-variance channel has better structural awareness. Right: The depth recovery effects for two sets of data. The columns from left to right are as follows: the ground truth (GT), the results of an adaptive auto-regressive method (AAR14), and our results. Compared to the average method used by AAR14 [13], our channel selection achieves better recovery.

$$w_{ij}^{\mathbf{I}} = exp(-\frac{\|\mathbf{F}_i \odot (\mathbf{C}_i - \mathbf{C}_j)\|_2^2}{2\sigma_2^2}), \mathbf{C} = \underset{\mathbf{C} \in r,g,b}{max}(var_{\mathbf{C}}). \tag{11}$$

Here, “$\odot$” represents the element-wise Hadamard product, $\mathbf{C}_i$ denotes the extracted patch centred at pixel i according to the selected channel, “var” is the variance of the patch for each channel, and σ_2 is the standard deviation of the structures, which controls the decay rate. The operator $\mathbf{F}_i$ is a bilateral filter kernel of the extracted patch $\mathbf{C}_i$, which is defined as

$$\mathbf{F}_{ij} = exp(-\frac{(i-j)^2}{2\sigma_3^2})exp(-\frac{(c_i - c_j)^2}{2\sigma_4^2}), j \in patch, \tag{12}$$

where c_i denotes the intensity at pixel i, σ_3 and σ_4 are the spatial and range parameters of the bilateral filter kernel to control the edge response. The benefit of introducing a local structure kernel $\mathbf{F}_i \in \mathbb{R}^{q \times q}$ is that it produces more candidate patches, which makes it possible to form a more well-determined NLR model. q is the width of the extracted patch. For explanations, please refer to Fig. 5.

For any pixel j in a sufficiently large search window around pixel i, we calculate its similarity w_{ij} to i. A pixel j is chosen as the similar pixel to i if $w_{ij} \geq 0.98$ or w_{ij} is within the first 15 pixels. Thus, the matrix $\mathbf{W}$ can be defined as

$$\mathbf{W}_{ij} = \begin{cases} w_{ij} & \text{if pixel } j \text{ is similar to pixel } i \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

D. Implementation of Optimization Algorithm

Note that the initialization of matrices $\mathbf{M}^v$ and $\mathbf{W}$ mainly depends on the unknown variables $\mathbf{D}$ and $\mathbf{I}$, respectively. Because Eq. 3 is non-convex, we can use the alternating

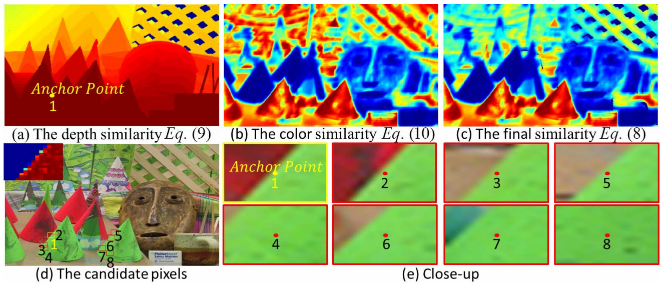

Anchor Point
1
(a) The depth similarity Eq. (9)
(b) The color similarity Eq. (10)
(c) The final similarity Eq. (8)
1
2
3
4
5
6
7
8
(d) The candidate pixels
Anchor Point
1
2
3
5
4
6
7
8
(e) Close-up

Fig. 5. Illustration of color-guided NLR model. (a) The depth similarity distribution for anchor point 1 calculated by Eq. 10. (b) The color similarity calculated by Eq. 11. (c) The final similarity calculated by Eq. 9. (d) The candidate pixels (2-8) selected by NLR model. The bilateral kernel of the patch 1 is shown at the corner. (e) Close-up effects. Pixel 2 is selected for only the patch-based method. The shape-based local regression model [13] can produce candidate pixels 2, 3, and 4. In contrast, the proposed shaped-based NLR model can choose the nonlocal pixels 5, 6, 7, and 8 while generating the candidate pixels 2, 3, and 4. Thus, our method produces more candidate pixels than other competing methods, and it shows better performance for depth SR.

minimization (AM) algorithm, which fixes one set of variables while updating another set of variables. If the objective function is slightly modified, it can also be optimized via Proximal Alternating Linearized Minimization [39], and the similar results would be obtained. Here, we encode the RGB and Depth images as a column vector, and we denote the RGB and Depth images as $\mathbf{I}$ and $\mathbf{D}$, respectively.

1) Updating HR RGB Image Estimate: We can fix $\mathbf{D}$ with $\hat{\mathbf{D}}$ in Eq. 3, and iteratively estimate the HR RGB image $\mathbf{I}$ by solving the following $\mathbf{I}$-problem optimization function:

$$\begin{aligned}&\underset{\mathbf{I}}{\operatorname{argmin}}\, E_d(\mathbf{I}) + \lambda_1 E_{NB}(\mathbf{I})\\ &\quad = \sum_{v=0}^{\nu} \|\mathbf{M}^v\mathbf{B}\mathbf{I} - \mathbf{I}^v\|_2^2 + \lambda_1 \frac{\sum_{m=1}^{\gamma}\|\Psi(m)\Phi(\mathbf{I},m)\|_1}{\sum_{m=1}^{\gamma}\|\Psi(m)\Phi(\mathbf{I},m)\|_2}, \qquad (14)\end{aligned}$$

where $B \in \mathbb{R}^{\beta^2N\times\beta^2N}$ is the matrix form of the blur kernel. Eq. 14 is also non-convex due to the normalized term $\frac{\sum_{m=1}^{\gamma}\|\Psi(m)\Phi(\mathbf{I},m)\|_1}{\sum_{m=1}^{\gamma}\|\Psi(m)\Phi(\mathbf{I},m)\|_2}$. When $\sum_{m=1}^{\gamma}\|\Psi(m)\Phi(\mathbf{I},m)\|_2$ is fixed, the sub-problem becomes a convex l_1-regularized problem:

$$\underset{\mathbf{I}}{\operatorname{argmin}} \sum_{v=0}^{\nu}\|\mathbf{M}^v\mathbf{B}\mathbf{I} - \mathbf{I}^v\|_2^2 + \lambda\sum_{m=1}^{\gamma}\|\Psi(m)\Phi(\mathbf{I},m)\|_1. \qquad (15)$$

When $\hat{\mathbf{D}} \in \mathbb{R}^{\beta^2N\times 1}$ is fixed, the matrix $\mathbf{M}^v \in \mathbb{R}^{N\times\beta^2N}$ can be initialized, and $\mathbf{I}^v \in \mathbb{R}^{N\times 1}$. We use the first-order primal-dual algorithm [38] to solve the general inverse problem, and we re-weight the parameter $\lambda = \lambda_1/\sum_{m=1}^{\gamma}\|\Psi(m)\Phi(\mathbf{I},m)\|_2$. Eq. 15 can be rewritten as

$$\begin{aligned}&\underset{\mathbf{I}}{\operatorname{argmin}} \left\| \begin{bmatrix}\mathbf{M}^0\mathbf{B}\\ \mathbf{M}^1\mathbf{B}\\ \vdots\\ \mathbf{M}^\nu\mathbf{B}\end{bmatrix}\mathbf{I} - \begin{bmatrix}\mathbf{I}^0\\ \mathbf{I}^1\\ \vdots\\ \mathbf{I}^\nu\end{bmatrix}\right\|_2^2 + \left\|\begin{bmatrix}\lambda\mu^1\mathbf{H}_{x1,1}\\ \lambda\mu^1\mathbf{H}_{y1,1}\\ \vdots\\ \lambda\mu^\gamma\mathbf{H}_{y2,\gamma}\end{bmatrix}\mathbf{I}\right\|_1\\ &\quad = \underset{\mathbf{I}}{\operatorname{argmin}}\|\mathbf{L}\mathbf{I} - \mathbf{J}\|_2^2 + \|\mathbf{S}\mathbf{I}\|_1, \qquad (16)\end{aligned}$$

where $\mathbf{L} \in \mathbb{R}^{(\nu+1)N\times\beta^2N}$, $\mathbf{J} \in \mathbb{R}^{(\nu+1)N\times 1}$, $\mathbf{H}_{x1,1} \in \mathbb{R}^{\beta^2N\times\beta^2N}$ is the convolution matrix of the discrete filter $\nabla_{x1,1}$, and analogous cases for other convolution matrices follow similar pattern (e.g., $\mathbf{H}_{y1,1}$, $\mathbf{H}_{y2,m}$), $\mathbf{S} \in \mathbb{R}^{(\nu+1)\beta^2N\times\beta^2N}$. Thus, Eq. 16 is rewritten as

$$\mathbf{I}_{opt} = \underset{\mathbf{I}}{\operatorname{argmin}}\, Q(\mathbf{I}) + P(\mathbf{S}\mathbf{I}), \qquad (17)$$

where $P(\mathbf{Z}) = \|\mathbf{S}\mathbf{I}\|_1$, $\mathbf{Z} = \mathbf{S}\mathbf{I}$, $Q(\mathbf{I}) = \|\mathbf{L}\mathbf{I} - \mathbf{J}\|_2^2$.

Algorithm 1 First-Order Primal-Dual Algorithm

```
input : $\tau = 1e^{-4}$, $\alpha = 0.9/(\tau\theta^2)^1$, $\mathbf{I}^0, \mathbf{Z}^0$, and $\bar{\mathbf{I}}^0 = \mathbf{0}$.
Iterations ($iter > 0$): Update $\mathbf{I}^{iter}$, $\mathbf{Z}^{iter}$, $\bar{\mathbf{I}}^{iter}$ as follows:
    $\begin{cases}\mathbf{Z}^{iter+1} & = (\mathbf{1} + \alpha\partial Q^\star)^{-1}(\mathbf{Z}^{iter} + \alpha\mathbf{S}\bar{\mathbf{I}}^{iter})\\ & = (\mathbf{1} + \alpha\partial Q^\star)^{-1}(\tilde{\mathbf{Z}})\\ \mathbf{I}^{iter+1} & = (\mathbf{1} + \tau\partial P)^{-1}(\mathbf{I}^{iter} - \tau\mathbf{S}^\star\mathbf{Z}^{iter+1})\\ & = (\mathbf{1} + \tau\partial P)^{-1}(\tilde{\mathbf{I}})\\ \bar{\mathbf{I}}^{iter+1} & = 2\mathbf{I}^{iter+1} - \mathbf{I}^{iter}\end{cases}$    (20)
```

$\mathbf{Z}_{opt} = \operatorname{argmax}_{\mathbf{Z}} -(Q^\star(-\mathbf{S}^\star\mathbf{Z}) + P^\star(\mathbf{Z}))$ is the dual problem of Eq. 17, where $\star$ denotes the convex conjugate, $\mathbf{Z} \in \mathbb{R}^{(\nu+1)\beta^2N\times 1}$. Algorithm 1 is used to solve the above problem, where θ is the square root of the largest eigenvalue of the symmetric matrix $\mathbf{S}^T\mathbf{S}$. We can use power iteration to obtain its value. Please refer to a previous study [38] for more details. The resolvent operator of Q is defined as

$$\mathbf{I} = (\mathbf{1} + \tau\partial P)^{-1}(\tilde{\mathbf{I}}) = \underset{\mathbf{I}}{\operatorname{argmin}} \frac{\|\mathbf{I} - \tilde{\mathbf{I}}\|_2^2}{2\tau} + Q(\mathbf{I}), \qquad (18)$$

where $\mathbf{I}$ is the estimated HR RGB result, and $\tilde{\mathbf{I}}$ is formulated in Eq. 20. By calculating the derivative of Eq. 18 with respect to $\mathbf{I}$ and setting the result to be 0, we obtain $(2\tau\mathbf{L}^T\mathbf{L} + \mathbf{1})\mathbf{I} = 2\tau\mathbf{L}^T\mathbf{J} + \tilde{\mathbf{I}}$. This is a squared matrix system, which can be solved by the Jacobi iterative method. Considering the computational cost, instead of using n-step iteration, we use 1-step iteration to approximate $\mathbf{I}^n$ in Algorithm 1, which has little effect on the final quality of image $\mathbf{I}$.

Similarly, the solution of $\mathbf{Z}$ is given by

$$\mathbf{Z} = (\mathbf{1} + \alpha\partial Q^\star)^{-1}(\tilde{\mathbf{Z}}) \Longleftrightarrow \mathbf{Z}_i = \frac{\tilde{\mathbf{Z}}_i}{max(1, |\tilde{\mathbf{Z}}_i|)}, \qquad (19)$$

where $\mathbf{Z}_i$ and $\tilde{\mathbf{Z}}_i$ denote the values of pixel i in $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$, respectively. $\tilde{\mathbf{Z}}$ is formulated in Eq. 20.

2) Updating the HR Depth Image Estimate: Next, we fix the updated HR RGB image $\mathbf{I}$ to further update the depth image $\mathbf{D}$ by solving the following energy function:

$$\begin{aligned}&\underset{\mathbf{D}}{\operatorname{argmin}}\, E_d(\mathbf{D}) + \lambda_2 E_{NL}(\mathbf{D})\\ &\quad = \underset{\mathbf{D}}{\operatorname{argmin}}\|\mathbf{M}^0\mathbf{D} - \mathbf{D}^0\|_2^2 + \lambda_2\|(\mathbf{1} - \mathbf{W})\mathbf{D}\|_2^2. \qquad (21)\end{aligned}$$

Here, Eq. 21 is convex. By calculating the derivative of Eq. 21 with respect to $\mathbf{D}$ and setting the result to be 0, we can compute the global minimum by solving the following squared matrix system:

$$((\mathbf{M}^0)^T\mathbf{M}^0 + \lambda_2(\mathbf{1} - \mathbf{W})^T(\mathbf{1} - \mathbf{W}))\mathbf{D} = (\mathbf{M}^0)^T\mathbf{D}^0, \qquad (22)$$

Algorithm 2 RGB-D Image Super-Resolution

```
input : Multi-observation RGB images $\mathbf{I}^v$, Depth image
        $\mathbf{D}^0$, $\lambda_1$, $\lambda_2$, $\sigma_1$, $\sigma_2$, $\sigma_3$, $\sigma_4$, $\gamma$.
output: HR RGB Image $\mathbf{I}$ and Depth Image $\mathbf{D}$.

for $OuterIter$ = 1 to 18 do
  Step I. Depth guided RGB Image Super-resolution
    1: $\hat{\mathbf{D}} = \mathbf{D}$ and back project $\mathbf{I}^0$ to 3D space;
    2: Construct the mapping matrix $\mathbf{M}^0$ and $\mathbf{M}^v$ (Eq. 2);
    3: Generate matrices $\mathbf{L}$, $\mathbf{S}$ in Eq. 16;
    4: for $iter$ = 1 to 100 do
          5: Estimate $\mathbf{I}^{iter}$ using algorithm 1;
          6: $\lambda = \lambda_1 / \sum_{m=1}^{\gamma} \|\Psi(m)\Phi(\mathbf{I}^{iter}, m)\|_2$;
       end
  Step II. Color guided Depth Super-resolution
    7: Calculate the similarity by using $\mathbf{I}^{iter}$ (Eq. 9);
    8: Construct the regression matrix $\mathbf{W}$ (Eq. 13);
    9: Estimate the super-resolved $\mathbf{D}$ using PCG method;
end
```

where the coefficient matrix $((\mathbf{M}^0)^T\mathbf{M}^0 + \lambda_2(\mathbf{1}-\mathbf{W})^T(\mathbf{1}-\mathbf{W}))$ is symmetric, positive definite and sparse. Eq. 22 can be efficiently solved by the Preconditioned Conjugate Gradients method (PCG) within a few iterations. Here, we use the Matlab function **pcg** to solve Eq. 22. Algorithm 2 documents the details of our RGB-D image SR.

E. Analysis of Computational Complexity

Here, we analyse the computational complexity of three ingredients (Preprocessing, RGB upsampling (*Pre.*), and Depth super-resolution). First, in the *Pre.* part, the mapping matrix $\mathbf{M}^0$ does not need any computational burden because it is a down-sampling operator under the same viewpoint. For the shifting viewpoints, 120 floating-point operations (FLO) are needed for the computing of $\mathbf{I}^v(\imath', \jmath')$. Thus, we need to perform $120\nu\beta^2 N$ FLOs to construct all the mapping matrices $\mathbf{M}^v$, where β denotes the magnification factor, ν and N are the numbers of viewpoints and pixels in the LR image, respectively. The computational complexity is $\mathcal{O}(N)$ in the *Pre.* step. In the RGB upsampling step, we first construct the matrix $\mathbf{L}$ by computing $\mathbf{M}^v\mathbf{B}$. Here, $\mathbf{M}$ and $\mathbf{B}$ are sparse matrices, and $\beta^2 N^2$ FLOs are needed for $\mathbf{M}^v\mathbf{B}$ computation. Thus, it is necessary to construct $\mathbf{L}$ to perform $\beta^2 N^2(\nu+1)$ FLO operations. Similarity, matrices $\mathbf{Z}$, $\tilde{\mathbf{Z}}$, and $\tilde{\mathbf{I}}$ must be computed to perform $2 * 8\beta^2 N$ FLOs, $2 * 8\beta^2 N + (\nu+1)\beta^2 N$ FLOs, and $2 * 8\beta^2 N + \beta^2 N$ FLOs, respectively. Here, the number 8 means eight discrete filters when $\gamma = 2$. Moreover, in the 1-step iteration of the Jacobi iterative method, we need to perform $(\nu+1)^2 N^2$ FLOs and $(\nu+1)N$ FLOs to compute matrix $\mathbf{L}^T\mathbf{L}$ and $\mathbf{L}^T\mathbf{J}$. Therefore, the computational complexity is $\mathcal{O}(N^2)$ in the RGB SR step. For depth super-resolution, $\beta^2 N^2$ FLOs, $15\beta^4 N^2$ FLOs, and N FLOs, respectively, are needed for computing $(\mathbf{M}^0)^T\mathbf{M}^0$, $(\mathbf{1}-\mathbf{W})^T(\mathbf{1}-\mathbf{W})$, and $(\mathbf{M}^0)^T\mathbf{D}^0$. Here, the number 15 is the number of candidate pixels in the nonlocal regression model. Therefore, the computational complexity is $\mathcal{O}(N^2)$ for constructing the squared matrix system. Table II documents the time statistics by running image on a computer with 24 GB RAM and an Intel i7-3770 3.4 GHz CPU. We list the time required for our method, including *Pre.*, RGB SR, and Depth SR. Time spent on RGB SR includes the costs of computing $\mathbf{L}$, $\mathbf{Z}$, and $\mathbf{I}$. Depth SR includes the time for constructing the squared matrix system ($Con.$), and for PCG method.

TABLE II

TIME PERFORMANCE (IN MINUTES) WITH 4X FOR DIFFERENT SIZE OF IMAGE

Size	Pre.	RGB SR			Depth SR		Sum
		L	**Z**	**I**	$Con.$	PCG	
160 × 240	0.12	0.57	0.04	0.52	0.41	1.54	3.20
450 × 370	0.31	0.83	0.10	2.20	2.44	3.96	9.84
640 × 480	0.47	1.03	0.23	4.60	4.26	8.96	19.55

V. EXPERIMENTS AND EVALUATIONS

We quantitatively evaluate and validate our algorithm on the Middlebury datasets [40], NYU datasets [41], and data reported by Yang *et al.* [13]. We use PSNR and the root mean squared error (RMSE) as ground truth based quality indicators, and use JPEG2000 compression (JP2K) [42], a blur metric (ReBlur) [43], and a blind/referenceless image spatial quality evaluator (BRISQUE) [44] as no-reference quality indicators. For JP2K and PSNR, higher values mean better quality, but lower values are better for ReBlur, BRISQUE, and RMSE.

Based on well-designed experiments, we compare our method with six state-of-the-art methods, including the bicubic method, Shan's method (Shan08) [32], Yang's method (Yang10) [7], the anchored neighborhood regression method (ANR13) [9], the super-resolution convolutional neural network method (SCRNN14) [24], and the jointly optimized regressors method (JOR15) [10]. These methods can provide only RGB image super-resolution; no depth images are involved. For the learning-based methods, because the released patch database is trained for 4× upsampling, we mainly present the comparison experiment with a magnification factor of 4. Specifically, for the experiments on color-guided depth super-resolution, we compare our method with four state-of-the-art methods, including joint geodesic filtering (JGF13) [12], total generalized variation (TGV13) [11], edge-weighted NLM-regularization (Edge14) [15], and the adaptive auto-regressive model method (AAR14) [13]. We present the results for 16× upsampling. In our experiments, the blur function B is used as the Gaussian kernel of size 11 with standard deviation 1.4.

This section is organized as follows. First, we detail the parameter selection in Section V-A. Section V-B shows RGB super-resolution evaluations based on comparisons, including 1) the comparison of RGB up-sampling; 2) the robustness evaluation over Gaussian noise case; 3) the visual quality evaluation of our method for high-magnification factor; 4) the performance analysis on a synthetic dataset; and 5) the analysis of the proposed prior (NBTV). We also evaluate the color-guided depth image super-resolution in Section V-C,

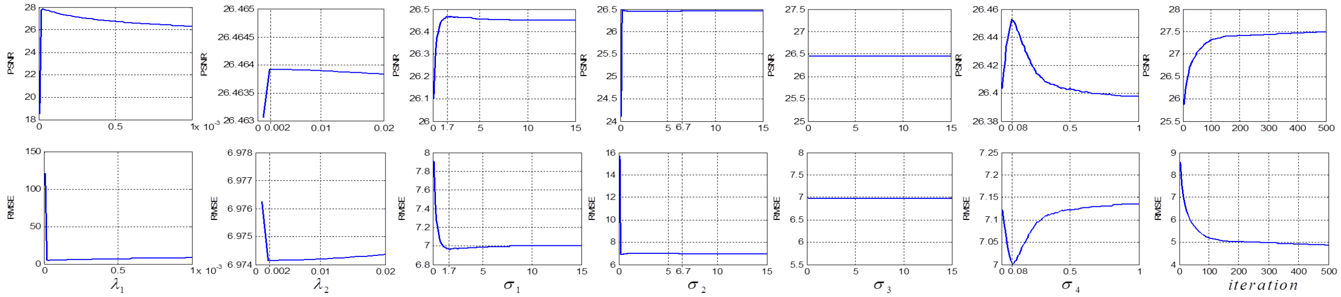


Fig. 6. The statistic curves of PSNR and RMSE with respect to the parameters λ_1, λ_2, σ_1, σ_2, σ_3, σ_4, and the number of iterations.

including: 1) the visual quality comparison of depth image super-resolution; 2) the comparison over the non-aligned case; 3) the robustness evaluation over ToF-like noise and Kinect-like noise cases.

A. Parameter Selection

To obtain the optimum parameters, we constructed a test dataset. First, we sample 50 image patches of size 20×20 from the *Middlebury* Dataset. We select only image patches with variance ranked within the first 10, and we remove patches with small variances, which tend to reduce the effectiveness of the parameter. For each patch, we evaluate the parameters using PSNR and RMSE for each test value. Then, for each test value, the average values of PSNR and RMSE for all test patches are used to construct the curves. To evaluate the effects of the involved parameters on the stability and accuracy, we first initialize these parameters as $\lambda_1 = 1e^{-5}$, $\lambda_2 = 1e^{-3}$, $\sigma_1 = 8$, $\sigma_2 = 9$, $\sigma_3 = 3$, $\sigma_4 = 0.1$. Then, we perform RGB-D SR $4X$ by varying a parameter while keeping other parameters being fixed. When one parameter reaches the optimum value, we update the initial value of the parameter, and then evaluate other parameters. Here, we briefly discuss how to determine these parameters.

λ_1 and λ_2: In RGB image SR, a very small λ_1 may indeed increase the instability of our method, which gives rise to lower PSNR value. As shown in Fig. 6, increasing the value of λ_1 yields a better result, and when $\lambda_1 \in [3e^{-5}, 1.2e^{-4}]$, our method achieves better values of PSNR and RMSE. Therefore, we empirically set $\lambda_1 = 3e^{-5}$. For depth image SR, in Eq. 22, $(\mathbf{M}^0)^T\mathbf{M}^0 + \lambda_2(\mathbf{1}-\mathbf{W})^T(\mathbf{1}-\mathbf{W})$ is equal to $(\mathbf{1}-\mathbf{W})^T(\mathbf{1}-\mathbf{W})$ for very large λ_2, and it is equal to $(\mathbf{M}^0)^T\mathbf{M}^0$ for very small λ_2, which will lead to a high condition number and increase the instability of our framework. We also find that $\lambda_2 \in [2e^{-3}, 1e^{-2}]$ can produce satisfactory results according to the curves in Fig. 6. Therefore, we set $\lambda_2 = 5e^{-3}$.

σ_1 and σ_2: In the depth term, σ_1 controls the importance of two neighboring different depth values. As shown in Fig. 6, PSNR has a peak at $\sigma_1 = 1.7$, and for the values larger than 1.7, there is a slight decrease. For the smaller σ_1, the depth weight plays a much more dominant role than the color term in Eq. 9. It may even make the color term completely ineffective, and it relies on only the depth term. We can see that the result reaches a steady state when $\sigma_1 = 10$. Therefore, we set $\sigma_1 = 10$. Being similar to σ_1, σ_2 in the color term controls the sharpness clue of the depth edge in the corresponding HR color information. As shown in Fig. 6, when $\sigma_2 \in [6.7, 15]$, a better PSNR can be achieved. Therefore, we set $\sigma_2 = 6.7$.

σ_3 and σ_4: These two parameters in Eq. 12 control the shape and size of the patch kernel, and they effect the selection of candidate nonlocal patches. In Fig. 6, the PSNR and RMSE curves are quite stable when σ_3 is varied. Therefore, we set $\sigma_3 = 3$. Parameter σ_4 controls the structure awareness and the supporting region of similarity. For a very small σ_4, the local structural kernel F_i tends to assign zero to most of the pixels. In extreme cases, the patch will degenerate to a single pixel. For a large σ_4, F_i will assign equal weights to pixels within the patch, and the structure kernel will degenerate to a mean kernel. The two cases above can easily lead to instability. In Fig. 6, the PSNR and RMSE curves have distinctive minimum around $\sigma_4 = 0.08$. Therefore, we set $\sigma_4 = 0.08$.

Number of Iterations: We analyse the relationship between the quantitative evaluation indicators and the number of iterations used in the first-order primal-dual algorithm. Fig. 6 presents the experimental curves. Note that RMSE drops rapidly for the first few iterations. As the number of iterations increases to 100, the curves reach a steady state. Therefore, we set the number of iterations to 100.

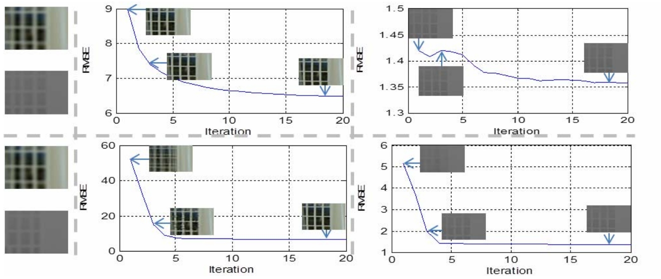


Fig. 7. Convergence analysis. Top row: the exactly aligned case. Bottom row: the non-aligned case. Left: Input RGB and Depth images. Middle: RMSE curves of RGB results for two cases. Right: RMSE curves of Depth SR results. Three representative super-resolved RGB and Depth results are presented for comparison and analysis, which are generated in 1, 3, 18 iterations.

Convergence Analysis: We analyse the convergence of our algorithm for two cases in Fig. 7. One case provides exact alignment between the RGB-D pairs (top row), and the other is non-aligned (bottom row). For the exactly aligned case, after 1 iteration, our method can generate an acceptable RGB and depth SR result. With increasing iterations, the RMSE

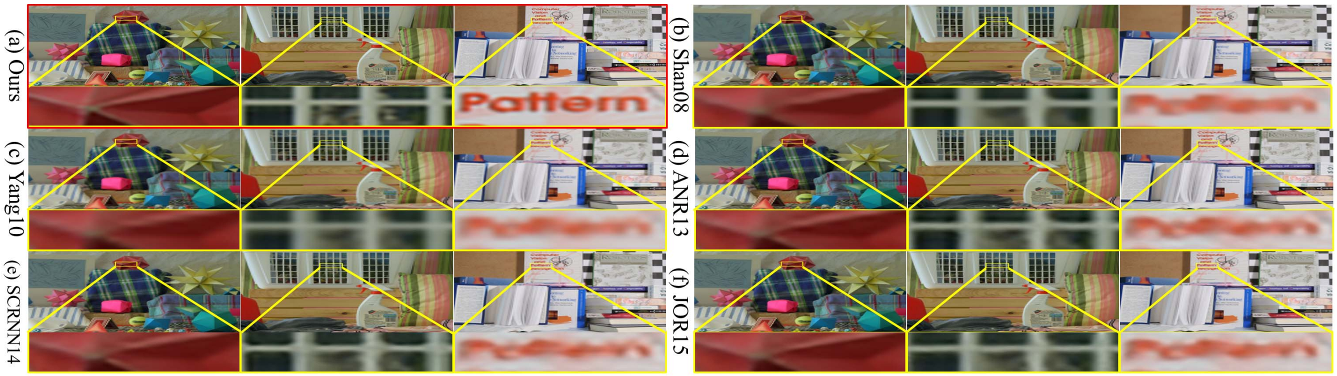

(a) Ours (b) Shan08 (c) Yang10 (d) ANR13 (e) SCRNN14 (f) JOR15

Fig. 8. Comparison of RGB image super-resolution results (4×) for the *Moebius*, *Laundry*, *Book* images (from left to right) from *Middlebury* datasets. (a) Our result; (b-f) The results from other methods. For the purpose of visual inspection, the regions highlighted with yellow rectangles are enlarged.

TABLE III

QUANTITATIVE EVALUATION OF MULTI-OBSERVED IMAGE SR RESULTS (4×) OVER THREE TESTING IMAGES. BOLD TEXT INDICATES THE BEST VALUE

Method	Book			Laundry			Moebius		
	JP2K	ReBlur	BRISQUE	JP2K	ReBlur	BRISQUE	JP2K	ReBlur	BRISQUE
Shan08	7.35	0.53	56.26	7.87	0.51	59.24	8.79	0.56	59.68
Yang10	8.24	0.50	52.28	8.38	0.49	56.36	8.92	0.53	52.91
ANR13	7.85	0.58	57.98	8.05	0.57	60.41	8.57	0.60	60.37
SCRNN14	8.32	0.48	51.11	8.23	0.47	54.51	8.87	0.53	53.86
JOR15	8.09	0.45	48.88	8.21	0.46	51.92	8.65	0.50	46.50
Ours	**9.08**	**0.35**	**43.61**	**8.95**	**0.38**	**50.25**	**9.58**	**0.40**	**39.74**

values of RGB and Depth decrease gradually. As the iteration number increases to 18, the curves reach a steady state for both RGB and Depth SR. For the non-aligned case, the offset is 4 pixels between the RGB-D pairs. Note that after 3 iterations, our algorithm can effectively correct the offset. Then, the super-resolved depth image can further improve the accuracy of the constructed mapping matrix M^{v}. Our algorithm can achieve convergence for non-aligned case within 20 iterations. Therefore, we set the number of iterations to 18.

In all of our experiments, we set the parameters according to the aforementioned analysis, unless otherwise indicated.

B. RGB Image Super-Resolution Evaluations

To demonstrate the high performance of our method through visual inspection and quantitative evaluation, we conducted SR experiments using two types of datasets: the originally-provided multi-viewpoints RGB-D images, and the down-sampled multi-observed RGB-D empirically obtained images. For the first case, we select four RGB images and a depth image from the *Middlebury* dataset [40] and NYU dataset [41] as the input images. For Li's empirical data [13], we select one RGB-D pair as the input images. To verify and evaluate the robustness of our algorithm, we also perturb the RGB images with Gaussian noise. For the second case, we generate the four LR images by convoluting the HR image with a Gaussian kernel and then down-sampling the smoothed images. Here, the sampling interval is in accordance with the magnification factor, and the offset between the multi-observation images is set to be half of the magnification factor. The depth information has been taken into account by our algorithm, however, the other competing methods all ignored such information. In this section, we will evaluate the RGB SR, while the depth map SR evaluation will be detailed in Section V-C.

1) Comparison for RGB Upsampling: For the experiments on the first type of dataset, Fig. 8 presents the SR results. From the zoom-in effects, the edges are blurred for Shan08 [32] and Yang10 [7]. This indicates that the gradient distribution prior (Shan08) cannot preserve edges very well, and the dictionary pair learned by Yang10 is not suitable for all testing images. Despite performing better than Yang10, the methods (ANR13 [9], SCRNN14 [24], and JOR15 [10]) still faces with some problems: the generated edges are not sharp enough (see the zoom-in effects of *Moebius* and *Laundry*) and some details are missing (see the enlarged effects of *Book*). There is an inherent defect in learning-based methods because of the lower similarity between the training set and testing set. In contrast, our method can reproduce sharp edges and achieve the best visual quality. Benefiting from the NBTV prior and the multi-observation data, our method can not only better maintain the sharpness of the edges but also reconstruct the small details, which gives rise to better visual perception. Moreover, the JP2K, Reblur, and BRISQUE values documented in Table III demonstrate that our method outperforms all other competing methods. Therefore, our method can produce more satisfactory HR results with sharper edges.

We also performed experiments using the NYU dataset [41] to verify the versatility of our method. The upsampling results (4×) are shown in Fig. 9. We observe that bicubic and Yang08 [7] tend to smooth out the sharp edges. As shown in Fig. 9 (c-e), ANR13 [9], SCRNN14 [24], and JOR15 [10] fail to preserve the small details (such as the rich textures of a table). In contrast, our algorithm produces the best results (Fig. 9 (f)); it preserves the sharp edges and

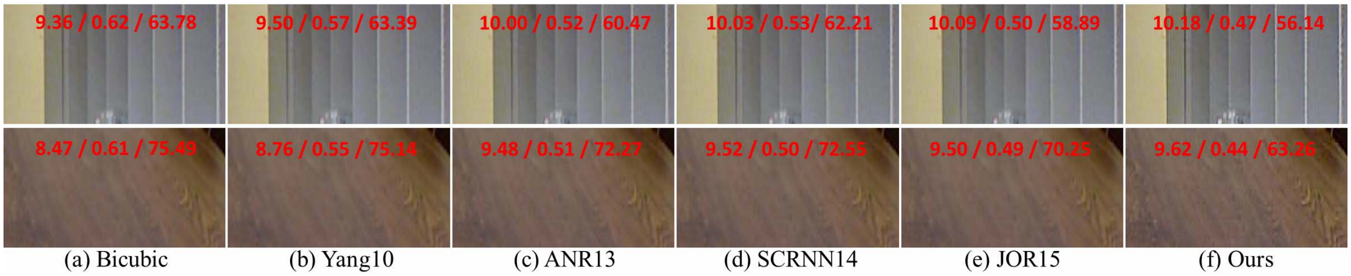


Fig. 9. Comparison of RGB image super-resolution results (4×) for the NYU data. (a) Bicubic, (b) Yang10 [7], (c) ANR13 [9], (d) SCRNN14 [24], (e) JOR15 [10], (f) our method. JP2K/ReBlur/BRISQUE values are presented by red letter.

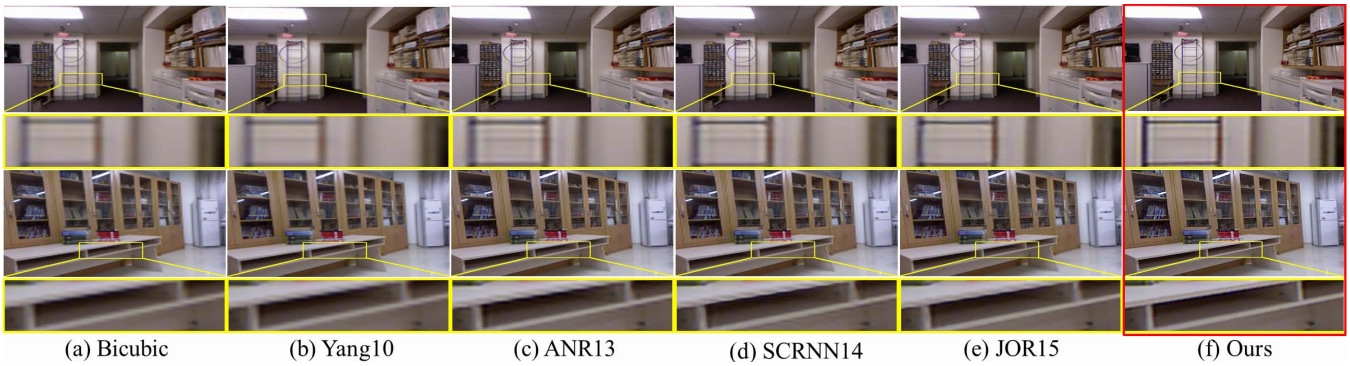


Fig. 10. Comparison of RGB image super-resolution results (4×) under one viewpoint on the NYU data (top row) and Li's empirical data (bottom row). (a) Bicubic, (b) Yang10 [7], (c) ANR13 [9], (d) SCRNN14 [24], (e) JOR15 [10], (f) our method.

effectively recovers the textures. Our method also achieves the best JP2K/ReBlur/BRISQUE values of the competing algorithms. Therefore, our method is better at preserving structure (guided by the newly proposed prior, NBTV), and recovers small details better (based on the multi-image information extraction).

To guarantee fairness, Fig. 10 shows the SR results under one viewpoint for the NYU data and Li's empirical dataset. Based on the zoom-in effects of the NYU data, we can see that all the competing methods produce severely blurred edges. In contrast, our algorithm produces better SR results with sharper edges, and it outperforms all competing methods. This demonstrates that the proposed prior, NBTV, can better maintain the sharpness of the edges. Meanwhile, from the super-resolved results of Li's empirical data, we can see that bicubic, Yang10 [7], ANR13 [9], and SCRNN14 [24] introduce staircase artifacts around the edges. Although JOR15 [10] has much better performance than those methods, it still produces some staircase artifacts. Benefiting from the NBTV prior, our method effectively suppresses the staircase artifacts and produces better results with sharper edges.

Fig. 11 shows a comparison between single image and multi-image cases to provide further insight into the effectiveness of the multi-image strategy. As shown in Fig. 11 (a), the proposed prior, NBTV, can effectively eliminate the staircase artifacts introduced by the original TV prior. However, the reconstructed edges are still not sharp, and some small details are smoothed (highlighted with yellow arrow). Introducing multiple images into the optimization function helps resolve these problems. Based on the zoom-in effects in Fig. 11 (b), we can see that the sharpness of edges can be effectively improved, and the small details are recovered. Therefore, the super-resolved results obtained by the multi-image method can guarantee better visual perception. This is because multiple images can provide enough redundant information for faithful RGB reconstruction, and the constraints of multiple images can guarantee that the objective optimization function will achieve better reconstruction.

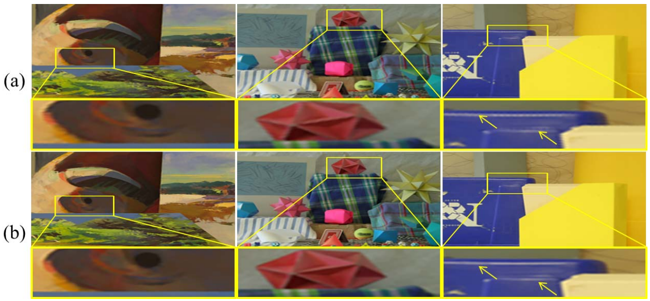


Fig. 11. Comparison of RGB super-resolution results (4×) under one viewpoint (a) and four viewpoints (b).

2) Robustness Verification: We perturbed RGB images with 15dB Gaussian noise, and Fig. 12 shows the RGB SR results obtained using different methods. The input RGB-D pairs were from the *Middlebury* dataset, NYU dataset, and Li's empirical data. Based on the enlarged effects in Fig. 12, we can see that the bicubic method tends to introduce color distortion around the edges that are affected by noise. Yang10 [7] and AAR13 [9] present the smoothed results, which suppress the

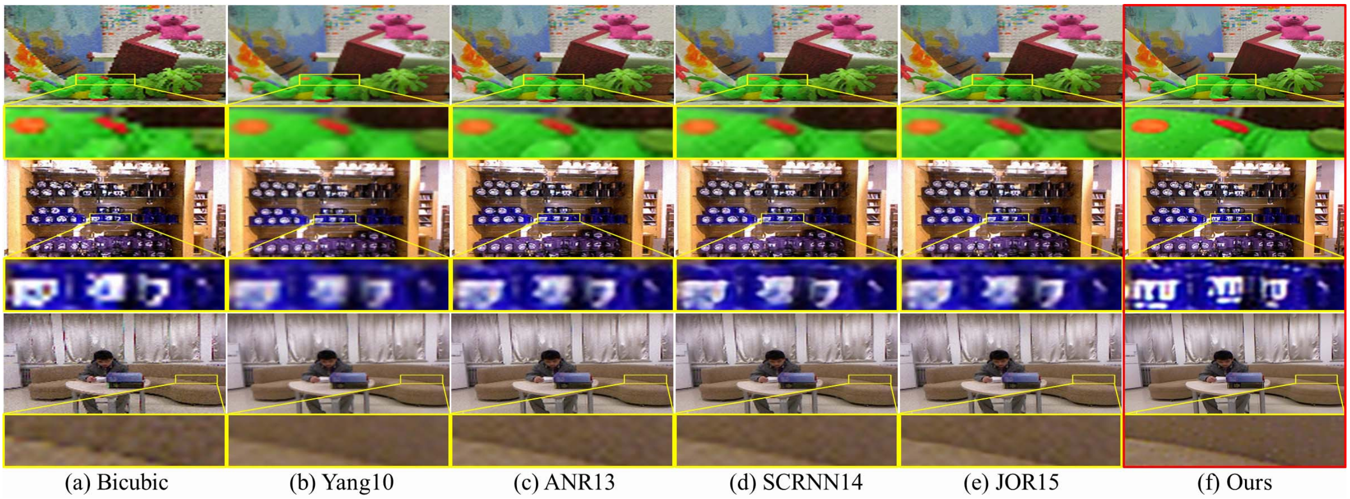

(a) Bicubic (b) Yang10 (c) ANR13 (d) SCRNN14 (e) JOR15 (f) Ours

Fig. 12. Comparison of Gaussian noise-perturbed RGB super-resolution results (4×) for the *Moebius* image (top row), the NYU data (middle row), and Li's empirical data (bottom row). (a) Bicubic, (b) Yang10 [7], (c) ANR13 [9], (d) SCRNN14 [24], (e) JOR15 [10], (f) our method.

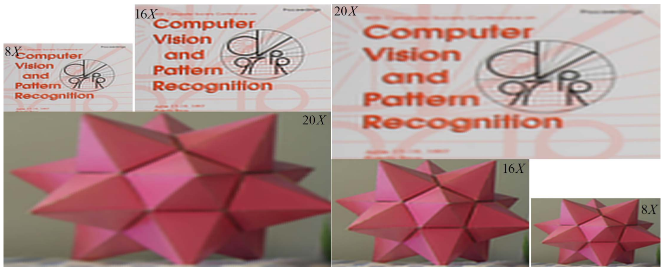


Fig. 13. Our RGB image SR results under three magnification factors on two image patches cropped from images *Book* and *Moebius*.

TABLE IV

QUANTITATIVE EVALUATION FOR DIFFERENT MAGNIFICATION FACTORS (8×, 16×, 20×) FOR THE *Book* AND *Moebius* IMAGES

	Book			Moebius		
	8×	16×	20×	8×	16×	20×
JP2K	8.16	7.33	6.57	9.64	8.17	7.26
ReBlur	0.46	0.74	0.82	0.59	0.72	0.80
BRISQUE	60.80	72.73	76.12	55.55	74.00	77.23

noise due to the inherent denoising of the learning-based method. SCRNN14 [24] and JOR15 [10] provide comparable results, with sharper edges due to the joint optimization strategy. Some artifacts are introduced in flat areas, even though the noise is effectively removed. In sharp contrast, benefiting from the constraint capacity of the NBTV prior and the redundant information contained in multiple images, our method is indeed capable of producing sharp edges and suppressing the noise.

3) High Magnification Results: Fig. 13 demonstrates the visual quality of our results under three magnification factors (8×, 16×, 20×), and the corresponding quality indicators are listed in Table IV. For the 8× and 16× cases, our method can generate desirable results with sharp edges and sufficient details. However, for the 20× case, close inspection shows that the edges are blurred, and ringing artifacts and staircase artifacts are introduced around the edges. This demonstrates that our method is reliable for RGB-D SR when the magnification factor is less than 16×.

4) Experiments Over the Synthetic Dataset (Second-Type Dataset): We present the SR results in Fig. 14, and list PSNR and RMSE values in Table V. Based on the zoom-in effects, we see that the bicubic method and Shan08 [32] fail to reconstruct satisfactory edges; they miss most of the image details due to a lack of well-defined structure awareness. The Yang10 [7] method generates many blurred edges and introduces artifacts around the edges. Additionally, based on the zoom-in effects on the right side of Fig. 14, ANR13 [9], SCRNN14 [24], and JOR15 [10] have eliminated the artifacts around the edges. However, blurring occurs (see the hair of the doll and the red edge of the bowling pin) because the ambiguity of correspondence between HR and LR patches is an inherent problem in learning-based methods. This leads to blurring at the edges and shape distortion, especially for thin edges (the crack between the rocks). Benefiting from the NBTV prior and multi-observation data, our method produces results more consistent with the empirical observations around the edges for all images. Meanwhile, the PSNR and RMSE statistics listed in Table V show that our method outperforms the competing methods. This proves that NBTV can effectively reconstruct sharp edges and eliminate the artifacts.

To verify the generic advantages of our method, Fig. 15 (a-b) present the statistic curves corresponding to the two indicators over *Middlebury* dataset. Fig. 15 (c-d) shows the average values of PSNR and RMSE. As shown in Fig. 15, the performance improves significantly when our method is used, indicating that an NBTV-based multi-image method can indeed help improve the accuracy of the super-resolved results, and it outperforms other competing methods.

5) Prior Analysis: To investigate the effects of our newly proposed prior, three images were used for evaluations. Table VI lists the PSNR and RMSE values for different priors. Fig. 16 presents the results generated by the TV, BTV, and NBTV on the *Cone* image. We find that the TV-based

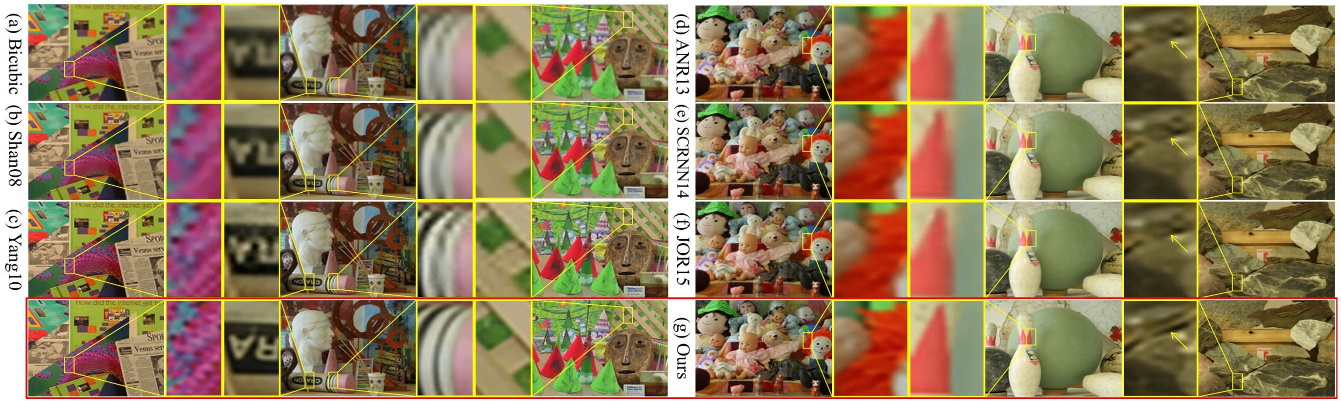
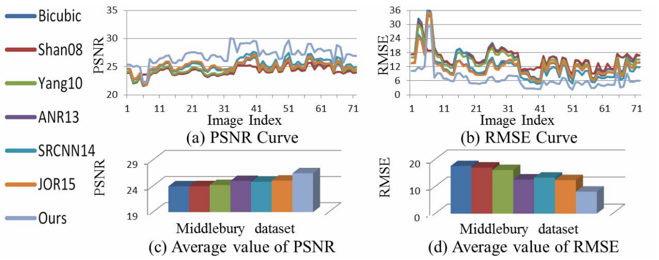

(a) Bicubic (b) Shan08 (c) Yang10
(d) ANR13 (e) SCRNN14 (f) JOR15
(g) Ours

Fig. 14. Visual quality comparison for RGB image super-resolution (4×) on *Middlebury* datasets. The testing images from left to right are *Poster*, *Art*, *Cone*, *Dolls*, *Bowling*, and *Rock*. (a) Bicubic, (b) Shan08 [32], (c) Yang10 [7], (d) ANR13 [9], (e) SCRNN14 [24], (f) JOR15 [10], (g) Our method. The regions highlighted with yellow rectangles are enlarged for better visual inspection.

TABLE V

QUANTITATIVE EVALUATION OF RGB IMAGE SR RESULTS (4×) ON SIX IMAGES. THE BEST VALUE IS HIGHLIGHTED IN BOLD

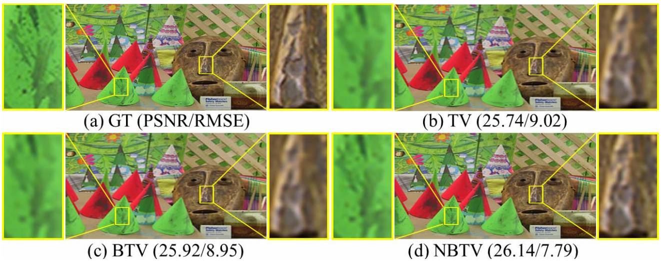
Method	Poster		Cone		Art		Dolls		Bowling		Rock	
	PSNR	RMSE	PSNR	RMSE	PSNR	RMSE	PSNR	RMSE	PSNR	RMSE	PSNR	RMSE
Shan08	23.51	19.31	24.05	15.99	24.62	13.71	24.16	15.44	25.54	9.58	24.91	11.93
Yang10	23.91	16.83	24.54	13.52	25.42	9.98	25.12	11.09	26.35	7.24	25.60	9.39
ANR13	24.01	16.24	24.69	12.82	25.75	8.89	25.38	10.13	26.57	6.71	25.79	8.78
SCRNN14	24.03	16.11	24.71	12.79	25.77	8.84	25.39	10.08	26.61	6.62	25.80	8.75
JOR15	24.07	15.93	24.75	12.59	25.86	8.56	25.45	9.88	26.65	6.52	25.85	8.61
Ours	**25.29**	**10.43**	**26.14**	**7.79**	**28.02**	**4.07**	**27.54**	**4.79**	**28.44**	**3.52**	**27.70**	**4.54**



(a) PSNR Curve (b) RMSE Curve

(c) Average value of PSNR (d) Average value of RMSE

Fig. 15. Quantitative evaluation of RGB super-resolution quality for the *Middlebury* dataset. (a) PSNR curve, (b) RMSE curve, (c) Average value of PSNR, (d) Average value of RMSE.



(a) GT (PSNR/RMSE) (b) TV (25.74/9.02)

(c) BTV (25.92/8.95) (d) NBTV (26.14/7.79)

Fig. 16. Visual quality comparison of color image super-resolution (4×) with different priors. (a) Ground truth, (b) TV, (c) BTV, (d) NBTV. For better visual inspection, regions highlighted with yellow rectangles are enlarged.

TABLE VI

QUANTITATIVE EVALUATION OF DIFFERENT-PRIOR RESULTS (4×) FOR THREE IMAGES. THE BEST VALUE IS HIGHLIGHTED IN BOLD

Prior	Poster		Cone		Art	
	PSNR	RMSE	PSNR	RMSE	PSNR	RMSE
TV	24.90	11.95	25.74	9.02	27.44	4.97
BTV	24.91	11.94	25.92	8.95	27.48	4.91
NBTV	**25.29**	**10.43**	**26.14**	**7.79**	**28.02**	**4.07**

method (Fig. 16(b)) tends to blur the edges and remove small details. That is an inherent defect for TV-based methods. Using higher-degree derivatives to exploit the local structure of the high-resolution image, the BTV-based method (Fig. 16(c)) can achieve better visual quality than the TV-based method. However, it is difficult to preserve the small details. In contrast, our NBTV prior is a normalized version of the BTV prior; it can better respect the sharpness and consistency of the edges and details. Specifically, the result (Fig. 16(d)) indicates that our method can achieve high visual quality, particularly around the edges. Table VI proves that our prior fully outperforms the competing priors according to two qualitative indicators.

C. Color-Guided Depth Image Super-Resolution Evaluations

1) Comparison for Depth Upsampling: We conducted more color-guided depth up-sampling experiments to verify the effectiveness of our nonlocal regression strategy. Three images (*Book*, *Dolls*, *Laundry*) were used for evaluations. Table VII lists the statistical values of the depth SR results (16×) for each RGB-D pair. Fig. 17 presents 16× depth results for *Dolls* and *Laundry*. As shown in Table VII, our method achieves the best PSNR and RMSE values, indicating that our method is effective for high magnification factors. For visual comparison, TGV13 introduces annoying artifacts in

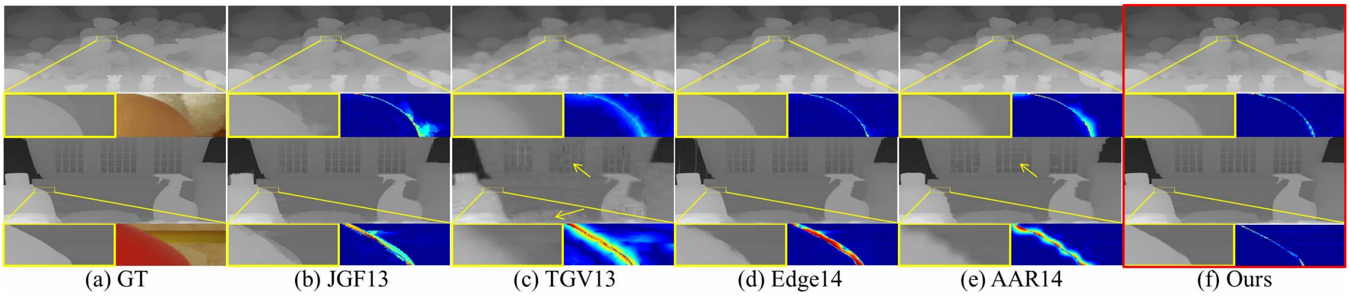


Fig. 17. Visual quality comparison of color-guided depth image super-resolution (16×) for two images (*Dolls* and *Laundry*). (a) Ground truth, the colored patches are the corresponding high-resolution ones used to guide depth up-sampling, (b) JGF13 [12], (c) TGV13 [11], (d) Edge14 [15], (e) AAR14 [13], (f) Our method. For better visual inspection, regions highlighted with yellow rectangles are enlarged, and the corresponding error maps are presented.

TABLE VII

QUANTITATIVE EVALUATION OF COLOR-GUIDED DEPTH SR RESULTS (16×). BOLD TEXT INDICATES THE BEST VALUE

Method	Book		Dolls		Laundry	
	PSNR	RMSE	PSNR	RMSE	PSNR	RMSE
JGF13	27.42	5.00	27.43	4.99	26.47	6.95
TGV13	27.32	5.17	28.41	3.56	26.73	6.36
Edge14	27.28	5.24	29.52	2.42	28.07	3.99
AAR14	28.97	2.93	29.64	2.32	27.70	4.55
Ours	**30.80**	**1.55**	**31.46**	**1.23**	**29.64**	**2.32**

most regions where the corresponding RGB image has rich textures (e.g., cloth and the plastic crate of *Laundry* highlighted with arrows). AAR14 fails to preserve tiny structures, and some fine structures are not recovered, such as the plastic crate of *Laundry* (highlighted with arrow). JGF13 introduces some artifacts around the edges, and TGV13 produces an over-smoothed edge. AAR14 and Edge14 also encountered smoothing problems for high magnification factors (16×) because of the average strategy used in the kernel of the color term. In contrast, our method retains the sharpness of the edges. According to close-up inspection and error maps, our method produces promising results with better geometrical structures; the reconstructed edges are much sharper than all other competing methods. Thus, the introduced channel selection in Eq. 11 can better distinguish the local geometrical structures, and our nonlocal regression can leverage more effective candidate pixels.

Fig. 18 presents 4× depth results for NYU data. Here, the input HR RGB images are generated by our algorithm. At low magnification, all methods can generate good visual quality in flat areas. However, most of the methods encounter difficulties in reconstructing sharp edges. In comparison, our method outperforms the competing methods, which can better recover structures along depth discontinuities. The sharper edges can be found in the zoom-in effects. Specifically, the JP2K/ReBlur/BRISQUE values indicate that our method performs far better than most of the state-of-the-art methods in terms of overall quality.

2) Visual Quality Evaluation for the Non-Aligned Case: Fig. 19 compares depth upsampling experiment results for Li's empirical data. Here, the input RGB and depth pair is non-aligned. According to the zoom-in effects, the bicubic, JGF13, and AAR14 introduce staircase artifacts around the discontinuity regions (such as edges) because the offset results in a mismatch between the texture details in the color image and depth structure. TGV13 tends to produce fuzzy edges. In contrast, our method achieves better visual quality around the edges, which means that the channel selection based nonlocal regression function is better at maintaining structural consistency between RGB and depth images.

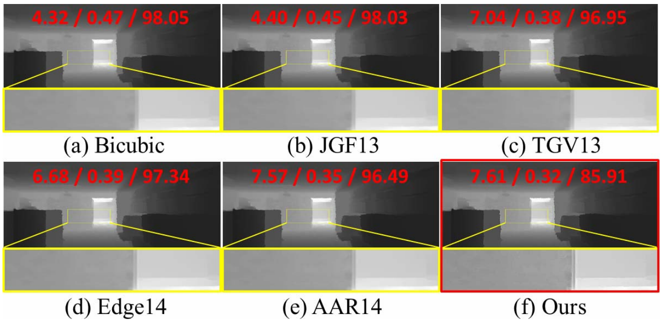


Fig. 18. Comparison of depth super-resolution results (4×) for NYU data. The high-resolution RGB image is our SR result. (a) Bicubic, (b) JGF13 [12], (c) TGV13 [11], (d) Edge14 [15], (e) AAR14 [13], (f) Our method. JP2K/ReBlur/BRISQUE values are presented by red letter.

3) Robustness Evaluation: Here, we simulate ToF-like noise by perturbing the original images with 15 dB Gaussian noise and downsampling the polluted images at 8-upsampling rates. Fig. 20 shows 8× upsampled depth results for ToF-like noise-perturbed RGB-D pairs for the *Art* and *Reindeer* images. We observe that both the bicubic method and JGF13 are more sensitive to noise in the upsampling because they lack a denoising capability, which seriously decreases the visual quality. TGV13 produces over-smoothed edges even though it suppressing the noise. AAR14 can produce good visual results, but it fails to preserve the sharpness of edges and small details. By comparison, it is obvious that our method can not only suppress ToF-like noise but also preserve the sharpness of edges, indicating the robustness of our method. For the depth discontinuities and flat regions, Kinect-like noises are respectively simulated with structurally-missing data and randomly-missing data. Fig. 21 presents the depth recovery results for Kinect-like noise cases in the NYU data and Li's empirical data. According to the comparison, AAR14 is very likely to introduce annoying staircase artifacts around the edges, and it also gives rise to blurred edges. However, our method performs

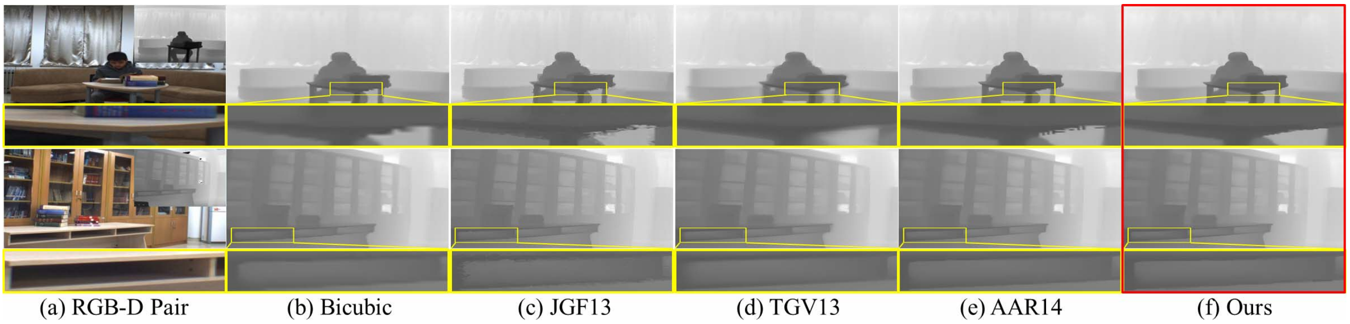


Fig. 19. Comparison of depth super-resolution results from a resolution of 200 × 200 to one of 640 × 480 for Li's empirical data. Here, RGB-Depth pair is non-aligned. (a) RGB-D pair, (b) Bicubic, (c) JGF13 [12], (d) TGV13 [11], (e) AAR14 [13], (f) our method.

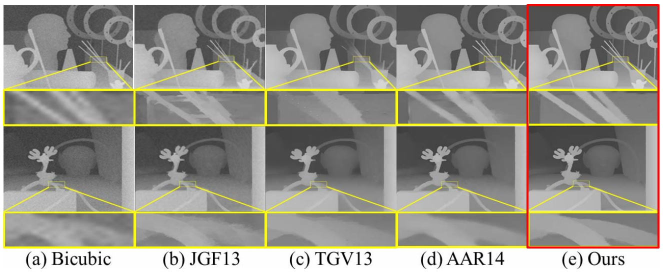


Fig. 20. Comparison of ToF-like noise-perturbed depth super-resolution results (8×) for the *Art* and *Reindeer* images. (a) Bicubic, (b) JGF13 [12], (c) TGV13 [11], (d) AAR14 [13], (e) Our method.

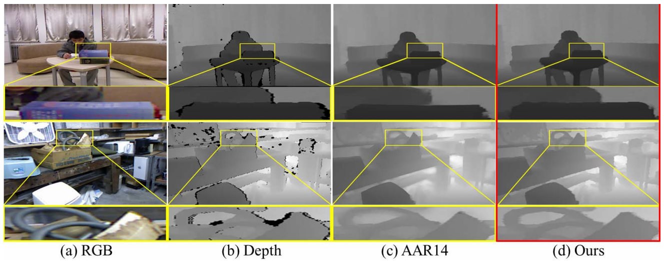


Fig. 21. Comparison of Kinect-like noise-perturbed depth recovery using Li's empirical data (top) and the NYU data (bottom). (a-b) the input RGB and Kinect-like noise perturbed depth images, (c) AAR14 [13], (d) Our method.

well in recovering and preserving the discontinuities in depth images. This demonstrates the robustness and versatility of our adaptive nonlocal regression model.

VI. Discussion and Conclusion

In this paper, we have systematically described a novel optimization framework to address a suite of research challenges in RGB-D image SR. Our technical solutions intrinsically leverage the depth-color correlations to effectively guide the SR process, improve accuracy and increase structural awareness. Specifically, we defined a normalized bilateral TV to regularize the RGB image SR, and we developed a nonlocal regression model to increase the stability of depth image estimation by coupling a channel selection strategy with a context-specific local filter kernel. We also explored the mapping of multi-view images based on the pin-hole imaging model. Extensive experiments on several public datasets, together with comprehensive quantitative evaluations, have demonstrated the superior performance of our method.

However, our method still has some limitations that should be addressed in our upcoming studies. First, the method is relatively time-consuming. As larger images are used, the computational cost will increase rapidly. Moreover, our method cannot effectively accommodate multi-observed images captured from arbitrary viewpoints because it requires a complicated and robust RGBD registration in the 3D scene. Our algorithm can still encounter difficulties for the effective handling of the occlusion problem, which may be solved by incorporating the idea of photo-consistency [17]. Additionally, it is equally important and deserves more research to develop a more general model for spatio-temporal RGB-D video SR.

References

[1] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1167–1183, Sep. 2002.

[2] G. Chantas, N. P. Galatsanos, R. Molina, and A. K. Katsaggelos, "Variational Bayesian image restoration with a product of spatially weighted total variation image priors," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 351–362, Feb. 2010.

[3] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[4] S. Korman and S. Avidan, "Coherency sensitive hashing," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1607–1614.

[5] M. K. Ng, H. Shen, E. Y. Lam, and L. Zhang, "A total variation regularization based super-resolution reconstruction algorithm for digital video," in *Proc. EURASIP J. Adv. Signal Process.*, Jun. 2007, pp. 1–16.

[6] J. Sun, J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[7] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[8] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Curves Surf.*, 2010, pp. 711–730.

[9] R. Timofte, V. De, and L. V. Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1920–1927.

[10] D. Dai, R. Timofte, and L. V. Gool, "Jointly optimized regressors for image super-resolution," *Eurographics*, vol. 34, no. 2, pp. 1–10, 2015.

[11] D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 993–1000.

[12] M.-Y. Liu, O. Tuzel, and Y. Taguchi, "Joint geodesic upsampling of depth images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 169–176.

[13] J. Yang, X. Ye, K. Li, C. Hou, and Y. Wang, "Color-guided depth recovery from RGB-D data using an adaptive autoregressive model," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3443–3458, Aug. 2014.

[14] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. S. Kweon, "High quality depth map upsampling for 3D-TOF cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1623–1630.

[15] J. Park, H. Kim, Y. Tai, M. S. Brown, and I. S. Kweon, "High-quality depth map upsampling and completion for RGB-D cameras," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5559–5572, Dec. 2014.

[16] L. Sheng, K. N. Ngan, C.-L. Lim, and S. Li, "Online temporally consistent indoor depth video enhancement via static structure," *IEEE Trans. Image Process.*, vol. 24, no. 7, pp. 2197–2211, Jul. 2015.

[17] U. Mudenagudi, A. Gupta, L. Goel, A. Kushal, P. Kalra, and S. Banerjee, "Super resolution of images of 3D scenecs," in *Proc. 8th Asian Conf. Comput. Vis.*, 2007, pp. 85–95.

[18] J. W. Silva, L. Gomes, K. A. Agüero, O. R. P. Bellon, and L. Silva, "Real-time acquisition and super-resolution techniques on 3D reconstruction," in *Proc. 20th IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 2135–2139.

[19] A. V. Bhavsar and A. N. Rajagopalan, "Resolution enhancement in multi-image stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1721–1728, Sep. 2010.

[20] H. S. Lee and K. M. Lee, "Simultaneous super-resolution of depth and images using a single camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogint.*, Jun. 2013, pp. 281–288.

[21] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3467–3478, Aug. 2012.

[22] Y. Zhu, Y. Zhang, and A. L. Yuille, "Single image super-resolution using deformable patches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2917–2924.

[23] C.-Y. Yang and M.-H. Yang, "Fast direct super-resolution by simple functions," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 561–568.

[24] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.

[25] H. A. Aly and E. Dubois, "Image up-sampling using total-variation regularization with a new observation model," *IEEE Trans. Image Process.*, vol. 14, no. 10, pp. 1647–1659, Oct. 2005.

[26] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1327–1344, Oct. 2004.

[27] D. Mitzel, T. Pock, T. Schoenemann, and D. Cremers, "Video super resolution using duality based TV-L^1 optical flow," in *Proc. 31st DAGM Symp. Pattern Recognit.*, vol. 5748. 2009, pp. 432–441.

[28] M. Unger, T. Pock, M. Werlberger, and H. Bischof, "A convex approach for variational super-resolution," in *Proc. 32nd DAGM Conf. Pattern Recognit.*, vol. 6376. 2010, pp. 313–322.

[29] W. Shao, "Image modeling based regularized multi-frame super-resolution reconstruction," Ph.D. dissertation, School Comput. Sci. Technol., Nanjing Univ. Sci. Technol., Nanjing, China, 2008.

[30] Y. Hu and M. Jacob, "Higher degree total variation (HDTV) regularization for image recovery," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2559–2571, May 2012.

[31] Z. Xu, X. Su, and Z. Zhang, "Multi-frame image super-resolution by total-variation regularization," *J. Inf. Comput. Sci.*, vol. 9, no. 4, pp. 945–953, 2012.

[32] Q. Shan, Z. Li, J. Jia, and C.-K. Tang, "Fast image/video upsampling," *ACM Trans. Graph.*, vol. 27, no. 5, 2008, Art. ID 153.

[33] Q. Yang, R. Yang, J. Davis, and D. Nistér, "Spatial-depth super resolution for range images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[34] J. Li, G. Zeng, R. Gan, H. Zha, and L. Wang, "A Bayesian approach to uncertainty-based depth map super resolution," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 205–216.

[35] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1052–1067, Jun. 2007.

[36] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1838–1857, Jul. 2011.

[37] W. Dong, L. Zhang, R. Lukac, and G. Shi, "Sparse representation based image interpolation with nonlocal autoregressive modeling," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1382–1394, Apr. 2013.

[38] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imag. Vis.*, vol. 40, no. 1, pp. 120–145, 2010.

[39] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex nonsmooth problems," *Math. Program.*, vol. 146, nos. 1–2, pp. 459–494, 2013.

[40] Middlebury Datasets. *Middlebury Stereo Datasets*. [Online]. Available: http://vision.middlebury.edu/stereo/data/, accessed Jan. 2016.

[41] NYU Datasets. *NYU Depth V2*. [Online]. Available: http://cs.nyu.edu/~silberman/datasets/, accessed Jan. 2016.

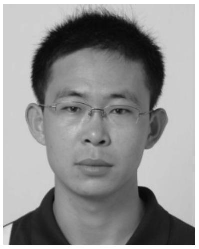
[42] Y. Horita, S. Arata, and T. Murai, "No-reference image quality assessment for JPEG/JPEG2000 coding," in *Proc. 12th Eur. Signal Process. Conf.*, Sep. 2004, pp. 1301–1304.

[43] F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas, "The blur effect: Perception and estimation with a new no-reference perceptual blur metric," *Proc. SPIE*, vol. 6492, pp. 6492–6503, Feb. 2007.

[44] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[45] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4723–4741, Oct. 2009.

Qingzheng Wang received the M.S. degree in computer science from Henan University, in 2010. He is currently pursuing the Ph.D. degree in technology of computer application with Beihang University, Beijing, China. His research interests include computer vision, pattern recognition, and image processing.



Shuai Li received the Ph.D. degree in computer science from Beihang University. He is currently an Assistant Professor with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University. His research interests include computer graphics, pattern recognition, computer vision, physics-based modeling and simulation, and medical image processing.

Hong Qin (SM'08) received the B.S. and M.S. degrees in computer science from Peking University, and the Ph.D. degree in computer science from the University of Toronto. He is currently a Professor of Computer Science with the Department of Computer Science, Stony Brook University. His research interests include geometric and solid modeling, graphics, physics-based modeling and simulation, computer-aided geometric design, visualization, and scientific computing.

Aimin Hao received the B.S., M.S., and Ph.D. degrees from Beihang University, all in computer science. He is currently a Professor with the School of Computer Science and the Associate Director of the State Key Laboratory of Virtual Reality Technology and Systems with Beihang University. His research interests are on virtual reality, computer simulation, computer graphics, geometric modeling, image processing, and computer vision.