The Visual Computer (2020) 36:1411–1422
https://doi.org/10.1007/s00371-019-01748-w

ORIGINAL ARTICLE

Depth image upsampling based on guided filter with low gradient minimization

Hang Yang[1] · Zhongbo Zhang[2]

Published online: 13 September 2019

Abstract
In this paper, we present a novel upsampling framework to enhance the spatial resolution of the depth image. In our framework, the upscaling of a low-resolution depth image is guided by a corresponding intensity images; we formulate it as a cost aggregation problem with the guided filter. However, the guided filter does not make full use of the information of the depth image. Since depth images have quite sparse gradients, it inspires us to regularize the gradients for improving depth upscaling results. Statistics show a special property of depth images, that is, there is a non-ignorable part of pixels whose horizontal or vertical derivatives are equal to ± 1. Based on this special property, we propose a low gradient regularization method which reduces the penalty for horizontal or vertical derivative ± 1, and well describes the statistics of the depth image gradients. Then, we present a solution to the low gradient minimization problem based on threshold shrinkage. Finally, the proposed low gradient regularization is integrated with the guided filter into the depth image upsampling method. Experimental results demonstrate the effectiveness of our proposed approach both qualitatively and quantitatively compared with the state-of-the-art methods.

Keywords Depth image · Upsampling · Low gradient minimization · Guided filter · Regularization method

1 Introduction

Over the last decade, RGB-D sensors have made rapid development, such as Microsoft Kinect, Intel Leap Motion and ASUS Xtion Pro. They enable a variety of applications based on the depth image of the scenes, for instance, pose estimation [1] and scene understanding [2]. Moreover, object tracking using depth information plays a vital role in several applications such as multimedia contexts, body-parts movements, video streaming, healthcare systems and smart indoor security systems [3].

In order to get better tracking and detection performance, researchers explore other sensors: RGB-D, bumblebee and stereo-cameras and their characteristics/properties with respect to object detection and tracking [4].

Moving objects detection and tracking are two important applications of depth cameras. Meanwhile, depth images are widely used for feature representation and extraction, and can assist RGB image to accomplish more complex tasks, such as action recognition, gait recognition, face recognition and behavior recognition [5,6].

However, current depth cameras are limited by manufacturing and physical constraints. Hence, depth images are affected by degenerations due to noise, missing values, and typically have a low resolution [7,8]. To mitigate these problems, we need to recover the corresponding high-resolution (HR) depth image from a given low-resolution (LR) one.

Depth image upsampling is a quite challenging task. Specifically, due to the limited spatial resolution, the LR image loses or distorts fine structures of the HR image. A brute-force upscaling method often makes those structures which have sharp edges become blurred in the upsampled image. In particular, for the case of single-image upscaling, the severely distorted fine structures often exist [9].

To address the above problem, a common approach is to utilize a corresponding HR intensity image as guidance

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00371-019-01748-w) contains supplementary material, which is available to authorized users.

✉ Hang Yang
yanghang09@mails.jlu.edu.cn

[1] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science, Changchun 130033, China

[2] Departments of Mathematics, Jilin University, Changchun 130012, China

[10,11]. This is based on the fact that a correspondence between a depth edge and an intensity edge can most likely be established. Some of the most successful algorithms for upsampling depth images aim at exploiting this correspondence assumption.

In this work, we present a novel method that combines the advantages of guided filter and the energy minimization model to compute an accurate high-resolution output from a LR depth map with the corresponding HR intensity image. In recent years, the guided filter as a new edge-preserving technology has also been employed in a wide range of applications, such as image deconvolution [12], image super-resolution [13] and image fusion [14]. Since the depth image is smooth, it is appropriate to be processed by guided filter which has shown to be effective for textureless image. And the guided filter can effectively fuse images from different sensors. Inspired by these, we attempt to use guided filter for depth image upsampling. However, the properties of depth images are not fully exploited by the guided filter. A shallow observation is that gradients of the depth image are 0 at most places [15]. Therefore, together with the textureless property, we also regularize the depth maps with the sparse gradient prior in the meanwhile.

Based on the statistics of the depth image gradient, we are surprised to find that the sparse gradient model is not accurate enough [16]. In the depth image, although more than 80% pixels have zero gradients (see Fig. 1), there is a non-ignorable part of pixels whose horizontal or vertical derivatives are equal to ± 1 (their proportion is about 15%, see Fig. 1). In other words, at many places, gradients of the depth map do not vanish but are very small [16]. This property has not been considered for depth image upsampling because it is not universal in natural images. Hence, we develop a specific gradient regularization which is denoted as l_0^t gradient regularization. Unlike the l_0 norm which penalizes the nonzero elements equally (the norm is always 1 if the element is not 0) [17,18], our proposed l_0^t measure reduces the penalty for horizontal or vertical derivative ± 1 and thus allows for gradual depth changes.

The main contributions of this work are threefold: (1) We present a specific gradient regularization l_0^t which well describes the statistical property of the depth image gradients. (2) We propose a solution to l_0^t gradient minimization problem based on threshold shrinkage. (3) We integrate the proposed l_0^t gradient regularization with the guided filter into the the depth image upsampling method. In the experiments, we demonstrate that the proposed method provides competitive results compared with the state-of-the-art algorithms.

The rest of the paper is organized as follows: Section 2 briefly introduces related work in depth image upsampling. Section 3 describes the proposed approach which considers the guided filter and the low gradient regularization. In Sect. 4, we perform simulations on the benchmark dataset and show the effectiveness of our method. We conclude the work in Sect. 5.

2 Related work

There are many methods to perform depth image upsampling in the literature. In general, they can be categorized into four classes:

Exemplars-based approaches These approaches build dictionaries for the LR and HR domains that are coupled by a common encoding. Yang et al. [19] seek the coefficients of this representation to obtain an upsampling result, but an important question is how to determine the optional dictionary size. To improve the inference speed, Timofte et al. [20] introduce the anchored neighborhood regression. Li et al. [21] present a joint examples-based upscaling approach. Ferstl et al. [22] present a dictionary learning method with edge priors for an anisotropic guidance. Schulter et al. [23] use random regression forests instead of the flat code-book of sparse coding methods. Mahmoudi et al. [24] denoise noisy samples and learn a depth dictionary from noisy and denoised samples. However, since the reconstruction is highly biased to the available training examples, these methods may not provide reliable results when no correspondence can be established.

Local image filtering Kopf et al. [25] propose a joint bilateral filter-based algorithm to smooth each depth pixel by considering the intensity similarity between the center pixel and its neighborhood. Yang et al. [26] present a method based on the bilateral filter that is iteratively used to generate an upsampled result. However, it is observed that the guidance of color image to upscaling of depth map runs the risk of texture copying and edge blurring, especially in smooth geometry regions. Geodesic distances are used to design the upsampling weights in [27], but it is designed without any consideration of the noise issue from depth sensors. Lu et al. [28] propose a smoothing approach to upscale depth map with the use of image segmentation, but segmentation errors will clearly disrupt the method. Li et al. [29] develop a fast guided interpolation (FGI) approach based on weighted least squares, which densifies depth maps by global interpolation with alternating guidances, but it may generate oversmooth results in some cases.

Global energy minimization methods These approaches formulate depth upscaling as an optimization problem which employs data fidelity and regularization term [30]. Diebel et al. [31] develop Markov Random Field (MRF)-based energy minimization framework, which fuses the LR depth map and the corresponding HR intensity image, but it tends to generate oversmooth results and is also sensitive to noises. In order to maintain local structures, Park et al. [32] use a non-local means filter (MRF+NLM) to regularize the depth map,

but jaggy artifacts occur in some boundaries. A more recent approach of Ferstl et al. [10] utilizes an anisotropic diffusion tensor to guide the depth map upsampling (TGV), and the tensor is calculated from a HR intensity image; however, it is not smooth in some internal areas. Aodha et al. [33] treat depth image upsampling as MRF labeling problem which matches LR depth image patches to HR patches from an ancillary database; the training of data, matching, and fusion are quite computationally intensive. In [34], an adaptive intensity-guided regression method is proposed for depth upsampling, but its relative high computational complexity would limit its practical applications.

Deep learning-based methods More recently, deep learning methods have become popular for single-image upsampling. A convolutional neural network (CNN) of three layers is trained in [35], and Kim et al. [36] improve this approach substantially. Dong et al. [37] present an end-to-end upsampling convolutional neural network (SRCNN) to achieve image super-resolution. These learning-based methods have mainly been applied to color images, and not suitable for depth map super-resolution. Xie et al. [38] propose a CNN framework for the single depth image upsampling guided by a reconstructed HR edge map. These methods have mainly been used to intensity images, where a great amount of training samples can be easily obtained. In contrast, huge datasets with dense, accurate depth maps have recently become available, e.g., [39]. Hui et al. [40] present a CNN framework in a multi-scale guidance architecture (MSG-Net). Riegler et al. [7] integrate an energy minimization model with anisotropic TGV regularization into a end-to-end convolutional network for a single depth image upsampling. However, CNN-based methods are highly dependent on training samples; these methods may not produce competitive results when the test data are not similar to the training set.

3 Depth image upsampling

Given an original HR intensity image I_H and a LR depth image I_L, we hope to obtain a HR depth map u. If I_H is a RGB image, it should be converted from the RGB space to the gray space. We first generate a coarse estimated depth image $D_\uparrow$ by bicubic interpolation from I_L; the resolutions of $D_\uparrow$ and I_H are equal.

In conventional image restoration problems, the guided filter performs very well in terms of both quality and efficiency [12,13]. The filter can smooth image with the edge-preserving property as the bilateral filter, but there are no gradient reversal artifacts. The advantage of guided filter is very suitable for processing depth image, so we introduce the guided filter into the upsampling algorithm. As discussed in Sect. 1, we first add the sparse gradient regularization for depth upsampling. Altogether, we construct a new formula as follows:

$$\min_u \| u - D_\uparrow \|_2^2 + \rho \| u - GF(u, I_H) \|_2^2 + \eta \| \nabla u \|_0 \tag{1}$$

where $\nabla u = (u_x, u_y) = (\partial_x * u, \partial_y * u)$ is the gradient of u, $\partial_x = [1, -1]$ and $\partial_y = [1, -1]^T$ are the horizontal and vertical derivative operators, respectively, ρ and η are two regularization parameters, $\mathbf{GF}(\cdot, \cdot)$ is the guided filter, and the output of $\| \nabla u \|_0$ is the number of nonzero elements in ∇u; we call this method **GFL0**.

Because $\mathbf{GF}(\cdot, \cdot)$ is highly nonlinear, it is difficult to solve the problem directly. Following the solution in [12], we employ a split variable approach to solve Eq. (1) and the variables are iteratively updated:

$$z = GF(u, I_H) \tag{2}$$

$$u = \arg\min_u \| u - D_\uparrow \|_2^2 + \rho \| u - z \|_2^2 + \eta \| \nabla u \|_0 \tag{3}$$

The minimization problem in Eq. (3) is widely used in image restoration models, and many approaches have been proposed to solve it directly and approximately [17,18,41].

l_0 gradient minimization does not always perform well [16]; one of the reasons may be the solution is only an approximation. That is to say, we do not make full use of the properties of the gradient maps. We count the gradient histogram of depth images and find out that the sparse gradient assumption is not accurate enough. In addition to 0, there is a non-ignorable part of pixels that has horizontal or vertical derivative ± 1. In l_0 regularization, all nonzero gradients are penalized equally [16]. Based on the special property of depth images, we propose a new gradient regularization algorithm to reduce the penalty for horizontal or vertical derivative ± 1.

3.1 l_0^t gradient minimization

In this subsection, we will show the statistics of depth image gradients and describe the proposed l_0^t gradient regularization method.

In this work, we use Middlebury Stereo Datasets (2001, 2003, 2005, 2006 and 2014) to do statistics on the depth gradients, and show the horizontal and vertical derivative magnitude histograms in Fig. 1. It is observed that depth image gradients cannot be simply described as sparse. We can see that most pixels have gradient magnitude 0 and a non-ignorable part whose gradients are $(\pm 1, \pm 1)$, $(0, \pm 1)$ or $(\pm 1, 0)$. Similar to the Total variation (TV) model, we propose a l_0^t norm to reduce the penalty for horizontal and vertical derivatives ± 1.

We define the $\| \cdot \|_{l_0^t}$ norm as:

$$\| \nabla u \|_{l_0^t} = \| u_x \|_{l_0^t} + \| u_y \|_{l_0^t} \tag{4}$$

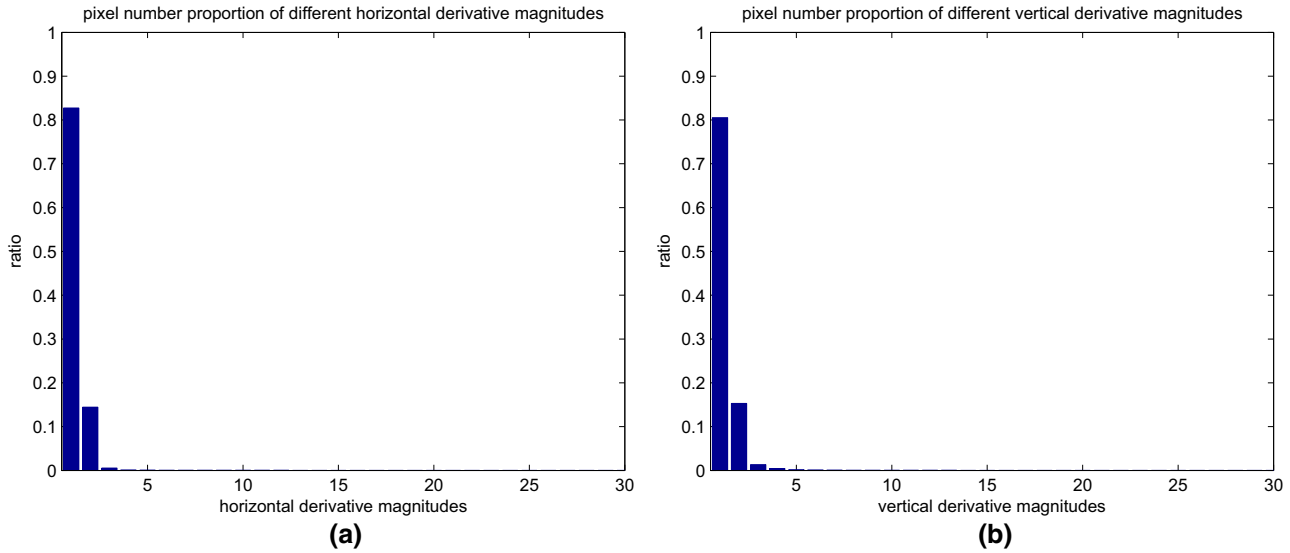


Fig. 1 Horizontal and vertical derivative magnitude histograms of ground-truth disparity map of Middlebury Stereo Datasets. **a** Horizontal derivative. **b** Vertical derivative. We can see that most pixels have gradient magnitude 0 and a non-ignorable part has magnitude 1

where

$$\| f \|_{l_0^t} = \sharp\{i| \mid f_i \mid > 1\} + t\sharp\{i| \mid f_i \mid = 1\} \tag{5}$$

and $0 < t < 1$, $\sharp\{\cdot\}$ is the number of elements in the data. In order to construct a problem that can be applied to a continuous domain, we revise the definition of $\| \cdot \|_{l_0^t}$ as

$$\| f \|_{l_0^t} = \sharp\{i| \mid f_i \mid > 1\} + t\sharp\{i|0 < \mid f_i \mid \leq 1\} \tag{6}$$

Actually the l_0^t "norm" is not a proper norm because it is not homogeneous; therefore, we call it a measure. Based on the statistics on the depth gradients, we set $t = 0.75$ in all the experiments.

Thus, we use the $\| \cdot \|_{l_0^t}$ measure in place of l_0 norm in Eq. (1), and it leads to the following optimization model:

$$\min_u \| u - D_\uparrow \|_2^2 + \rho \| u - GF(u, ref) \|_2^2 + \eta \| \nabla u \|_{l_0^t} \tag{7}$$

where the ref is the I_H in this work.

Similar to Eq. (1), we extend the split variable method to solve Eq. (7). In this case, two subproblems are as follows:

$$z = GF(u, I_H) \tag{8}$$

$$u = \arg\min_u \| u - D_\uparrow \|_2^2 + \rho \| u - z \|_2^2 + \eta \| \nabla u \|_{l_0}^t \tag{9}$$

To solve Eq. (9), we introduce auxiliary variables h and v, corresponding to u_x and u_y, respectively, and rewrite the cost function as:

$$\begin{aligned} \{u, h, v\} = \arg\min_{u,h,v} & \| u - D_\uparrow \|_2^2 + \rho \| u - z \|_2^2 \\ & + \beta(\| h - u_x \|_2^2 + \| v - u_y \|_2^2) \\ & + \gamma(\| h \|_{l_0}^t + \| v \|_{l_0}^t) \end{aligned} \tag{10}$$

where β is an automatically adapting parameter. Equation (11) is solved through alternatively minimizing (h, v) and u, and it is split into three subproblems in this work:

$$\begin{aligned} u = \arg\min_u & \| u - D_\uparrow \|_2^2 + \rho \| u - z \|_2^2 \\ & + \beta(\| u_x - h \|_2^2 + \| u_y - v \|_2^2) \end{aligned} \tag{11}$$

$$h = \arg\min_h \| h - u_x \|_2^2 + \lambda \| h \|_{l_0}^t \tag{12}$$

$$v = \arg\min_v \| v - u_y \|_2^2 + \lambda \| v \|_{l_0}^t \tag{13}$$

Equation (11) is quadratic and thus has a global minimum. We use fast Fourier transform (FFT) to speedup the diagonalization of derivative operators. These yield solutions in the Fourier domain

$$\mathcal{F}(u) = \frac{\mathcal{F}(D_\uparrow) + \rho\mathcal{F}(z) + \beta(\mathcal{F}(\partial_x)^*\mathcal{F}(h) + \mathcal{F}(\partial_y)^*\mathcal{F}(v))}{1 + \rho + \beta(|\mathcal{F}(\partial_x)|^2 + |\mathcal{F}(\partial_y)|^2)} \tag{14}$$

where $\mathcal{F}$ and $\mathcal{F}(\cdot)^*$ denote the FFT operator and the complex conjugate, respectively. The plus, multiplication, and division are all component-wise operators.

Now, the remaining question is how to solve Eqs. (12) and (13). In next subsection, we will show that these two apparently sophisticated subproblems have closed-form solutions and can be solved quickly.

The L_0 gradient minimization leads to better results in depth upscaling results in most cases. However, it does not

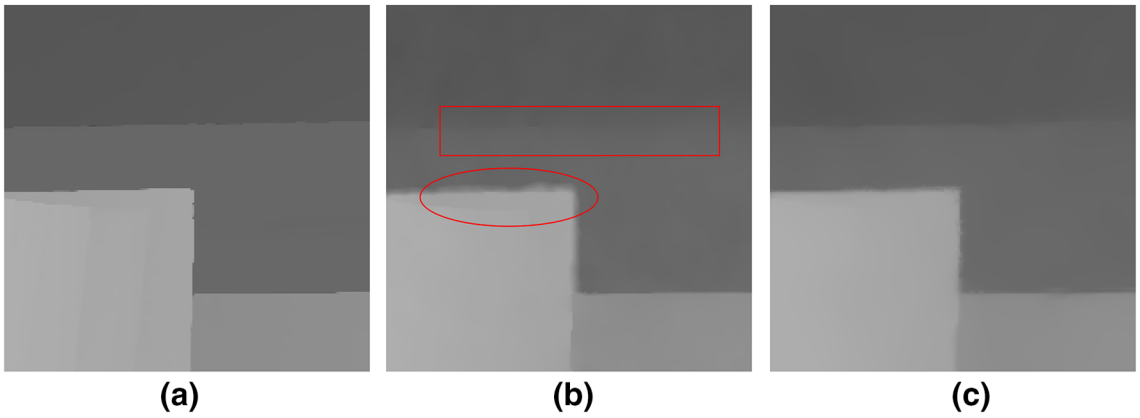


Fig. 2 Illustrations for contrast between the gradient minimization-based approach and our proposed method (scaling factor = 4). **a** Ground truth, **b** GFL0 result, **c** our result

always perform well (see Fig. 2). Figure 2 shows a failure case of gradient minimization.

3.2 The l_0^t measure minimization

Without loss of generality, Eqs. (12) and (13) are written in a unified way:

$$\sum_i \min_{x_i}\{(x_i - p_i)^2 + \alpha H^t(p_i)\} \tag{15}$$

where

$$H^t(p) = \begin{cases} 0 & \text{if p = 0} \\ t & \text{if } 0 < \mid p \mid \leq 1 \\ 1 & \text{if } \mid p \mid > 1 \end{cases} \tag{16}$$

Each single term w.r.t. pixel p_i in Eq. (16) is

$$E(p) = \min_p (x - p)^2 + \alpha H^t(p) \tag{17}$$

For Eq. (15), we can obtain its closed-form solution based on the following theorem. In this work, we discuss two cases, that is, the situation of $|x| \geq 1$ and the situation of $|x| < 1$. Theorem 3.1 and Theorem 3.2 give the solution of Eq. (15) in the case of $|x| \geq 1$, and Theorem 3.3 shows the solution of Eq. (15) in the case of $|x| < 1$.

Theorem 1 *When* $|x| \geq 1$, *Eq.* (17) *reaches its minimum* E_p^* *under the condition*

$$p = \begin{cases} 0 & if \mid x \mid \leq \min(\frac{1+\alpha t}{2}, \sqrt{\alpha}) \\ sgn(x) & if \frac{1+\alpha t}{2} < \mid x \mid \leq 1 + \sqrt{\alpha(1-t)} \\ x & if \mid x \mid > \max(1 + \sqrt{\alpha(1-t)}, \sqrt{\alpha}) \end{cases} \tag{18}$$

Proof Without loss of generality, we suppose $x \geq 0$, and the proof of $x < 0$ case is similar.
(1) When $x > \max(1+\sqrt{\alpha(1-t)}, \sqrt{\alpha}) > 1$, nonzero $p > 1$ yields

$$E_p = (x - p)^2 + \alpha \tag{19}$$

when $p = x$, Eq. (19) achieves minimal value α.

Note that $p = 0$ leads to

$$E_p = x^2 > \alpha \tag{20}$$

And we can find that $0 < p \leq 1$ yields

$$E_p = (x - p)^2 + \alpha t \tag{21}$$

when $p = 1$, Eq. (19) achieves minimal value $(x-1)^2 + \alpha t$.

Because $x > 1 + \sqrt{\alpha(1-t)}$, that is to say $(x-1)^2 > \alpha(1-t)$, and $(x-1)^2 + \alpha t > \alpha$.

Comparing Eqs. (19) and (20), the minimal energy E_p is produced when $p = x$.
(2) When $\frac{1+\alpha t}{2} < x \leq 1 + \sqrt{\alpha(1-t)}$, $p > 1$ yields

$$E_p = (x - p)^2 + \alpha \geq \alpha \geq (x-1)^2 + \alpha t \tag{22}$$

$p = 0$ leads to

$$E_p = x^2 > (x-1)^2 + \alpha t \tag{23}$$

When $0 < p \leq 1$, Eq. (21) still holds. we find that, when $x \geq 1$, the minimal energy Eq. (21) is produced when $p = 1$, the minimal value is $(x-1)^2 + \alpha t$.

So, in this case, comparing these three values, when $p = sgn(x) = 1$, E_p achieves minimal.
(3) When $x \leq \min(\frac{1+\alpha t}{2}, \sqrt{\alpha})$, Eq. (19) still holds, the minimal energy α is greater than x^2.

When $p = 0$, E_p has its minimum value x^2.

When $0 < p \leq 1$, if $x \geq 1$, Eq. (21) achieves minimal value $(x-1)^2 + \alpha t \geq x^2$.

So, the minimum energy E_p is produced when $p = 0$ □

In Theorem 3.1, the relationship of three functions $1 + \sqrt{\alpha(1-t)}$, $\sqrt{\alpha}$ and $\frac{1+\alpha t}{2}$ is not determined, and therefore, the reliability of the conclusion cannot be guaranteed. Through further research, we show the three function curves in Fig. 3, and one can clearly see the relationship between them (strict mathematical proofs of the relationship can be found in the supplementary material).

Now, according to the above analysis, we can rewrite Theorem 3.1 as follows:

Theorem 2 *When* $|x| \geq 1$, *and* $\alpha > \frac{2-t+2\sqrt{1-t}}{t^2}$, *Eq.* (17) *reaches its minimal* E_p^* *under the condition*

$$p = \begin{cases} 0 & if \mid x \mid \leq \sqrt{\alpha} \\ x & if \mid x \mid > \sqrt{\alpha} \end{cases} \tag{24}$$

When $|x| \geq 1$, *and* $\frac{2-t+2\sqrt{1-t}}{t^2} \geq \alpha \geq \frac{2-t-2\sqrt{1-t}}{t^2}$, *Eq.* (17) *reaches its minimal* E_p^* *under the condition*

$$p = \begin{cases} 0 & if \mid x \mid \leq \frac{1+\alpha t}{2} \\ sgn(x) & if \frac{1+\alpha t}{2} < \mid x \mid \leq 1 + \sqrt{\alpha(1-t)} \\ x & if \mid x \mid > 1 + \sqrt{\alpha(1-t)} \end{cases} \tag{25}$$

When $|x| \geq 1$, *and* $\frac{2-t-2\sqrt{1-t}}{t^2} > \alpha > 0$, *Eq.* (17) *reaches its minimal* E_p^* *under the condition*

$$p = \begin{cases} sgn(x) & if\, 1 \leq \mid x \mid \leq 1 + \sqrt{\alpha(1-t)} \\ x & if \mid x \mid > 1 + \sqrt{\alpha(1-t)} \end{cases} \tag{26}$$

The detailed proofs of the Theorem 3.2 can be found in the supplementary material.

Theorem 3 *When* $|x| < 1$, *Eq.* (17) *reaches its minimal* E_p^* *under the condition*

$$p = \begin{cases} 0 & if \mid x \mid \leq \sqrt{\alpha t} \\ x & if \mid x \mid > \sqrt{\alpha t} \end{cases} \tag{27}$$

Proof Without loss of generality, we suppose $x \geq 0$, and the proof of $x < 0$ case is similar.
(1) when $x^2 \leq \alpha t$, nonzeros $0 < p \leq 1$ yield

$$E_p = (x-p)^2 + \alpha t \geq \alpha t \geq x^2 \tag{28}$$

Note that $p = 0$ leads to $E_p = x^2$. Comparing Eq. (28), the minimal energy E_p is produced when $p = 0$.
(2) when $\alpha t < x^2 < 1$, $p = 0$ leads to $E_p = x^2$. But when $p = x$, E_p has its minimal value αt. Comparing these two values, the minimal energy E_p is produced when $p = x$. □

An overview of our depth image upsampling framework is provided in Fig. 4, and the proposed algorithm is sketched in Algorithm 1. Once the scaling factor is determined, we use the bicubic interpolation to obtain the initial estimation, and then, guided filter and gradient minimization method are used as follows. Parameter β is automatically adapted in iterations starting from a small value β_0, and it is multiplied by κ each time. This scheme is inspired by Wang et al. [42], which shows that this scheme is effective to speed up convergence. In all the experiments, we fix the regularization parameters $\beta_0 = 0.0025$ and set $\kappa = 2$ to balance the efficiency and performance.

Algorithm 1: Depth Image Upsampling Algorithm

1. Input: LR depth image I_L, HR intensity image I_H, parameter β_0

initialize $u \leftarrow D_\uparrow, \beta = \frac{1}{2}\beta_0, \lambda = 255 \times \frac{\beta}{\beta_0}, \rho = 0.1 \times \beta$ and $\kappa = 2$.

2. Repeat $p = 1 : MAX_{iter}$:

solve for z^p using Eq. (8);

solve for u^p using Eq. (14);

solve for h^p and v^p base on Theorem 3.2 and Theorem 3.3;

update parameters: $\beta \leftarrow \kappa\beta$, $\lambda \leftarrow 255 \times \frac{\beta}{\beta_0}$ and $\rho \leftarrow 0.1 \times \beta$.

3. Output: HR depth image u.

4 Experimental results

To verify the superiority of our method, we evaluate the performance of the proposed method with respect to some state-of-the-art depth image upsampling methods. We perform experiments on a PC with Intel i7-5600U CPU (2.6 GHz) and 8 GB RAM using MATLAB 2012b. $20 - 30$ iterations are generally performed in our method. Most computation is spent on FFT in Eq. (14) and guided image filtering in Eq. (8). Overall, it takes 1.4 seconds per iteration to upsample $\times 4$ to 345×272 depth images.

In this experiment, we evaluate our method on two standard benchmark datasets for depth map super-resolution: Following [2,10,32], we evaluate our results on the noisy Middlebury 2007 dataset. Additionally, in the second evaluation, we compare our method on the challenging ToFMark dataset *Books*, *Devil* and *Shark* which are proposed in [10].

4.1 Noisy Middlebury

In this experiment, we evaluate the performance of the proposed method on *Art*, *Books*, *Moebius*, *Dolls*, *Laundry* and *Reindeer* of the Middlebury dataset [10]. Each set contains a disparity image obtained from structured light and

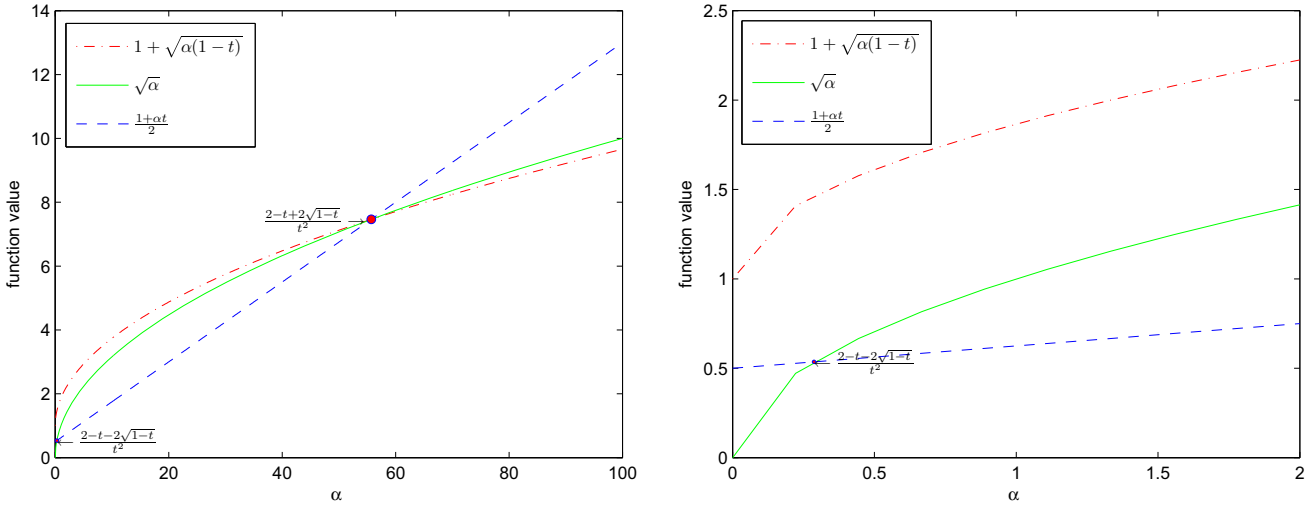


Fig. 3 The relation of the three curves when $t = 0.25$. Left subplot: $\alpha \in [0, 100]$. Right subplot: $\alpha \in [0, 2]$

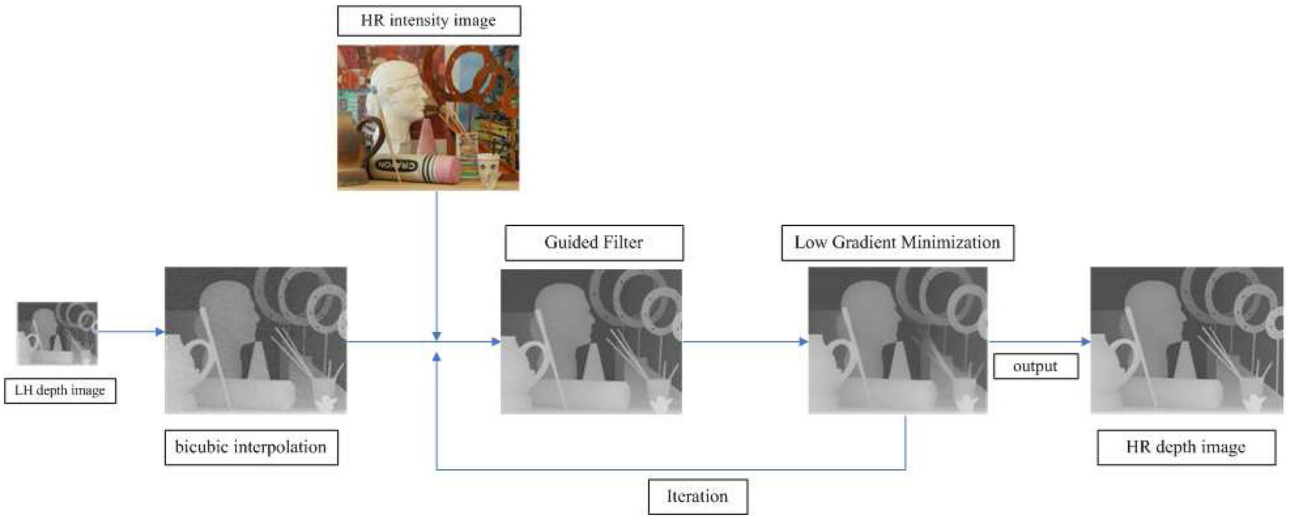


Fig. 4 Framework of our upsampling method

a registered RGB image. We use the RGB image as the HR color image for guidance and the disparity image as the ground truth. They can be available from "http://vision.middlebury.edu/stereo/data/."

To simulate the acquisition process of a time-of-flight sensor, depth-dependent Gaussian noise is added to these input images, as proposed by Park et al. [32]. This dataset consists of three time-of-flight (ToF) depth maps of three different scenes. For each scene, there exists an accurate high-resolution structured light scan as ground truth. The ToF depth maps have a resolution of 120×160 pixels, and the target resolution, given by the guidance intensity image, is 610×810 pixels. This corresponds to an upsampling factor of approximately 5.

We compare our method to simple upsampling methods, such as bicubic interpolation. We compare our proposed method to other approaches that utilize an additional intensity image as guidance. Those methods include the Markov Random Field (MRF)-based approach in [31], the joint bilateral filtering with cost volume (JBFcv) in [26], cross-based local multipoint filtering (CLMF) in [43] the guided image filter (GIF) in [13], the nonlocal means filter (MRF+NLM) in [32], the variational model (TGV) in [10] and fast guided global interpolation (FGI) in [29]. In addition, to illustrate the effectiveness of the low gradient regularization, we also compare with GFL0 proposed in Eq. (1). The GFL0 approach is to minimize the l_0 norm of the gradient, which penalizes the nonzero elements equally. The l_0^t norm proposed in our work also penalizes the nonzero elements, but reduces the penalty for horizontal or vertical derivative ± 1.

Table 1 reports quantitative results in terms of the root mean squared error (RMSE) between ground-truth depth

Table 1 Error as root mean squared error (RMSE) comparison on Middlebury 2007 datasets with added noise for magnification factors (×2 and ×4)

	Art		Books		Moebius		Dolls		Laundry		Reindeer	
	×2	×4	×2	×4	×2	×4	×2	×4	×2	×4	×2	×4
Bicubic	4.78	5.54	4.20	4.38	4.16	4.31	4.16	4.30	4.37	4.74	4.51	4.95
MRF [31]	3.49	4.51	2.06	3.00	2.31	3.11	–	–	–	–	–	–
JBFcv [26]	3.01	4.02	1.87	2.23	1.92	2.42	–	–	–	–	–	–
CLMF [43]	3.29	4.03	1.80	2.38	1.79	2.29	1.83	2.37	2.36	2.91	2.52	3.15
GIF [13]	3.55	4.41	2.37	2.74	2.48	2.83	1.79	2.64	2.33	3.22	2.63	3.43
MRF+NLM [32]	3.74	4.56	1.95	2.61	1.96	2.51	2.06	2.61	2.99	3.63	3.11	3.86
TGV [10]	3.19	4.06	1.52	2.21	**1.47**	2.03	1.49	2.85	2.62	3.44	2.78	3.20
FGI [29]	3.13	4.14	1.48	1.92	1.65	**1.94**	1.47	1.84	1.94	**2.59**	2.22	2.86
GFL0	2.78	3.98	1.40	1.89	1.63	2.04	1.42	1.86	1.90	2.65	2.07	2.79
Ours	**2.71**	**3.87**	**1.34**	**1.82**	1.57	2.01	**1.38**	**1.79**	**1.83**	2.60	**2.01**	**2.76**

We highlight the best result in boldface

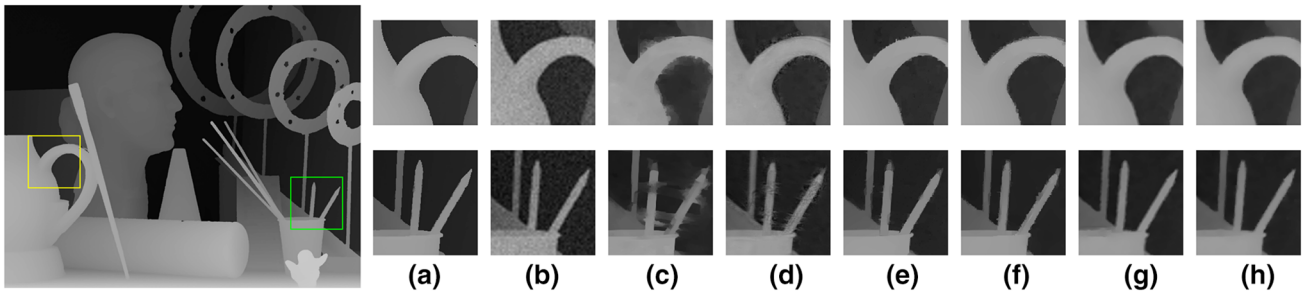


Fig. 5 Visual comparison of *Art* with cropped zoomed regions (scaling factor = 4). **a** Ground truth, **b** Bicubic, **c** MRF+NLM [32], **d** GIF [13], **e** TGV [10], **f** FGI [29], **g** GFL0, **h** our proposed method

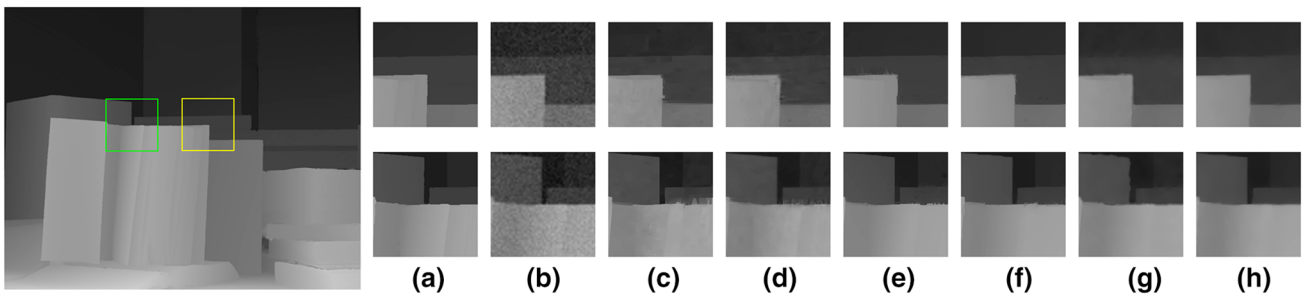


Fig. 6 Visual comparison of *Books* with cropped zoomed regions (scaling factor = 4). **a** Ground truth, **b** Bicubic, **c** MRF+NLM [32], **d** GIF [13], **e** TGV [10], **f** FGI [29], **g** GFL0, **h** our proposed method

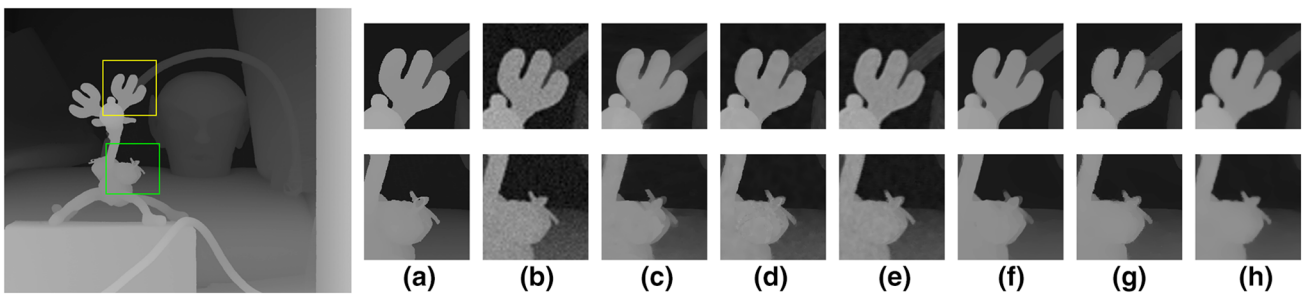


Fig. 7 Visual comparison of *Reindeer* with cropped zoomed regions (scaling factor = 4). **a** Ground truth, **b** Bicubic, **c** MRF+NLM [32], **d** CLMF [43], **e** GIF [13], **f** TGV [10], **g** FGI [29], **h** our proposed method

Table 2 Results on real Time-of-Flight data from the ToFMark benchmark dataset

	Books	Devil	Shark
Bicubic	27.78	26.25	31.68
CLMF [43]	25.67	23.86	28.93
TGV [10]	24.68	23.19	29.89
MSG-Net [40]	25.03	23.07	30.43
Ours	**24.53**	**23.04**	**28.46**

We report the error as RMSE in mm and highlight the best result in boldface

maps and the results by various depth upsampling methods including ours.

From the quantitative results in Table 1, we observe that the proposed method clearly performs better than the state-of-the-art methods that utilize an additional guidance input for most images and upsampling factors.

The proposed method clearly outperforms several existing methods such as GF [13], MRF+NLM [32], CLMF [43] and TGV [10] that used different color-guided upsampling or optimization techniques. Our method also yields much smaller error rates than FGI [29]. Note that the GFL0 approach performs generally better than TGV and achieves similar results to the proposed method. Our approach always achieves the best results in RMSE because it allows for gradual pixel value variation which is common in depth images.

Figures 5, 6 and 7 show the visualize qualitative results of the Middlebury data with zoomed cropped regions. From the figures, we can notice that the proposed algorithm also generates more visual appealing results than the state-of-the-art approaches. Our algorithm can preserve thin structures of the scene in regions. Edges in our results are generally smoother and sharper along the depth boundaries. Our algorithm also preserves thin structures in regions. Although FGI [29], TGV [10] and GFL0 also generate promising RMSE scores, the results of them suffer from artifacts around boundaries visually.

In our final experiment, we evaluate our method on the challenging ToFMark dataset [10] consisting of three time-of-flight (ToF) pairs, *Books*, *Shark* and *Devil*, with ground-truth depth maps. The depth maps are of size 120 × 160, and the intensity images are of size 610 × 810. This corresponds to an upsampling factor of approximately ×5. In the low-resolution depth maps, we add depth-dependent noise and backproject the remaining points to the target camera coordinate system.

We compare our results to simple bicubic interpolation and three state-of-the-art depth map super-resolution methods that utilize an additional guidance image as input. The quantitative results measured with RMSE in *mm* are presented in Table 2. Even on this difficult dataset, we are at least on par with other four classic or state-of-the-art methods for all the three test cases.

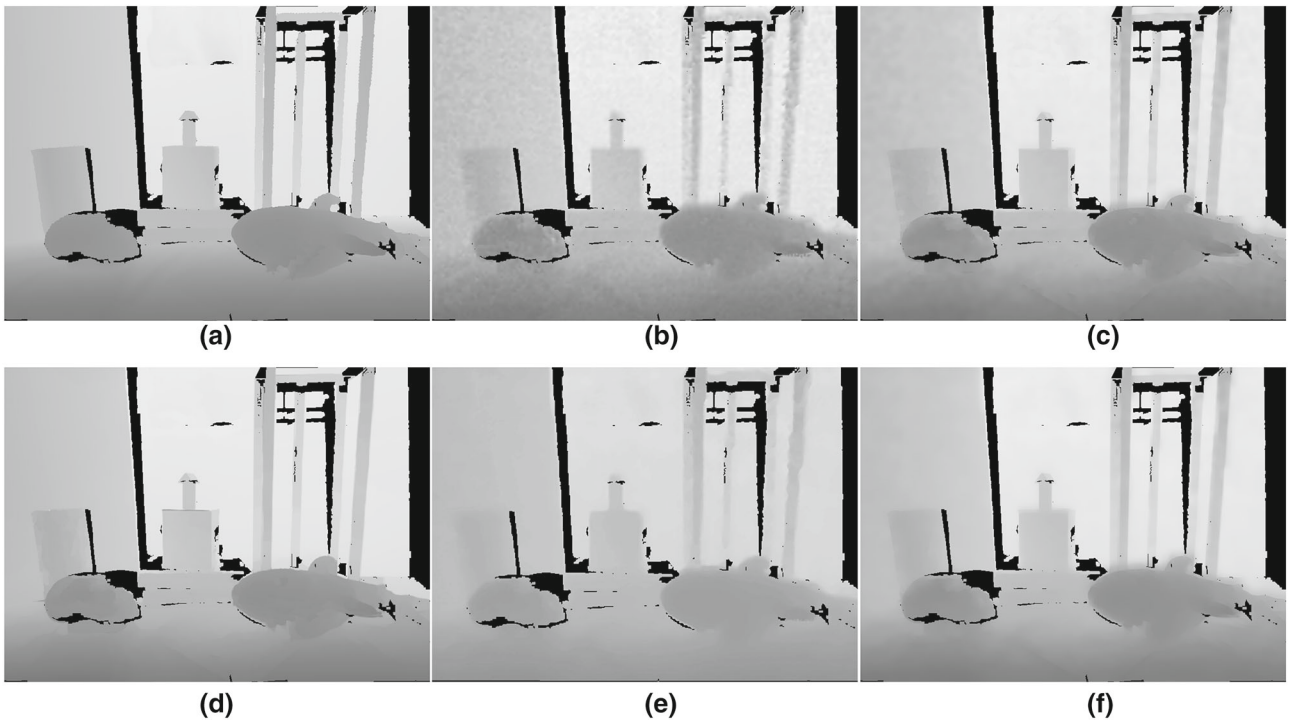


Fig. 8 Qualitative results for the ToFMark dataset sample *Shark*. **a** Ground truth, **b** Bicubic, **c** CLMF [43], **d** TGV [10], **e** MSG-Net [40], **f** our proposed method

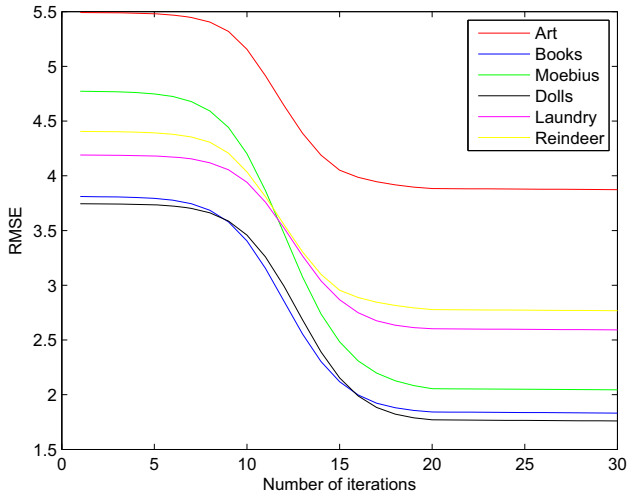


Fig. 9 Convergence illustration. The curves show how RMSE changes with iteration with upscaling factor 4. Curves of different colors represent different experimental images. The results get stable in less than 30 iterations

In Fig. 8, we show the result of upsampling the depth image *Shark*. It is observed that depth maps recovered by Bicubic and CLMF [43] still contain noises, while the results obtained by TGV [10] and our method are much clearer. By closer inspection, the TGV method in some cases introduces faulty structures in regions where the associated intensity image has rich textures, e.g., the shape of the shark's dorsal fin is deformation. MSG-Net [40] obtains a non-sharp result, and one can see that the region of the shark's teeth is blurred.

4.2 Convergence

Since the guided image filter is highly nonlinear, it is difficult to prove the global convergence of our algorithm in theory. In this work, we only provide empirical evidence to show the stability of the proposed algorithm.

In Fig. 9, we show the convergence of the proposed method for test images with upscaling factor 4. One can see that all the RMSE curves reduce monotonically with the increase in iteration number, and finally become stable and flat. One can also find that 30 iterations are typically sufficient for convergence.

4.3 Limitations

Although our method can suppress edge-blurring artifacts, sometimes over-smoothing appears in the upscaling depth images. This is because we do not have enough information to predict the high-resolution edges from the low-resolution depth input, when edges in RGB images do not correspond to those of depth images, neither Eq. (8) nor Eq. (9) can obtain sharp high-resolution depth images, and over-smoothing will appear in the results.

5 Conclusion

This work presents a new framework to recover depth maps from low-quality measurements. Based on the gradient statistics of depth images, we propose the low gradient regularization and combine it with the guided filter into the depth upsampling approach. And we present a solution to the proposed low gradient minimization problem based on threshold shrinkage. In a quantitative evaluation using widespread datasets (Middlebury and TofMark), we show that the proposed algorithm clearly performs better than the existing state-of-the-art methods in terms of RMSE. In this work, we reveal the statistical characteristics of depth image gradients, and provide a new optimization regularization method to depth image upscaling. Our method is not limited to depth image upsampling, and as a future perspective, the proposed method can be extended to process various depth image reconstructions, such as inpainting and denoising.

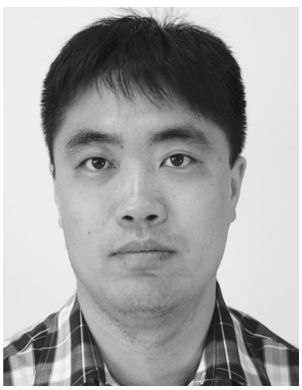
Compliance with ethical standards

Conflict of interest We declare that we have no conflict of interest.

References

1. Jalal, A., Kim, Y.: Dense depth maps-based human pose tracking and recognition in dynamic scenes using ridge data. In: IEEE International Conference on Advanced Video and Signal Based Surveillance (2014)
2. Gupta, S., Girshick, R., Arbelaez, P., Malik, J.: Learning rich features from RGB-D images for object detection and segmentation. In: European Conference on Computer Vision (ECCV) (2014)
3. Jalal, A., Kamal, S., Kim, D.: A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments. Sensors **14**(7), 11735–11759 (2014)
4. Jalal, A., Kamal, Y.H., Kim, Y.J.: Robust human activity recognition from depth video using spatiotemporal multi-fused features. Pattern Recognit. **61**, 295–308 (2017)
5. Jalal, A., Uddin, M., Kim, T.S.: Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home. IEEE Trans. Consum. Electron. **58**(3), 863–871 (2012)
6. Jalal, A., Uddin, M., Kim, J.T.: Recognition of human home activities via depth silhouettes and r transformation for smart homes. Indoor Built Environ. **21**, 184–190 (2012)
7. Riegler, G., Rüther, M., Bischof, H.: ATGV-net: accurate depth super-resolution. In: IEEE European Conference on Computer Vision (ECCV), pp. 268–284 (2016)
8. Lasang, P., Kumwilaisak, W., Liu, Y.: Optimal depth recovery using image guided tgv with depth confidence for high-quality view synthesis. J. Visual Commun. Image Represent. **39**, 24–39 (2016)

9. Hui, T.W., Chen, C.L., Tang, X.: Depth map super-resolution by deep multi-scale guidance. In: IEEE European Conference on Computer Vision (ECCV), pp. 353–369 (2016)
10. Ferstl, D., Reinbacher, C., Ranftl, R., Rüther, M., Bischof, H.: Image guided depth upsampling using anisotropic total generalized variation. In: IEEE International Conference on Computer Vision, pp. 993–1000 (2013)
11. Kwon, H., Tai, Y.W., Lin, S.: Data-driven depth map refinement via multi-scale spare representations. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
12. Yang, H., Zhang, Z., Guan, Y.: An adaptive parameter estimation for guided filter based image deconvolution. Signal Process. **138**(1), 16–26 (2017)
13. He, K., Sun, J., Tang, X.: Guided image filtering. IEEE Trans. Pattern Anal. Mach. Intell. **35**(6), 1397–1409 (2013)
14. Li, S., Kang, X., Hu, J.: Image fusion with guided filtering. IEEE Trans. Image Process. **22**(7), 2864–2875 (2013)
15. Ding, K., Wu, X., Chen, W.: Optimum inpainting for depth map based on l0 total variation. Visual Comput. **30**(12), 1311–1320 (2014)
16. Xue, H., Zhang, S., Cai, D.: Depth image inpainting: improving low rank matrix completion with low gradient regularization. IEEE Trans. Image Process. **26**(9), 4311–4320 (2017)
17. Xu, L., Lu, C., Xu, Y.: Image smoothing via l0, gradient minimization. In: SIGGRAPH Asia Conference, ACM, p. 174. (2011)
18. Nguyen, R.M.H., Brown, M.S.: Fast and effective l0 gradient minimization by region fusion. In: IEEE International Conference on Computer Vision (ICCV), pp. 208–216 (2015)
19. Yang, J., Wright, J., Huang, T.S.: Image super-resolution via sparse representation. IEEE Trans. Image Process. **19**(11), 2861–2873 (2010)
20. Timofte, R., De, V., Gool, L.V.: Anchored neighborhood regression for fast example-based super-resolution. In: IEEE International Conference on Computer Vision, pp. 1920–1927 (2013)
21. Li, Y., Xue, T., Sun, L.: Joint example-based depth map super-resolution. In: IEEE International Conference on Multimedia and Expo, pp. 152–157 (2012)
22. Ferstl, D., Rüther, M., Bischof, H.: Variational depth super-resolution using example-based edge representations. In: IEEE International Conference on Computer Vision, pp. 513–521 (2015)
23. Schulter, S., Leistner, C., Bischof, H.: Fast and accurate image upscaling with super-resolution forests. In: IEEE Computer Vision and Pattern Recognition, pp. 3791–3799 (2015)
24. Mahmoudi, M., Sapiro, G.: Sparse representations for range data restoration. IEEE Trans. Image Process. **21**(5), 2909–2915 (2012)
25. Kopf, J., Cohen, M.F., Lischinski, D.: Joint bilateral upsampling. ACM Trans. Graph. **26**(3), 96 (2011)
26. Yang, Q., Yang, R., Davis, J.: Spatial-depth super resolution for range images. In: IEEE Computer Vision and Pattern Recognition, pp. 1–8 (2011)
27. Liu, M.Y., Tuzel, O., Taguchi, Y.: Joint geodesic upsampling of depth images. In: IEEE Computer Vision and Pattern Recognition, pp. 169–176 (2013)
28. Lu, J., Forsyth, D.: Sparse depth super resolution. In: IEEE Computer Vision and Pattern Recognition, pp. 2245–2253 (2015)
29. Li, Y., Min, D., Do, M.N.: Fast guided global interpolation for depth and motion. In: European Conference on Computer Vision, pp. 717–733 (2016)
30. Jung, C., Yu, S., Kim, J.: Intensity-guided edge-preserving depth upsampling through weighted l0 gradient minimization. J. Visual Commun. Image Represent. **42**, 132–144 (2017)
31. Diebel, J., Thrun, S.: An application of markov random fields to range sensing. Adv. Neural Inf. Process. Syst. **2006**, 291–298 (2006)
32. Park, J., Kim, H., Tai, Y.W.: High quality depth map upsampling for 3d-tof cameras. In: IEEE International Conference on Computer Vision (ICCV), pp. 1623–1630 (2011)
33. Aodha, M., Campbell, N.D.F., Nair, A.: Patch based synthesis for single depth image super-resolution. In: European Conference on Computer Vision (ECCV), pp. 71–84 (2012)
34. Yang, J., Ye, X., Li, K.: Color-guided depth recovery from rgb-d data using an adaptive autoregressive model. IEEE Trans. Image Process. **23**(8), 3443–3458 (2014)
35. Dong, C., Chen, C.L., He, K.: Learning a deep convolutional network for image super-resolution. In: European Conference on Computer Vision (ECCV), pp. 184–199 (2014)
36. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1646–1654 (2016)
37. Dong, C., Chen, C.L., He, K.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2014)
38. Xie, J., Feris, R.S., Sun, M.T.: Edge-guided single depth image super resolution. IEEE Trans. Image Process. **25**(1), 428–438 (2016)
39. Handa, A., Patraucean, V., Badrinarayanan, V.: Understanding real world indoor scenes with synthetic data. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4077–4085 (2016)
40. Hui, T.W., Loy, C.C., Tang, X.: Depth map super-resolution by deep multi-scale guidance. In: European Conference on Computer Vision (ECCV), pp. 353–369 (2014)
41. Storath, M., Weinmann, A., Demaret, L.: Jump-sparse and sparse recovery using potts functionals. IEEE Trans. Signal Process. **62**(14), 3654–3666 (2014)
42. Wang, Y., Yang, Y., Yin, W.: A new alternating minimization algorithm for total variation image reconstruction. SIAM J. Imaging Sci. **1**(3), 248–272 (2008)
43. Lu, J., Shi, K., Min, D., Lin, L., Do, M.N.: Cross-based local multipoint filtering. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 430–437 (2012)



Hang Yang received his B.S. and Ph.D. degrees in mathematics from the Jilin University in 2007 and 2012, respectively. He is currently an associate research fellow at the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science. His current research interests include image deblurring, super-resolution and visual tracking.

Zhongbo Zhang received his M.S and Ph.D. degrees in mathematics from the Jilin University in 2000 and 2003, respectively. He is currently an associate professor at the School of Mathematics, Jilin University, China. His current research interests include image denoising, image deconvolution and pattern recognition.