Journal of
Electronic Imaging

JElectronicImaging.org

Integrated cosparse analysis model with explicit edge inconsistency measurement for guided depth map upsampling

Yifan Zuo
Qiang Wu
Ping An
Xiwu Shang

SPIE. IS&T imaging.org

Yifan Zuo, Qiang Wu, Ping An, Xiwu Shang, "Integrated cosparse analysis model with explicit edge inconsistency measurement for guided depth map upsampling," *J. Electron. Imaging* **27**(4), 043004 (2018), doi: 10.1117/1.JEI.27.4.043004.

Journal of Electronic Imaging 27(4), 043004 (Jul/Aug 2018)

Integrated cosparse analysis model with explicit edge inconsistency measurement for guided depth map upsampling

Yifan Zuo,[a,b] Qiang Wu,[b] Ping An,[a,*] and Xiwu Shang[a]
[a]Shanghai University, School of Communication and Information Engineering, Shanghai, China
[b]University of Technology Sydney, Sydney, New South Wales, Australia

Abstract. A low-resolution depth map can be upsampled through the guidance from the registered high-resolution color image. This type of method is so-called guided depth map upsampling. Among the existing methods based on Markov random field (MRF), either data-driven or model-based prior is adopted to construct the regularization term. The data-driven prior can implicitly reveal the relation between color-depth image pair by training on external data. The model-based prior provides the anisotropic smoothness constraint guided by high-resolution color image. These types of priors can complement each other to solve the ambiguity in guided depth map upsampling. An MRF-based approach is proposed that takes both of them into account to regularize the depth map. Based on analysis sparse coding, the data-driven prior is defined by joint cosparsity on the vectors transformed from color-depth patches using the pair of learned operators. It is based on the assumption that the cosupports of such bimodal image structures computed by the operators are aligned. The edge inconsistency measurement is explicitly calculated, which is embedded into the model-based prior. It can significantly mitigate texture-copying artifacts. The experimental results on Middlebury datasets demonstrate the validity of the proposed method that outperforms seven state-of-the-art approaches. [DOI: 10.1117/1.JEI.27.4.043004]

Keywords: depth map upsampling; depth map enhancement; analysis sparse coding; Markov random field.

Paper 180129 received Feb. 7, 2018; accepted for publication Jun. 15, 2018; published online Jul. 5, 2018.

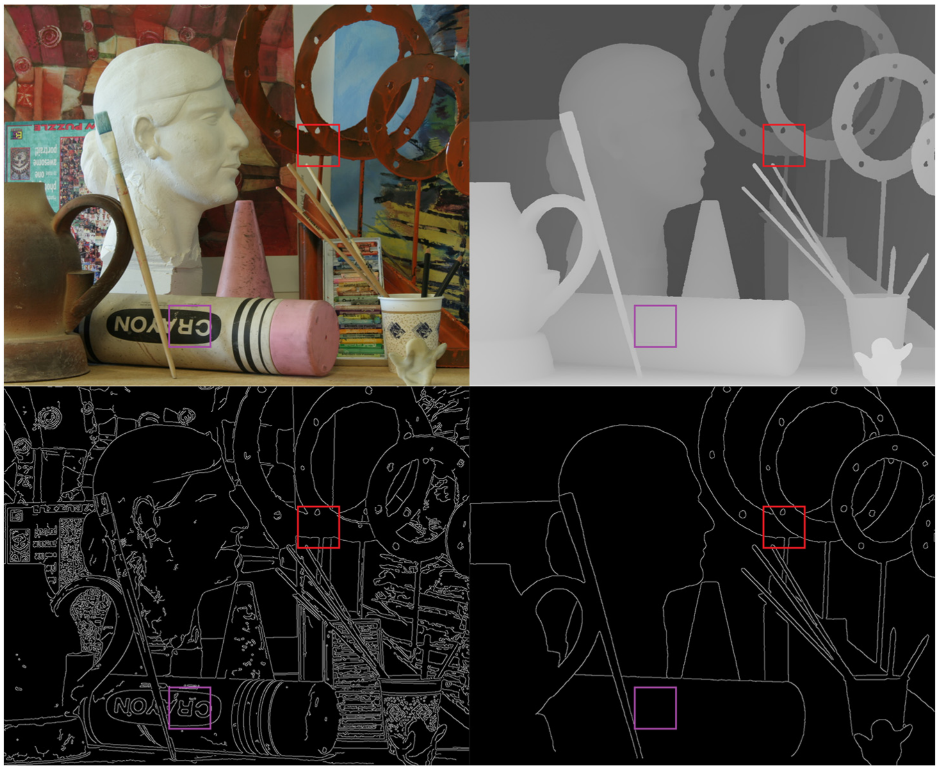
1 Introduction

Depth information is widely used in many advanced applications, e.g., image-based rendering, robot navigation, and three-dimensional (3-D) reconstruction. However, compared with shooting conventional color image, high-quality depth map acquisition is more difficult. The methods for shooting depth map are classified into two types, such as passive and active. The passive methods estimate the depth map through stereomatching that is to find correspondences among the color images in different views. In spite of developments in decades, the difficulty of estimating the depth values in texture-less and occlusion regions is still open. Alternatively, as the cost of sensors is reduced, the active methods that directly capture depth map using such sensors receive more and more attention. Laser scanner[1] normally is the best choice to obtain a highly accurate depth map for static scenes. For dynamic scenes, ToF sensors (e.g., Swissranger 4500, Kinect V2) obtain real-time depth sequences by measuring the phase difference between the emitted light and the reflected light.[2] However, these depth maps are noisy in low resolutions (e.g., 176×144, 200×200). So, the quality of the raw depth maps captured by such sensors cannot meet the requirements in real applications. Typically, the high-resolution (HR) color image that is the description of the same scene as the low-resolution (LR) depth map is always used to guide LR depth map upsampling. This type of method is so-called guided depth map upsampling.

Many approaches for guided depth map upsampling are proposed. The early works upsample the LR depth map by assuming that the depth and the color edges on the corresponding locations are consistent. The relation between neighboring pixels on the HR color image can provide the prior for the corresponding pixels on the depth map that is to be upsampled. Such prior is model-based without training on external data. Nevertheless, the edge-consistent assumption is not always true leading to texture-copying artifacts and blurring depth edges. Specially, the texture-copying artifacts are caused by the situation that the smooth depth regions correspond to the color regions with rich texture. By contrast, the blurring depth edges are normally observed in the case that the smooth color regions correspond to the depth regions with edges. Figure 1 shows an illustration for the two cases of edge inconsistency explained above. The edge maps are computed by Canny operator.[3] For decades, there have been several works trying to solve these challenges.[4–8] They balance the contribution from the original depth map and the companion color image. Garcia et al.[4] proposed a credibility map to identify the smooth regions on the depth map where texture copying might occur. The credibility map is defined as a Gaussian function that is related to the gradient of the LR depth map. However, this method may blur depth edges when there are no corresponding edges on the color image. In addition, the credibility map only uses the local feature on the LR depth map, which cannot precisely evaluate the edge inconsistency between the color image and the depth map. Lo et al.[6] presented a framework for depth

*Address all correspondence to: Ping An, E-mail: anping@shu.edu.cn

1017-9909/2018/$25.00 © 2018 SPIE and IS&T



(a) (b)

Fig. 1 An illustration of edge inconsistency: (a) the color image and its edge map and (b) the depth map and its edge map. (Pink window: the edges occur on the color image but not on the depth map, red window: the edges occur on the depth map but not on the color image.)

map SR problem by solving a Markov random field (MRF) labeling optimization problem, which considers preserving depth edges and suppressing texture-copying artifacts. The main constraints are twofold. (1) Reference 6 only uses the range information of local patches on the depth map to classify smooth and unsmooth regions. Depending on the classification above, different weight schemes are carried out. As the local structure is used, the improvement is limited. (2) Reference 6 only detects the edges on the depth map and it does not explicitly measure inconsistency between the color edge map and the depth edge map. Therefore, the guidance may not be correct in certain situations. Xiang et al.[8] proposed a method that validates the pair of edges on the depth map and the corresponding color image. Such validation is only based on the local structure. An empirical threshold is adopted to filter unreliable edge pairs. Overall, the common drawback of these methods is that they do not explicitly evaluate the edge inconsistency between the color image and the corresponding depth map by taking local and global structure into account. Therefore, they cannot adaptively control the efforts of the guidance from the HR color image when enhancing the LR depth map. Our previous work[9] for the first time explicitly calculates the inconsistency between the depth edges and the corresponding color edges through global optimization. Although the performance of this type of method is significantly improved, the improvement is limited because only the handcraft features (e.g., edge, segment) are used.

Recently, some works were proposed to learn the relation between the HR color image and the LR depth map by training on external data. Such relation is implicitly learned through joint sparse coding. There are two types of sparse coding, which are based on synthesis and analysis models. The synthesis model assumes that the signal $\mathbf{X}$ can be represented by sparse atoms $\{\mathbf{u_i}, \dots, \mathbf{u_j}\}$ that are columns of learned dictionary $\mathbf{U}$. The sparse coefficient $\boldsymbol{\alpha}$ is to identify these atoms with a sparsity constraint, which is always defined by L1 norm. That is, $\mathbf{X} = \mathbf{U}\boldsymbol{\alpha}$ s.t. $g(\boldsymbol{\alpha}) < \tau$ in which g is a function for evaluating sparsity and τ is a threshold. Therefore, the signal is represented by the elements that are nonzero in α. In joint synthesis sparse coding, a pair of dictionaries is learned for the depth map and color image, respectively. The relation between them is represented by sharing the sparse coefficient α. In the analysis perspective, the original signal $\mathbf{X}$ is transformed by a learned operator $\boldsymbol{\Omega}$, which is called analyzed signal. The analyzed signal $\boldsymbol{\Omega}\mathbf{X}$ is assumed to be sparse [i.e., $g(\boldsymbol{\Omega}\mathbf{X}) < \tau$]. Unlike the situation of the synthesis model, the signal is represented by the elements that are zero in $\boldsymbol{\Omega}\mathbf{X}$ in the analysis counterpart. So, the cosparsity is introduced to distinguish this situation from that in the synthesis model.[10] Similarly, a pair of operators is learned for color–depth image pair in a joint analysis sparse coding. The relation between them can be represented by constraining the locations of "zero" on the analyzed signals to be the same. As the depth map is indirectly reconstructed in a patchy manner using the optimized sparse coefficient α in the reconstruction phase of the synthesis model, the consistent reconstructed depth values cannot be guaranteed on the overlapping regions of the adjacent patches. Therefore, the final result is more likely to be over-smooth by averaging them. Compared with the synthesis model, the analysis counterpart can directly reconstruct the depth map. That is, the reconstructed depth values are directly determined by the solution to the optimization problem of reconstruction. Overall, both synthesis and analysis models provide a data-driven way to regularize the depth map. But the edge inconsistency is not explicitly considered in this type of methods due to the implicit relation learning. The texture-copying artifacts and blurring depth edges can be observed on their results.

Based on the analysis above, this paper proposes an approach that combines model-based prior (i.e., a more accuracy model to represent the property of edge cooccurrence on color–depth image pair) with the data-driven one to better mitigate texture-copying artifacts and restore more details for upsampling the LR depth map. Compared with our previous work,[9] the main contribution of this article is to consider model-based and data-driven priors as regularizations on the depth map upsampling. Such two parts complement each other, which are not considered in Ref. 9. This is investigated in Sec. 4.1. In addition, Sec. 4.2 shows the improvement compared with other state-of-the-art methods. The rest of the paper is organized as follows: Sec. 2 reviews the related works and motivation. The proposed method is explained in Sec. 3. Section 4 shows the experimental results and corresponding analyses. Section 5 concludes this paper.

2 Related Works and Motivation

The existing methods for guided depth map upsampling can be classified into filter-based, optimization-based, and learning-based categories. First, the related works of each category are reviewed. Then, the motivation of the proposed method is presented.

2.1 *Filter-Based Methods*

The pioneer work of filter-based methods is a joint bilateral upsampling (JBU) framework proposed by Kopf et al.[11] In this approach, bilateral filtering is used to refine the edges of the LR depth map according to the ones of the registered HR color image. Yang et al.[12] iteratively refined the depth maps via the cost volume computed by JBU with depth candidates. He et al.[13] proposed a guided filtering that models a linear relationship between the output and the guidance image. It is based on the assumption that the output has an edge only if the guidance image has one. Garcia et al.[14] reduced the weight of unreliable depth samples by analyzing the gradient of LR depth values. Min et al.[15] proposed a weighted mode filtering method (WMF) based on joint histogram of the depth candidates, which enforces the result to satisfy the requirement of L1 norm minimization. Liu et al.[16] proposed a variant of JBU, which computes weights for average based on geodesic. The geodesic is a joint space of color and distance instead of separating spaces in JBU. Lo et al.[17] extended JBU to joint trilateral filtering (JTF) by considering local gradient information of the LR depth map. Hua et al.[18] proposed an approach that adopts onion-peeling filtering procedure to exploit local gradient information of the depth map. Recently, Barron and Poole[19] presented a bilateral solver for edge-preserving smoothing, which combines the flexibility and speed of simple filtering approaches with the accuracy of domain-specific optimization algorithms.

2.2 *Optimization-Based Methods*

By comparison with the filter-based methods, the optimization-based ones are more robust to noise. Diebel and Thrun[20] modeled depth map upsampling as solving a multilabeling optimization via MRF framework. Zhu et al.[21] updated the traditional spatial MRF to the dynamic MRF, which introduces both spatial and temporal information in a MRF energy function. It improves the accuracy and the robustness of depth map upsampling for dynamic scenes. Lu et al.[22] designed a data term in MRF energy function that can better fit to the characteristics of depth maps. Ferstl et al.[23] designed a second-order total generalized variation smoothness constraint to regularize the depth map and use an anisotropic diffusion tensor computed from the registered HR color image to guide depth map upsampling. Lo et al.[6] defined the data term and the regularization term based on sigmoid function. Park et al.[24] proposed a nonlocal regularization term and combined it with a weight scheme involving the edge, the gradient, and the segmentation guidance extracted from the HR color images. Yang et al.[5] proposed a color-guided autoregression model (AR) to upsample depth map. Yang et al.[25] proposed an optimization framework for guided depth map upsampling, which takes local, nonlocal regularization term on the LR depth map and the features of HR color image into account. Li et al.[26] proposed a hierarchical global optimization framework that is based on fast weighted least squares solver.[27] Liu et al.[28] adopted a robust M-estimator to define the regularization term that implicitly handles the edge inconsistency between the depth map and the registered color image. The previous work[9] proposed a quantitative measurement on edge inconsistency between the registered color image and the depth map. Then such inconsistency measurement is explicitly embedded into the MRF energy function.

2.3 *Learning-Based Methods*

In addition to the methods in filter-based and optimized-based categories that only use model-based prior, some approaches based on data-driven prior are proposed. As a pioneer work, Li et al.[29] introduced synthesis sparse coding model to guide depth upsampling, which jointly trains three dictionaries for registered patches of the LR depth maps, the HR depth maps, and the HR color images. In the reconstruction phase, the depth maps are reconstructed in a patchy manner using the sparse representation on the learned dictionary. Kwon et al.[30] proposed a multiscale learning scheme to train three dictionaries of the HR depth patches, the LR depth patches, and the HR color patches. Furthermore, it explicitly constrains consistent reconstruction between the overlapping patches in the object function. Unlike the methods above which use synthesis model, Kiechle et al.[31] used analysis model to exploit the cosparsity of the transformed depth–color image pair using jointly learned operators, the HR depth map is reconstructed through data fidelity and color-guided sparsity constraint.

2.4 *Motivation*

Based on the analysis above, sparse coding models can learn data-driven prior to regularize the HR depth map. However, such data-driven prior does not explicitly consider the edge inconsistency between the color image and the depth map. Some texture-copying artifacts and blurring depth edges can be observed on the upsampled depth map. In addition, the performance of learning-based models is significantly affected by the training data. Compared with such data-driven prior, the model-based prior (e.g., the anisotropic L2 norm used in Refs. 9 and 28) explicitly handles the edge inconsistency above.

In this paper, an MRF-based method for guided depth map upsampling is proposed which combines the data-driven prior with the model-based ones. As synthesis sparse coding model only optimizes the sparse coefficient, the proposed method adopts the analysis model that can directly optimize HR depth map. To handle the edge inconsistency between the color image and the depth map, motivated by the previous work,[9] the model-based prior is embedded with explicit calculation of the edge inconsistency by taking global and local features into account. It can significantly mitigate texture-copying artifacts and better preserve depth edges.

3 Proposed Method

This paper proposes a method for guided depth map upsampling through solving a MRF optimization problem, which is defined as

$$\mathbf{D}^{\star} = \arg\min_{\mathbf{D}} E_{\text{data}}(\mathbf{D}_{\text{sub}}, \mathbf{D}^{\mathbf{o}}) + \lambda E_{\text{regl}}(\mathbf{D}) + \gamma E_{\text{regp}}(\mathbf{D}), \tag{1}$$

where $\mathbf{D}$ is the HR depth map, E_{data} is the data term that maintains the consistency between the reconstructed depth values and the initial observed ones. $\mathbf{D}^{\mathbf{o}}$ is the set of the observed depth values. $\mathbf{D}_{\text{sub}}$ is the subset of $\mathbf{D}$ sharing the same locations as $\mathbf{D}^{\mathbf{o}}$. E_{regl} and E_{regp} form the regularization term, which are data-driven and model-based priors respectively. In fact, the data term E_{data} is related to maximum likelihood estimation (MLE). The model-based and data-driven priors (i.e., E_{regp} and E_{regl}) guided by high-resolution color image provide regularizations on the distribution of the depth map itself, which updates MLE to maximum *a posteriori* estimation (MAP). Therefore, the MRF optimization is equivalent to MAP. Figure 2 shows the block diagram representation of the proposed method. As shown in Fig. 2, the edge inconsistency measurement is applied to construct the model-based prior. In addition, the data-driven prior is learned from external datasets. Both of them are used to regularize the depth map. Finally, the high-quality depth map is reconstructed via solving the MRF optimization. The data term is defined as Eq. (2). E_{regl} and E_{regp} are explained in the following equation:

$$E_{\text{data}}(\mathbf{D}_{\text{sub}}, \mathbf{D}^{\mathbf{o}}) = \|\mathbf{D}_{\text{sub}} - \mathbf{D}^{\mathbf{o}}\|_F. \tag{2}$$

3.1 *Learned Prior*

We adopt the analysis model of sparse coding to learn the relation between the color image $\mathbf{I}$ and the depth map $\mathbf{D}$. There are many functions that can be used to evaluate the sparsity of signals. In this paper, the sparsity of signal $\mathbf{X} \in \mathbf{R}^{\mathbf{n}}$ is evaluated by the following equation:

$$g(\mathbf{X}) = \sum_{i=1}^{n} \log(1 + 400x_i^2), \tag{3}$$

where x_i is the i'th element of $\mathbf{X}$. We generate a pair of local patches for each pixel on the HR depth map $\mathbf{D}$ and the HR color image $\mathbf{I}$, respectively. The patch size is 5×5. Such relation is exploited by assuming that the j'th vectorized color and depth patches ($\mathbf{S}_{\mathbf{I}}^{\mathbf{j}} \in \mathbf{R}^{\mathbf{m}}, \mathbf{S}_{\mathbf{D}}^{\mathbf{j}} \in \mathbf{R}^{\mathbf{m}}$) transformed by a learned pair of operators ($\mathbf{\Omega}_{\boldsymbol{I}} \in \mathbf{R}^{\mathbf{n} \times \mathbf{m}}, \mathbf{\Omega}_{\boldsymbol{D}} \in \mathbf{R}^{\mathbf{n} \times \mathbf{m}}$) allow a cosparse representation. Such operators are trained according to the method proposed by Kiechle et al.[31] That is, the positions of the zero elements are shared in the pair of the analyzed vectors $\mathbf{\Omega}_{\boldsymbol{I}}\mathbf{S}_{\mathbf{I}}^{\mathbf{j}}$, $\mathbf{\Omega}_{\boldsymbol{D}}\mathbf{S}_{\mathbf{D}}^{\mathbf{j}}$. However, this constraint is ideal, the sparsity evaluation function g [i.e., Eq. (3)] is redefined as Eq. (4) in a joint analysis

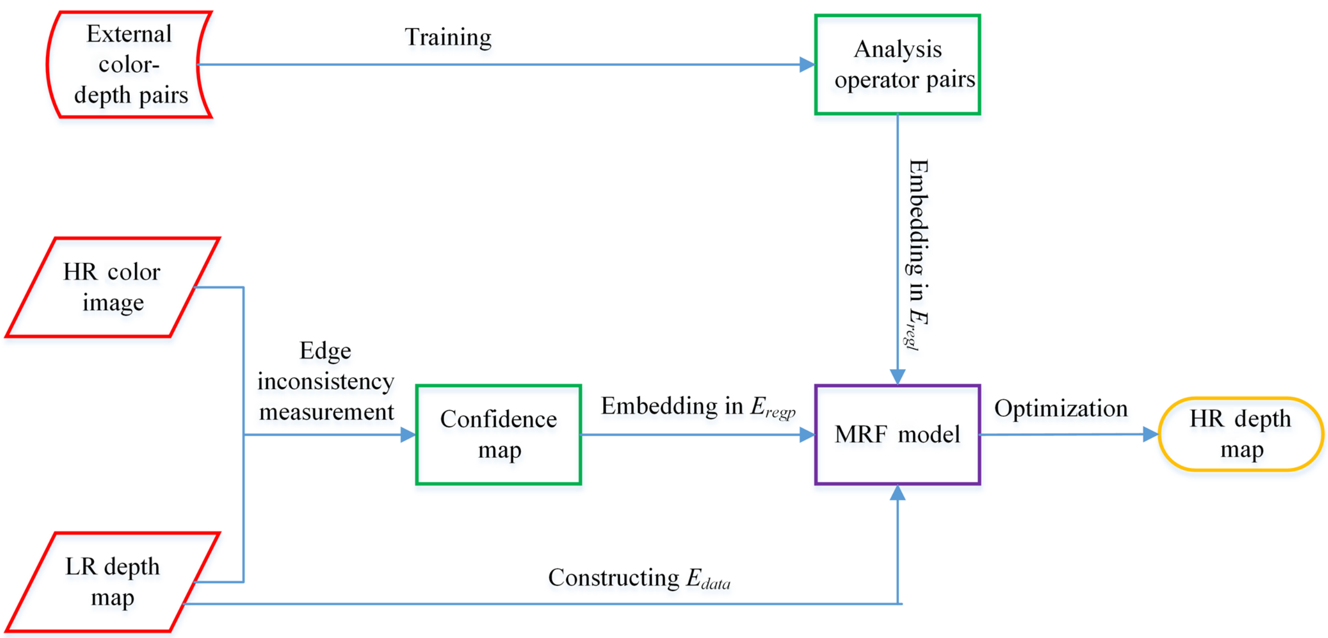


Fig. 2 The block diagram representation of the proposed method.

sparse coding to embed such cosparse constraint in real situations

$$g(\mathbf{\Omega}_I\mathbf{S}_\mathbf{I}^\mathbf{j}, \mathbf{\Omega}_D\mathbf{S}_\mathbf{D}^\mathbf{j}) = \sum_{i=1}^{n} \log\{1 + 400[(\mathbf{\Omega}_\mathbf{I}\mathbf{S}_\mathbf{I}^\mathbf{j})_i^2 + (\mathbf{\Omega}_\mathbf{D}\mathbf{S}_\mathbf{D}^\mathbf{j})_i^2]\}. \tag{4}$$

All of such patch-based constraints are considered to regularize the whole depth map. Therefore, the data-driven regularization term E_{regl} is defined as

$$E_{\text{regl}}(\mathbf{D}) = \sum_{j=1}^{M} g(\mathbf{\Omega}_I\mathbf{S}_\mathbf{I}^\mathbf{j}, \mathbf{\Omega}_D\mathbf{S}_\mathbf{D}^\mathbf{j}), \qquad \mathbf{S}_\mathbf{D}^\mathbf{j} \in \mathbf{D}, \tag{5}$$

where M is the total number of pixels on the HR depth map.

3.2 *Predefined Prior*

The model-based prior $E_{\text{regp}}(\mathbf{D})$ is usually defined as[9,20,22,24]

$$E_{\text{regp}}(\mathbf{D}) = \sum_{p} \sum_{q\in\mathbf{N_p}} \boldsymbol{\phi}(p,q)(d_p - d_q)^2, \tag{6}$$

where $\mathbf{N_p}$ is the eight-connected neighborhood of pixel p. d_p, d_q are elements in $\mathbf{D}$, which represent the depth values of pixels p, q. $\boldsymbol{\phi}$ is the set of anisotropic affinities, which includes the guidance from the color image. The following part first explains the edge inconsistency measurement between the color image and the depth map. Then, $\boldsymbol{\phi}$ is constructed by embedding such explicit measurement to mitigate texture-copying artifacts and preserve depth edges.

To detect two cases of inconsistency between the HR color image and the LR depth map, which are shown in Fig. 1, the LR depth map is coarsely interpolated to the same resolution as the HR color image followed by a bidirection evaluation between the edge maps of them. These edge maps are obtained by a Canny operator. Along each direction, it presents an evaluation for the target edge map tar against the reference edge map ref. The ref/tar means the color/depth edge map along one direction, and the depth/color edge map along the other direction.

As the quality of LR depth map is low, the edge pixels on the depth edge map may shift from the real positions. Therefore, they are not strictly aligned with the ones on the color edge map, even if these edge pixels are consistent. However, it is assumed that the displacement between the consistent edge pixels from ref to tar should be constrained in a small range. So, the proposed method searches the optimal correspondence of each edge pixel on ref within a local region on tar. Specifically, the local regions are set to 5×5, 7×7, and 11×11 for the upsampling factors of 2×, 4×, and 8×, respectively. In addition, the strength and the orientation of the displacements of the matched edge pixels in a nearby region should be smooth. Based on these assumptions, the best matching pairs of edge pixels can be determined through an MRF optimization as

$$\mathbf{L}^\star = \arg\min_{\mathbf{L}=\{l_p\}} \sum_{p\in\text{ref}} C(p, p + l_p) + \mu \sum_{p\in\text{ref}} \sum_{q\in\mathbf{N_p}} R(l_p, l_q), \tag{7}$$

where $\mathbf{L}$ is the set of displacements between corresponding edge pixels on the depth map and the color image. μ is a balance factor between the data term C and the regularization term R. p means the position of an edge pixel on ref. $\mathbf{N_p}$ represents the set of eight-connected neighboring pixels of p. l_p stands for the displacement for p, which is an element

Table 1 Quantitative comparison of components (in MAE).

Datasets	Cones			Teddy			Tsukuba			Venus		
Variants	2×	4×	8×	2×	4×	8×	2×	4×	8×	2×	4×	8×
E_{regl}	0.49	1.10	2.16	0.37	0.77	1.54	0.87	1.72	3.53	0.18	0.39	0.91
E_{regp}	0.43	0.79	1.65	0.35	0.70	1.32	0.71	1.25	2.40	0.10	0.18	0.44
$E_{\text{regl}} + E_{\text{regp}}$	0.39	0.73	1.52	0.31	0.64	1.21	0.70	1.20	2.20	0.09	0.18	0.36

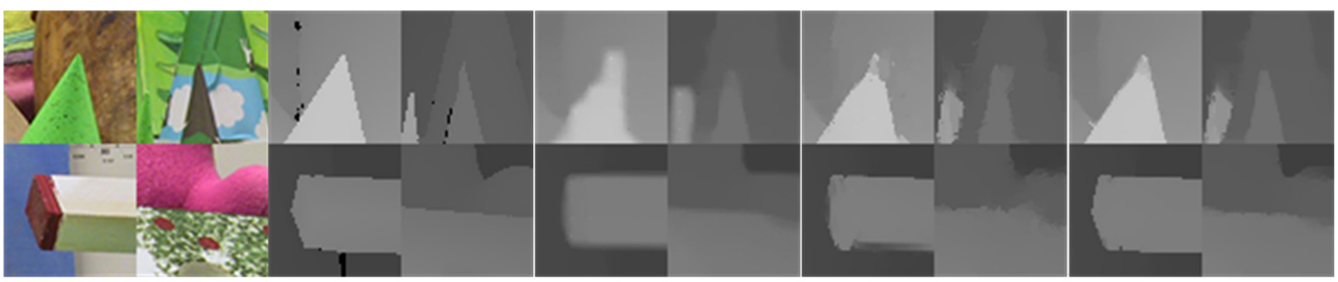

(a) (b) (c) (d) (e)

Fig. 3 The comparison between the variants of the proposed method: (a) the color patches, (b) the ground truth depth patches, the LR depth patches are upsampled (8×) by (c) E_{regl} only, (d) E_{regp} only, and (e) $E_{\text{regl}} + E_{\text{regp}}$.

of $\mathbf{L}$. So, $p + l_p$ represents the position of the edge pixel k on tar.

As minimum weighted bipartite matching[32] considers local structure and constrains each edge pixel on the reference patch to match at most one edge pixel on the target patch, it is a more robust approach than mean of absolute error (MAE). So, the proposed method constructs the data term C based on minimum weighted bipartite matching. The regularization term R implies global structure. The details are explained as below.

Table 2 Quantitative comparison (in MAE) on Middlebury datasets.

Datasets	Cones			Teddy			Tsukuba			Venus		
Methods	2×	4×	8×	2×	4×	8×	2×	4×	8×	2×	4×	8×
Bicubic	0.95	1.81	4.36	0.91	1.40	2.35	1.41	2.31	4.80	0.34	0.62	1.19
NLMR[24]	0.87	1.78	3.23	0.73	1.38	2.65	1.45	2.48	4.32	0.27	0.53	1.14
TGV[23]	0.90	1.07	1.91	0.75	0.89	1.53	1.38	1.75	3.42	0.16	0.27	**0.35**
WMF[15]	0.87	1.48	3.17	0.85	1.23	1.87	1.07	1.57	3.39	0.31	0.52	0.89
JBUV[12]	1.23	1.78	3.80	1.28	1.50	2.13	2.14	2.42	4.37	0.55	0.66	1.07
SUSR[38]	0.88	1.37	N/A	0.70	1.14	N/A	1.53	2.68	N/A	0.30	0.45	N/A
VSER[35]	0.51	1.05	N/A	0.38	0.75	N/A	0.95	1.87	N/A	0.12	0.28	N/A
Ours	**0.39**	**0.73**	**1.52**	**0.31**	**0.64**	**1.21**	**0.70**	**1.20**	**2.20**	**0.09**	**0.18**	0.36

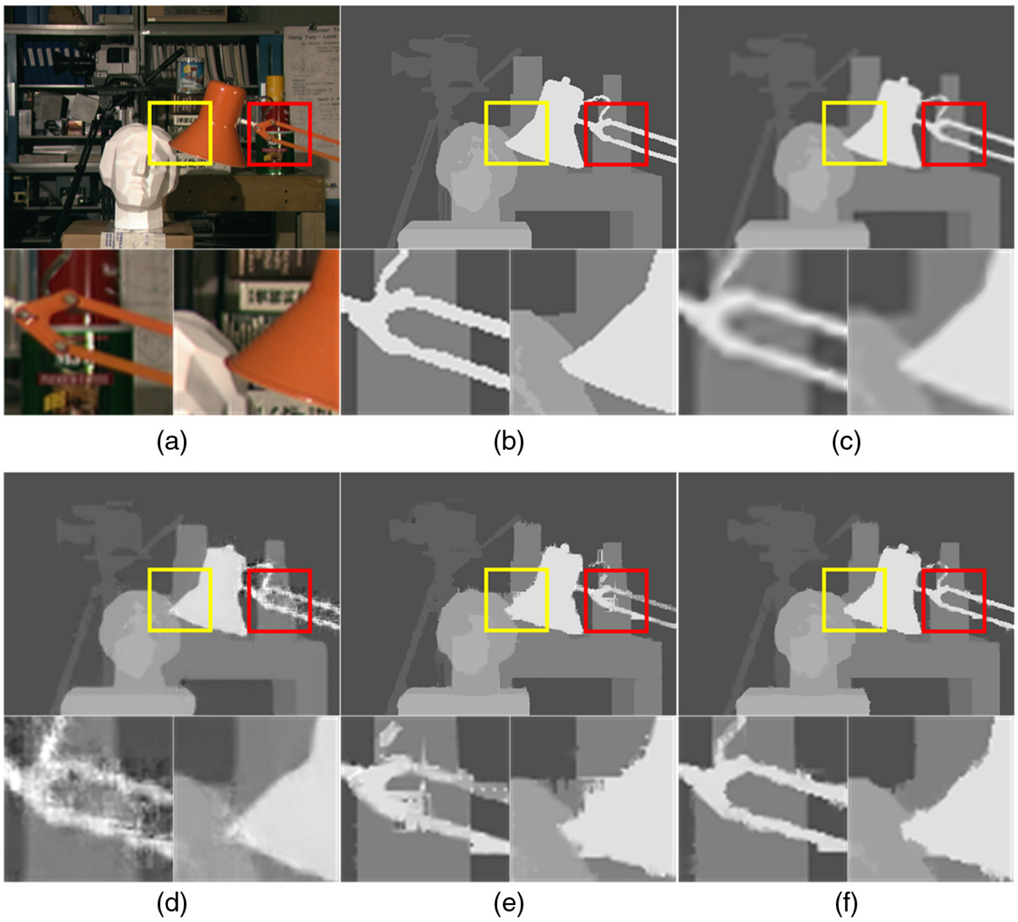


Fig. 4 The visual results of "Tsukuba"; (a) the color image, (b) the ground truth depth map, the LR depth maps are upsampled (4×) by (c) VSER,[35] (d) SUSR,[38] (e) TGV,[23] and (f) the proposed method.

By given p, if the target pixel k is not an edge pixel, it is treated as definite inconsistency in which $C(p,k)$ is assigned to the maximum inconsistency value. It is set to 1 in this work. Otherwise, $C(p,k)$ is defined by minimum weighted bipartite matching[32] between two patches centered by p and k. The patch size is 3×3. Specially, the bipartite graph $\mathbf{G}(\mathbf{V^p},\mathbf{V^k},\mathbf{E},\mathbf{W})$ is defined as follows: $\mathbf{V^p}=\{ep_1,ep_2,\ldots,ep_M\}$ and $\mathbf{V^q}=\{ek_1,ek_2,\ldots,ek_N\}$ are vertices representing the sets of edge pixels in the two patches (excluding the centers p and k). $\mathbf{E}$ is the graph edges linking vertices. As the edge pixels of matched pair in $\mathbf{V^p}$ and $\mathbf{V^q}$ are more likely to be near with each other, the weight set $\mathbf{W}$ of $\mathbf{E}$ is defined as

$$\kappa(ep_i,ek_j)=f(|ep_i^x-ek_j^x|+|ep_i^y-ek_j^y|), \tag{8}$$

where $f(0)=0$, $f(1)=1$, $f(2)=1.6$ and $f(x)=2$ when $x>2$. ep_i^x and ep_i^y are the coordinates of the edge pixel ep_i.

For robust matching, the difference in the total number of edge pixels between $\mathbf{V^p}$ and $\mathbf{V^q}$ should be considered as well. In this work, the data term $C(p,k)$ computed by minimum weighted bipartite matching is defined as Eq. (9). The range of it is [0, 1]

$$C(p,k)=\sum_{\substack{(mp_s,mk_s)\\ \in \mathbf{V^{pk'}}}}[\kappa(mp_s,mk_s)/2+|M-N|]/8 \tag{9}$$

where $\mathbf{V^{pk'}}=\{(mp_1,mk_1),(mp_2,mk_2),\ldots,(mp_r,mk_r)\}$ is the set of edge pixel pairs determined by minimum weighted bipartite matching. M,N are the total edge pixels in $\mathbf{V^p}$ and $\mathbf{V^q}$, respectively.

$R(l_p,l_q)$ is the regularization term in Eq. (7), which penalizes the case that the adjacent edge pixels have different displacements. It is defined as follows:

$$R(l_p,l_q)=\begin{cases}0, & l_p=l_q\\ 1, & l_p\neq l_q\end{cases}. \tag{10}$$

Graph cut[33] is adopted to solve the discrete MRF optimization problem [i.e., Eq. (7)]. The inconsistency measurement for each edge pixel p is represented by $C(p,p+l_p^\star)$, which can be computed by the optimized displacement $l_p^\star\in\mathbf{L}^\star$. $\mathbf{C_{color}}$ (the color edge map is regarded as the reference edge map) and $\mathbf{C_{depth}}$ (the depth edge map is regarded as the reference edge map) are the sets of the inconsistency measurement for all edge pixels. As the positions of edge pixels on the depth edge map are inaccurate, $\mathbf{C_{depth}}$ should be registered to $\mathbf{C_{color}}$ through the optimal displacement set

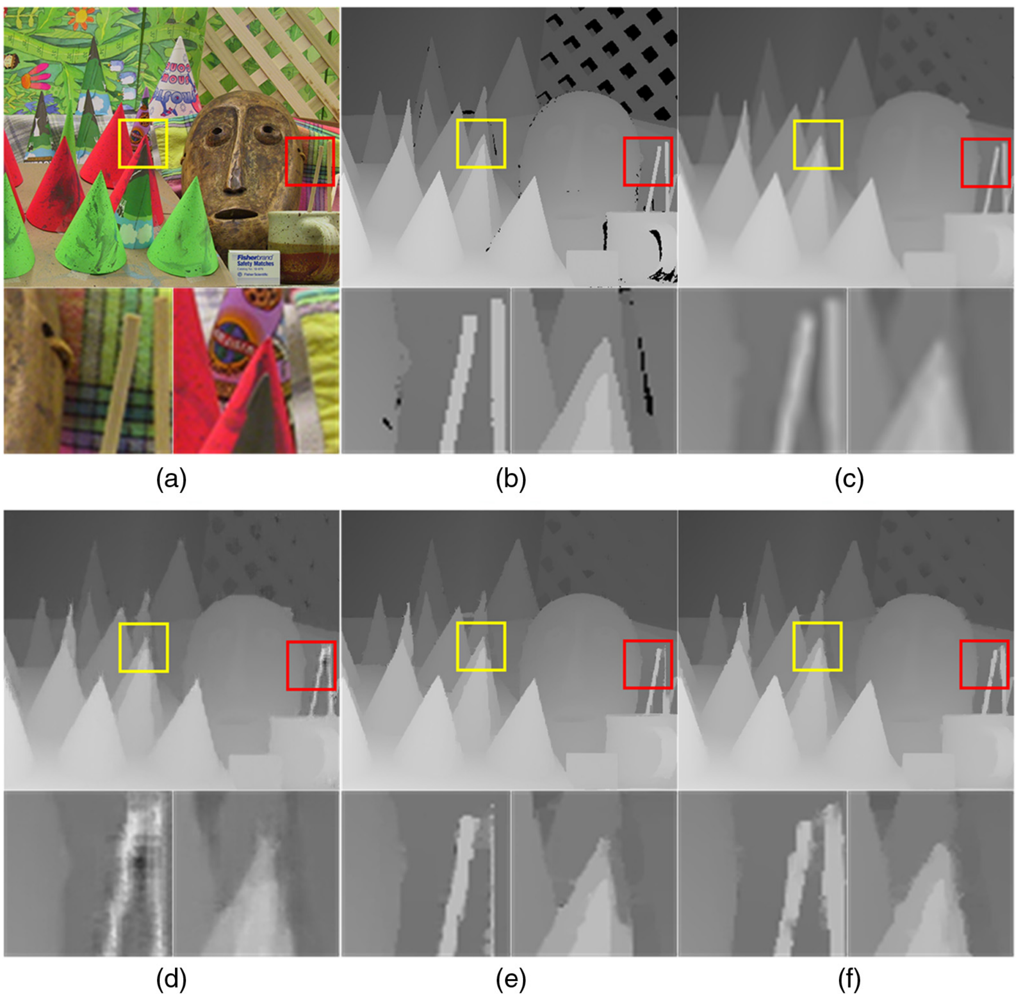


Fig. 5 The visual results of “cones”; (a) the color image, (b) the ground truth depth map, the LR depth maps are upsampled (4×) by (c) VSER,[35] (d) SUSR,[38] (e) TGV,[23] and (f) the proposed method.

$\mathbf{L}_{\mathbf{depth}}$. $\mathbf{L}_{\mathbf{depth}}$ is the solution to Eq. (7) when the depth edge map is the reference edge map. If there are more than one depth edge pixels mapping to the same color edge pixel, the best mapping with the lowest $C(p,k)$ is adopted. We use $\mathbf{C}'_{\mathbf{depth}}$ to represent the registered $\mathbf{C}_{\mathbf{depth}}$. To take the bidirectional evaluation into account, a confidence map $\boldsymbol{\alpha}$ is defined as

$$\boldsymbol{\alpha} = \max(\mathbf{C}'_{\mathbf{depth}}, \mathbf{C}_{\mathbf{color}}). \tag{11}$$

This confidence map is explicitly embedded into the construction of $\boldsymbol{\phi}$ to mitigate texture-copying artifacts and preserve depth edges. Specially, the affinity $\boldsymbol{\phi}(p,q)$ between neighboring pixels p and q is defined as

$$\boldsymbol{\phi}(p,q) = \exp\left[-\frac{(1-\alpha_{pq})|\nabla I_{pq}| + \alpha_{pq}|\nabla \hat{D}_{pq}|}{2\delta^2}\right], \tag{12}$$

where α_{pq} is defined as $\alpha_{pq} = \max(\boldsymbol{\alpha}_p, \boldsymbol{\alpha}_q)$ which integrates $\boldsymbol{\alpha}_p$ and α_q together to better mitigate texture-copy artifacts and preserve depth edges because of the single pixel width edges detected by Canny operator. ∇I_{pq} and $\nabla \hat{D}_{pq}$ are the differences between p and q on the color image and the coarsely interpolated depth map, respectively. It is observed that ∇I_{pq} is able to play more important role when the color edge map is more consistent with the depth edge map (i.e., α_{pq} is closer to zero) and vice versa.

The optimization problem [i.e., Eq. (1)] is nonconvex. In this paper, we use the coarsely interpolated depth map $\hat{\mathbf{D}}$ as the initial values and L-BFGS is adopted to solve Eq. (1).

4 Experimental Results

In this section, we evaluate different state-of-the-art methods on Middlebury datasets.[34] The LR depth map is generated by downsampling the ground truth through nearest neighbor interpolation, which is also used in methods.[5,35–37] The downsampling scales are 2×, 4×, and 8×. So, the observed depth values are ground-truth depth values, which mean that they have high confidence. The proposed method expects the reconstructed depth values at such locations are consistent with the observed ones. Therefore, relative small values are chosen for λ and γ to increase the contribution of the data term E_{data}. In addition, to balance the contributions of E_{regl} and E_{regp}, γ should be smaller than λ based on the observation of the cost values of these terms. More specifically, λ and γ are empirically assigned to 0.01 and 0.001, respectively. These values are tuned to balance the performance and the robustness of computation on “cones” dataset and they are fixed for other datasets.

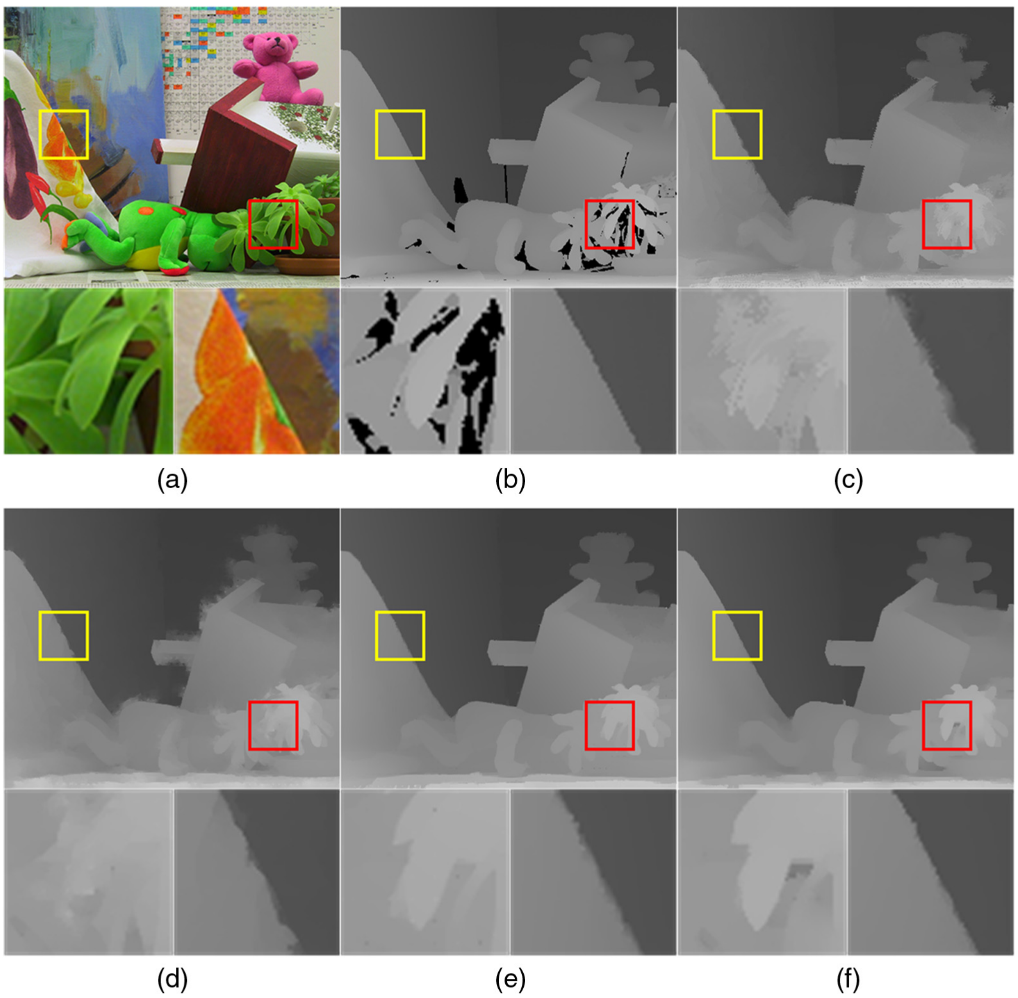


Fig. 6 The visual results of “teddy”; (a) the color image, (b) the ground truth depth map, the LR depth maps are upsampled (8×) by (c) WMF,[15] (d) NLMR,[24] (e) TGV,[23] and (f) the proposed method.

To validate the advantage of integrating cosparse analysis model with explicit edge inconsistency measurement, the performance comparison against the models that are adopted only one component (i.e., either cosparse analysis model or edge inconsistency measurement) is conducted. Then, the proposed method is compared with other state-of-the-art approaches.

4.1 *Comparison Between Variants of the Proposed Method*

We compare the proposed method with two variants, which uses either E_{regl} or E_{regp} to define regularization term in Eq. (1) on Middlebury datasets in the cases of 2×, 4×, and 8×, respectively. As shown in Table 1 which evaluates results by MAE, the complex regularization term $(E_{\mathrm{regl}} + E_{\mathrm{regp}})$ outperforms the others. The data-driven prior is not robust when the upsampling factor is large, but it still provides certain complementary to the model-based prior. Figure 3 shows the subjective comparison between such variants. It is observed that the results are over-smooth when only E_{regl} is used to regularize the depth map. Edge leaking artifacts are shown on the results of the model that uses E_{regp} only. Compared with the two above, the proposed method is closest to the ground truth.

4.2 *Comparison with Other State-of-the-Art Methods*

The proposed method is further compared with seven benchmark and state-of-the-art approaches, which are bicubic interpolation, JBUV,[12] WMF,[15] NLMR,[24] TGV,[23] SUSR,[38] and VSER.[35] Table 2 shows the results of these methods evaluated by MAE. SUSR and VSER are the approaches to single depth upsampling, which is a more changeling problem. Only the results of the cases 2× and 4× are available for them. The optimal one is marked in bold. It is observed that the proposed method can significantly improve the performance of depth map upsampling compared with other state-of-the-art approaches. Figures 4 and 5 show the visual results of 4× case on "Tuskuba" and "cones" datasets. The proposed method is compared with two approaches of single depth map upsampling (SUSR, VSER) and a method of guided depth map upsampling (TGV). From the highlighted regions, the results of SUSR and VSER always blur the depth edges. There are texture-copying artifacts presented in the results of SUSR. In addition, TGV cannot preserve tiny details (e.g., the sticks in the cup) of the depth map. Compared with these approaches, the proposed method performs better in preserving depth edges and mitigating texture-copying artifacts.

To further evaluate the robustness of the proposed method in the case of large upsampling factor, Figs. 6 and 7 show

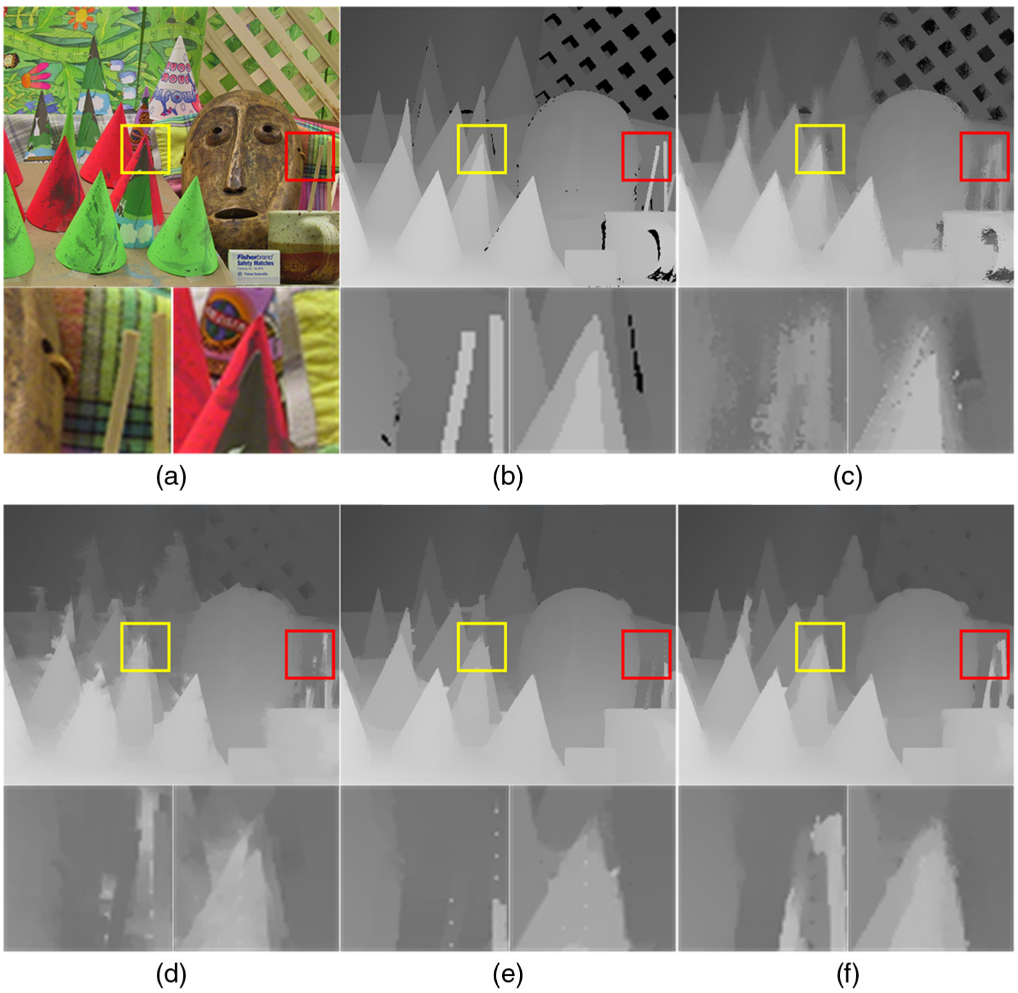


Fig. 7 The visual results of "cones"; (a) the color image, (b) the ground truth depth map, the LR depth maps are upsampled (8×) by (c) WMF,[15] (d) NLMR,[24] (e) TGV,[23] and (f) the proposed method.

the visual results of 8× case on "cones" and "teddy" datasets. The highlighted regions show that WMF and NLMR perform poor in preserving depth edges. Although TGV can better preserve edges than WMF and NLMR, it may fail restoring tiny details of the depth map (e.g., highlighted by red square in Fig. 7). Compared with them, the proposed method provides more robust performance of depth map upsampling.

5 Conclusion

This paper proposes a method for guided depth map upsampling, which integrates data-driven prior with model-based prior to regularize the HR depth map. The data-driven prior based on analysis sparse coding can be treated as the regularization of the high-order MRF model. The model-based prior embedding with edge inconsistency measurement model can mitigate texture-copying artifacts. The two parts of regularization complement each other to restore high-quality depth map. The experimental results show the improvement in the proposed method compared with other state-of-the-art approaches.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China, under Grant Nos. U1301257, 61571285, 61422111, and 61172096.

References

1. P. Thanusutiyabhorn, P. Kanongchaiyos, and W. S. Mohammed, "Image-based 3D laser scanner," in *8th Int. Conf. on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pp. 975–978, IEEE (2011).
2. A. Kolb et al., "Time-of-flight cameras in computer graphics," *Comput. Graphics Forum* **29**(1), 141–159 (2010).
3. J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-8**(6), 679–698 (1986).
4. F. Garcia et al., "Unified multi-lateral filter for real-time depth map enhancement," *Image Vision Comput.* **41**, 26–41 (2015).
5. J. Yang et al., "Color-guided depth recovery from RGB-D data using an adaptive autoregressive model," *IEEE Trans. Image Process.* **23**(8), 3443–3458 (2014).
6. K.-H. Lo, K.-L. Hua, and Y.-C. F. Wang, "Depth map super-resolution via Markov random fields without texture-copying artifacts," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1414–1418, IEEE (2013).
7. O. Choi and S.-W. Jung, "A consensus-driven approach for structure and texture aware depth map upsampling," *IEEE Trans. Image Process.* **23**(8), 3321–3335 (2014).
8. S. Xiang, L. Yu, and C. W. Chen, "No-reference depth assessment based on edge misalignment errors for T+D images," *IEEE Trans. Image Process.* **25**(3), 1479–1494 (2016).
9. Y. Zuo et al., "Explicit edge inconsistency evaluation model for color-guided depth map enhancement," *IEEE Trans. Circuits Syst. Video Technol.* **28**(2), 439–453 (2018).
10. S. Hawe, M. Kleinsteuber, and K. Diepold, "Analysis operator learning and its application to image reconstruction," *IEEE Trans. Image Process.* **22**(6), 2138–2150 (2013).
11. J. Kopf et al., "Joint bilateral upsampling," *ACM Trans. Graphics* **26**(3), 96 (2007).
12. Q. Yang et al., "Spatial-depth super resolution for range images," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE (2007).
13. K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(6), 1397–1409 (2013).
14. F. Garcia et al., "Real-time distance-dependent mapping for a hybrid ToF multi-camera rig," *IEEE J. Sel. Top. Signal Process.* **6**(5), 425–436 (2012).
15. D. Min, J. Lu, and M. N. Do, "Depth video enhancement based on weighted mode filtering," *IEEE Trans. Image Process.* **21**(3), 1176–1190 (2012).
16. M.-Y. Liu, O. Tuzel, and Y. Taguchi, "Joint geodesic upsampling of depth images," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 169–176 (2013).
17. K.-H. Lo, Y.-C. F. Wang, and K.-L. Hua, "Joint trilateral filtering for depth map super-resolution," in *Visual Communications and Image Processing (VCIP)*, pp. 1–6, IEEE (2013).
18. K.-L. Hua, K.-H. Lo, and Y.-C. F. F. Wang, "Extended guided filtering for depth map upsampling," *IEEE Multimedia* **23**(2), 72–83 (2016).
19. J. T. Barron and B. Poole, "The fast bilateral solver," *Lect. Notes Comput. Sci.* **9907**, 617–632 (2016).
20. J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," in *Proc. of the 18th Int. Conf. on Neural Information Processing Systems*, pp. 291–298 (2005).
21. J. Zhu et al., "Spatial-temporal fusion for high accuracy depth maps using dynamic MRFs," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(5), 899–909 (2010).
22. J. Lu et al., "A revisit to MRF-based depth map super-resolution and enhancement," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 985–988, IEEE (2011).
23. D. Ferstl et al., "Image guided depth upsampling using anisotropic total generalized variation," in *Proc. of the IEEE Int. Conf. on Computer Vision*, pp. 993–1000 (2013).
24. J. Park et al., "High-quality depth map upsampling and completion for RGB-D cameras," *IEEE Trans. Image Process.* **23**(12), 5559–5572 (2014).
25. Y. Yang et al., "Depth map super-resolution via local and nonlocal priors," *J. Electron. Imaging* **23**(2), 023019 (2014).
26. Y. Li et al., "Fast guided global interpolation for depth and motion," *Lect. Notes Comput. Sci.* **9907**, 717–733 (2016).
27. D. Min et al., "Fast global image smoothing based on weighted least squares," *IEEE Trans. Image Process.* **23**(12), 5638–5653 (2014).
28. W. Liu et al., "Robust color guided depth map restoration," *IEEE Trans. Image Process.* **26**(1), 315–327 (2017).
29. Y. Li et al., "Joint example-based depth map super-resolution," in *IEEE Int. Conf. on Multimedia and Expo (ICME)*, pp. 152–157, IEEE (2012).
30. H. Kwon, Y.-W. Tai, and S. Lin, "Data-driven depth map refinement via multi-scale sparse representation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 159–167 (2015).
31. M. Kiechle et al., "A bimodal co-sparse analysis model for image processing," *Int. J. Comput. Vision* **114**(2–3), 233–247 (2015).
32. H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logist.* **2**(1–2), 83–97 (1955).
33. Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(11), 1222–1239 (2001).
34. D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Proc.*, Vol. 1, I-195–I-202, IEEE (2003).
35. D. Ferstl, M. Ruther, and H. Bischof, "Variational depth superresolution using example-based edge representations," in *Proc. of the IEEE Int. Conf. on Computer Vision*, pp. 513–521 (2015).
36. M. Hornácek et al., "Depth super resolution by rigid body self-similarity in 3D," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1123–1130, IEEE (2013).
37. O. M. Aodha et al., "Patch based synthesis for single depth image super-resolution," *Lect. Notes Comput. Sci.* **7574**, 71–84 (2012).
38. R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," *Lect. Notes Comput. Sci.* **6920**, 711–730 (2010).

Yifan Zuo received his bachelor's degree in electronic information engineering from Nanchang University, Nanchang, China, in 2008, and his master's degree in signal processing from Shanghai University, Shanghai, China, in 2012. Currently, he is pursuing his PhD with the School of Communication and Information Engineering, Shanghai University, China, and also with the faculty of engineering and information technology, University of Technology, Sydney, Ultimo, NSW, Australia. His current research interests include depth map refinement and depth map estimation.

Qiang Wu received his BEng and MEng degrees from Harbin Institute of Technology, Harbin, China, in 1996 and 1998, respectively, and his PhD from University of Technology Sydney, Ultimo, NSW, Australia, in 2004. He is an associate professor and a core member of the Global Big Data Technologies Center, University of Technology Sydney. His research interests include computer vision, image processing, pattern recognition, machine learning, and multimedia processing.

Ping An received her BA and MS degrees from the Hefei University of Technology, Hefei, China, in 1990 and 1993, respectively, and her PhD from Shanghai University, Shanghai, China, in 2002. Currently, she is a professor with the Video Processing Group, School of Communication and Information Engineering, Shanghai

University. Her research interests include image and video processing, with a focus on 3D video processing.

Xiwu Shang received his BS degree in electronic information engineering from Hubei University, Hubei, China, in 2010, and his MEng degree and PhD degrees in signal and information processing from Shanghai University, Shanghai, China, in 2014 and 2018, respectively. His research interests are three-dimensional video coding and high-efficiency video coding.