Yuan *EURASIP Journal on Image and Video Processing* (2018) 2018:66
https://doi.org/10.1186/s13640-018-0310-x

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Robust semi-automatic 2D-to-3D image conversion via residual-driven optimization

Hongxing Yuan

Abstract

Semi-automatic 2D-to-3D conversion provides a cost-effective solution to the problem of 3D content shortage. The performance of most methods degrades significantly when cross-boundary scribbles are present due to their inability to remove unwanted input. To address this problem, a residual-driven energy function is proposed to remove unwanted input introduced by cross-boundary scribbles while preserving expected user input. Firstly, confidence of user input is computed from residuals between the estimation and user-specified depth values, and it is applied to the data fidelity term. Secondly, the residual-driven optimization is performed to estimate dense depth from user scribbles. The procedure is repeated until a maximum number of iterations is exceeded. Input confidence based on residuals avoids the propagation of unwanted scribbles and thus enables to generate high-quality depth even with cross-boundary input. Experimental results demonstrate that the proposed method removes unwanted scribbles successfully while preserving expected input, and it outperforms the state-of-the-art when presented with cross-boundary scribbles.

Keywords: 3D video, 2D-to-3D conversion, Depth, Cross-boundary scribbles, Optimization

1 Introduction

2D-to-3D conversion aims to estimate depth from 2D images and generates stereoscopic views from the depth, which is a key technology to produce 3D content [1]. Existing approaches are mainly categorized into two groups: automatic and semi-automatic methods.

Automatic methods try to create depth from 2D images using various depth cues, such as dark channel [2], motion [3], lighting bias [4], defocus [5], geometry [6], boundary [7], etc. Each cue is only applicable to certain scenes [8], and thus, these methods are hard to provide acceptable results in any general content. Recently, neural networks have been employed to learn the implicit relation between depth and color values [9–12]. However, these learning-based methods are limited to the trained image types [13].

Semi-automatic methods address these issues by introducing human interactions. The objective of these approaches is to produce a dense depth-map from user scribbles which indicate the labeled pixels are farther or closer from the camera [14]. In order to solve the problem of 3D content shortage, many methods have been developed for depth estimation from user input. Guttmann et al. [15] employed user scribbles to train a support vector machines (SVM) classifier that assigns depth to image patches, but results may be inaccurate due to misclassifications. S'ykora et al. [16] proposed an interactive method for user adding depth (in)equalities information and formulated depth propagation as an optimization problem, but it may produce several artifacts due to the incorrect estimation of contour thickness. Rzeszutek et al. [17] utilized the random-walks (RW) algorithm to generate dense depth-maps from user input, but RW has problems in preserving strong edges. Phan et al. [18] appended graph-cuts (GC) segmentation to the neighbor cost in RW to preserve depth boundaries. Xu et al. [19] proposed a similar method which uses a fast watershed segmentation to replace GC. Zhang et al. [20] combined automatic depth estimation from multiple cues and interactive object segmentation to obtain the final depth. Zeng et al. [21] utilized occlusion cues and shape priors to obtain a rough approximation of depth and refined the estimation using an interactive ground fitting. These segmentation-based methods can preserve strong edges but may generate artifacts due to incorrect segments.

Correspondence: yuanhx@mail.ustc.edu.cn
School of Electronics and Information Engineering, Ningbo University of Technology, Fenghua Road, 315211 Ningbo, China

Yuan et al. [22] incorporated non-local neighbors into the RW algorithm to improve depth quality. Liang et al. [23] extended this scheme to support video conversion using spatial-temporal information. Wang et al. [24] propagated user-specified sparse depth into dense depth using an optimization method originally used for colorization [25]. Wu et al. [26] improved this method with depth consistency between superpixels. Liao et al. [27] used a diffusion process to generate a depth map from user coarse annotations.

Depth-map is typically made of smooth regions separated by sharp transitions along the boundaries between different objects [28]. Therefore, existing semi-automatic methods require that user scribbles do not cross object boundaries; otherwise, the quality of produced depth degrades significantly. As shown in Fig. 1, when user scribbles cross object boundaries, the state-of-the-art methods [18, 22, 24] will produce depth artifacts. In 2D-to-3D conversion, the cross-boundary scribbles are introduced by users carelessly. As for a cross-boundary scribble, its longer part is usually user expected input and shorter part is unwanted input. It can be seen from Fig. 1f that the proposed method can remove depth artifacts caused by unwanted user input from cross-boundary scribbles.

Semi-automatic image segmentation methods have addressed the problem of cross-boundary scribbles [29–31]. Although Subr et al. [29] and Bai et al. [30] can reduce artifacts caused by cross-boundary scribbles, they focus on the foreground object segmentation and are hard to apply in 2D-to-3D conversion. Oh et al. [31] used occurrence and co-occurrence probability (OCP) of color values at labeled pixels to estimate the confidence of user input. This method can be used for 2D-to-3D conversion, but it may mistake expected scribbles for unwanted ones.

Surprisingly, there are few methods to consider the impact of cross-boundary scribbles on 2D-to-3D conversion. To address this problem, we propose a robust method based on residuals between the user-specified and estimated depth values during the iteratively solving process. Thanks to the confidence of user scribbles measured by the residuals, experimental results show that the proposed method can remove depth artifacts caused by cross-boundary scribbles. The two most relevant to this work are Wang et al. [24] and Hong et al. [32]. Unlike the optimization model in Wang et al. [24], the proposed method utilizes residuals to eliminate the depth artifacts caused by cross-boundary scribbles. The main difference to Hong et al. [32] is that they use residuals to determine the relative weight between data fidelity and regularization, whereas this paper leverages residuals to compute the confidence of user scribbles.

Recently, Ham et al. [33] proposed a static dynamic filter (SDF) to reduce artifacts caused by structural differences between guidance and input signals. Although SDF [33] can handle differences in structure, it is not robust to outliers introduced by cross-boundary scribbles. Yuan et al. [34] proposed an ℓ_1 optimization method to remove user erroneous scribbles. However, ℓ_1 norm assumes that input image can be approximated by the sum of a piecewise-constant function and a smooth function [35]. Depth artifacts will be introduced when the assumption does not hold.

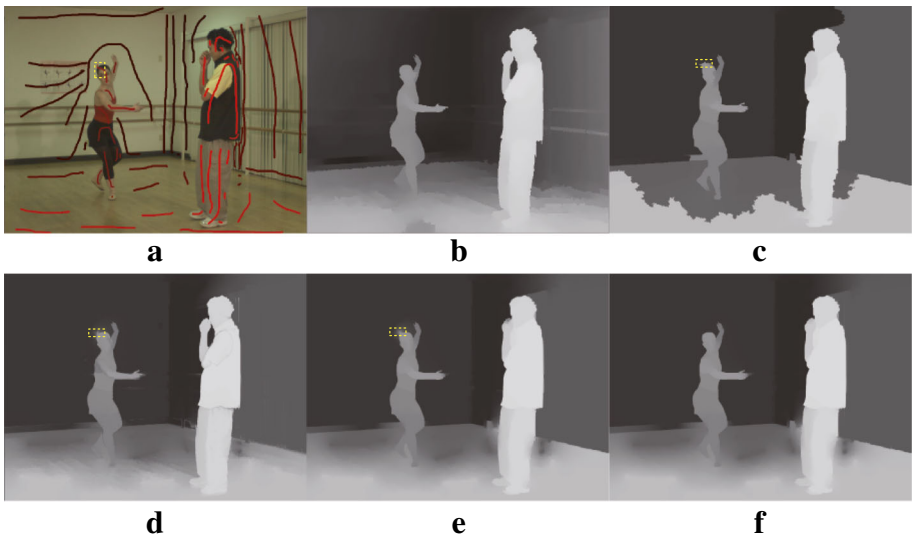

Fig. 1 Depth estimation with cross-boundary user input (depth artifacts caused by cross-boundary scribbles are marked by yellow rectangles). **a** Input image with user scribbles (the cross-boundary scribble is marked by the yellow rectangle). **b** Groundtruth. **c** Hybrid GC and RW [18]. **d** Nonlocal RW [22]. **e** Optimization [24]. **f** Proposed. Please zoom in to see details

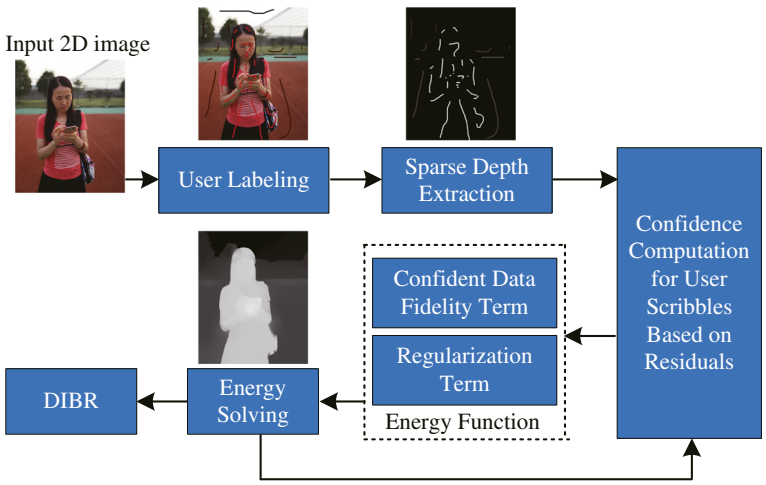

Fig. 2 A flowchart of the semi-automatic 2D-to-3D image conversion with the proposed method

The remainder of this paper is organized as follows. In Section 2, the proposed method is described. The experimental results are given in Section 3. Finally, conclusion is given in Section 4.

2 Method

The workflow of 2D-to-3D image conversion based on the proposed method is shown in Fig. 2. Firstly, the user specifies sparse depth on an input image, where scribbles indicate the labeled pixels are closer or farther from the camera. Secondly, a sparse depth-map is extracted according to the intensities of user scribbles. Thirdly, the confidence of user scribbles is calculated based on the residuals between the estimated and user-specified depth values. Then, an energy function constraint by the confidence is designed and minimized to obtain the estimated dense depth-map. The procedure is repeated from the confidence computation step, until a maximum number of iterations is exceeded. Finally, the stereoscopic 3D image is generated by depth image-based rendering (DIBR).

2.1 Model

Let $\mathbf{O}$ be the set consisting of pixels with user-specified depth values. The objective of this paper is to estimate

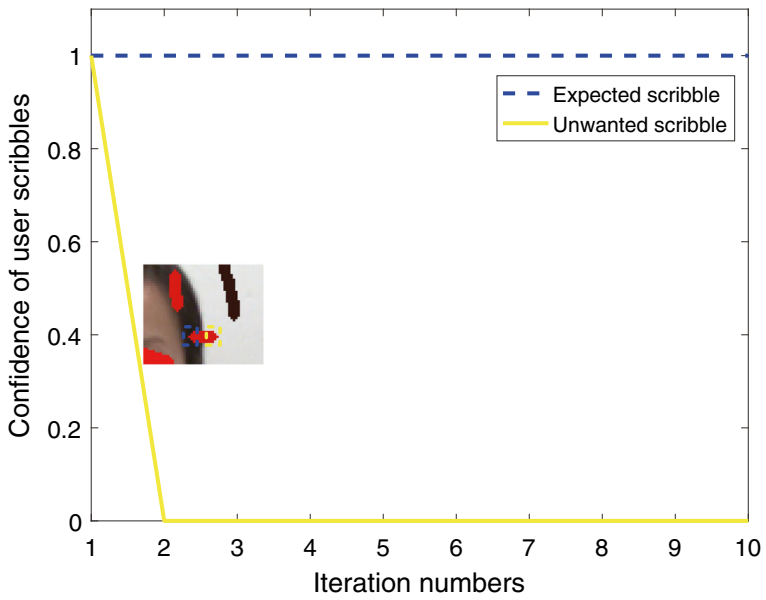

Fig. 3 Change curve of the confidence from user scribbles during iterative solution process where blue and yellow curves are for scribbles inside the blue and yellow rectangles, respectively

Table 1 SSIM of estimated depth on RGBZ datasets when cross-boundary scribbles are present

	RW	HGR	NRW	OPT	OCP	SDF	ℓ_1	Proposed
RGBZ_ 01	0.836	0.776	0.852	0.842	0.873	0.839	*0.894*	**0.932**
RGBZ_02	0.817	0.849	0.827	0.821	0.821	0.817	*0.878*	**0.931**
RGBZ_03	0.853	*0.868*	*0.868*	0.867	0.861	0.855	0.808	**0.890**
RGBZ_04	0.911	0.911	0.923	*0.928*	0.897	0.917	0.808	**0.944**
RGBZ_05	0.901	0.904	0.916	0.906	0.911	0.903	**0.935**	*0.929*
RGBZ_06	0.907	0.885	0.906	0.879	0.880	0.905	*0.909*	**0.921**
RGBZ_07	0.895	0.872	0.889	0.886	*0.906*	0.892	0.904	**0.912**
RGBZ_08	0.933	0.906	0.934	0.933	*0.944*	0.932	0.934	**0.951**
RGBZ_09	0.927	0.750	*0.938*	0.916	0.926	0.925	0.900	**0.950**
Average	0.887	0.858	*0.895*	0.886	0.891	0.887	0.885	**0.929**

The first and second best SSIM at each row are shown in bold and italics, respectively

an accurate dense depth-map $\mathbf{d}$ from the user input and the given image $\mathbf{I}$ even when cross-boundary scribbles are present. It can be expressed as solving the energy minimization problem:

$$\mathbf{d} = \arg\min_{\mathbf{d}\in\mathbb{R}^n} \underbrace{\sum_{i\in\mathbf{O}} r_i(d_i - u_i)^2}_{\text{data fidelity}} + \underbrace{\sum_{i=1}^{n}\sum_{j\in\mathcal{N}_i} w_{ij}(d_i - d_j)^2}_{\text{regularization}}, \quad (1)$$

where d_i and u_i denote the estimated and user-specified depth values at pixel i, respectively. n is the size of the input image $\mathbf{I}$. $\mathcal{N}_i$ represents the set of 8-connected neighbors for pixel i. w_{ij} is a weighting function to make pixels with similar colors have similar depth values and is defined as

$$w_{ij} = \begin{cases} \exp\left(-\beta \left\|\mathbf{I}_i - \mathbf{I}_j\right\|^2\right) & \text{if } j \in \mathcal{N}_i, \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

where $\mathbf{I}_i$ and $\mathbf{I}_j$ are the color values of image $\mathbf{I}$ at pixel i and j, respectively. β in Formula (2) is a parameter controlling the strength of the weight w_{ij}.

r_i in Formula (1) is a confidence measure of the user-specified depth value at pixel i and is defined asz

$$r_i = \begin{cases} \exp\left(-\eta(d_i - u_i)^2\right) & \text{if } i \in \mathbf{O}, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

Here, η is a constant that controls how dissimilar two depth values are. In Formula (1), the data fidelity term enforces the estimated depth values of labeled regions to approximate the user-specified ones. Unlike Wang et al. [24], the proposed method maintains this consistency only when user inputs are confident. The confidence r_i is low when the residual $(d_i - u_i)^2$ is high. The regularization term is used to penalize the difference of the estimated depth values between each pixel and its neighbors.

2.2 Solver

Formula (1) is nonlinear to $\mathbf{d}$ and thus is an unconstrained, non-linear optimization. A fixed point iteration strategy is adopted to solve Formula (1). Let $\mathbf{d}^k = \left[d_i^k\right]_{n\times 1}$ and $\mathbf{u}$ denote vectors representing the estimated depth image in iteration k and user-specified depth values, respectively. The i-th element of $\mathbf{u}$ is user-specified depth value u_i if $i \in \mathbf{O}$ and 0 otherwise. Then, in iteration k, the objective function to be minimized is expressed as

$$E\left(\mathbf{d}^k\right) = \left(\mathbf{d}^k - \mathbf{u}\right)^T \mathbf{R}^{k-1}\left(\mathbf{d}^k - \mathbf{u}\right) + \lambda \mathbf{d}^{k,T}\mathbf{L}\mathbf{d}^k, \quad (4)$$

where $\mathbf{R}^{k-1}$ is a $n \times n$ diagonal matrix and its i-th diagonal element is r_i^{k-1}. Here, $r_i^{k-1} = \exp\left(-\eta\left(d_i^{k-1} - u_i\right)^2\right)$ if $i \in \mathbf{O}$ and 0 otherwise. $\mathbf{L}$ is the $n \times n$ sparse Laplacian matrix. Its element $L_{ij} = -w_{ij}$ $(i \neq j)$ and $L_{ii} = \sum_{j\in\mathcal{N}_i} w_{ij}$. To

Table 2 SSIM of estimated depth on Middlebury datasets when cross-boundary scribbles are present

	RW	HGR	NRW	OPT	OCP	SDF	ℓ_1	Proposed
Tsukuba	0.724	0.716	0.723	*0.727*	0.708	0.724	0.722	**0.731**
Venus	0.969	0.961	0.968	0.970	0.966	0.969	*0.971*	**0.974**
Teddy	0.861	0.846	*0.865*	0.862	0.860	0.861	**0.868**	*0.865*
Cones	0.900	0.871	0.900	*0.902*	0.850	0.897	0.885	**0.903**
Average	0.864	0.848	0.864	*0.865*	0.846	0.863	0.862	**0.868**

The first and second best SSIM at each row are shown in bold and italics, respectively

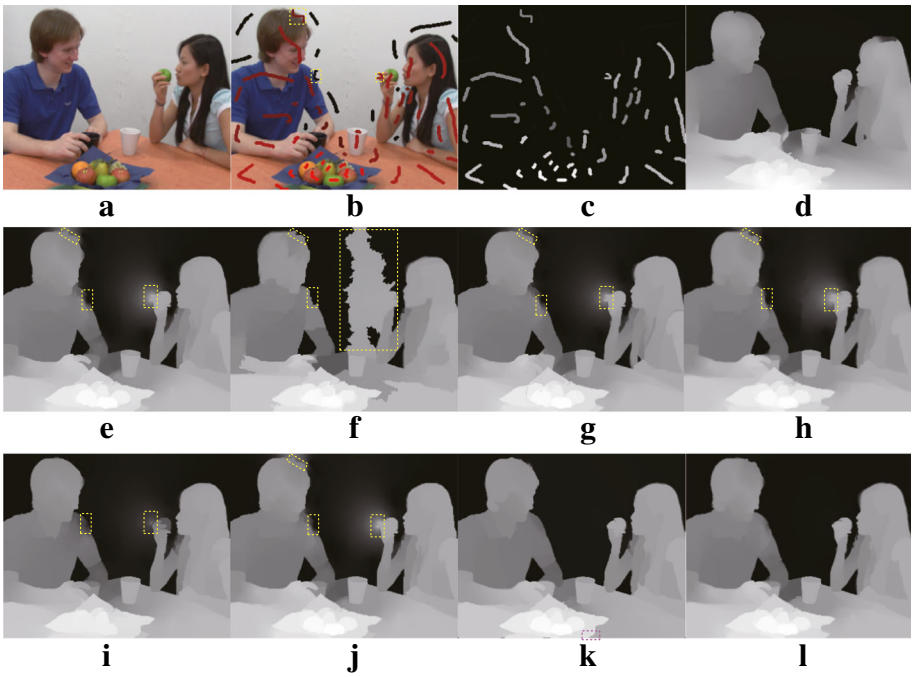

Fig. 4 Results of RGBZ_01 with cross-boundary input. **a** Input image. **b** User-labeled image. **c** Sparse depth. **d** Groundtruth depth. **e** Depth of RW. **f** Depth of HGR. **g** Depth of NRW. **h** Depth of OPT. **i** Depth of OCP. **j** Depth of SDF. **k** Depth of ℓ_1. **l** Depth of the proposed method. Please zoom in to see details

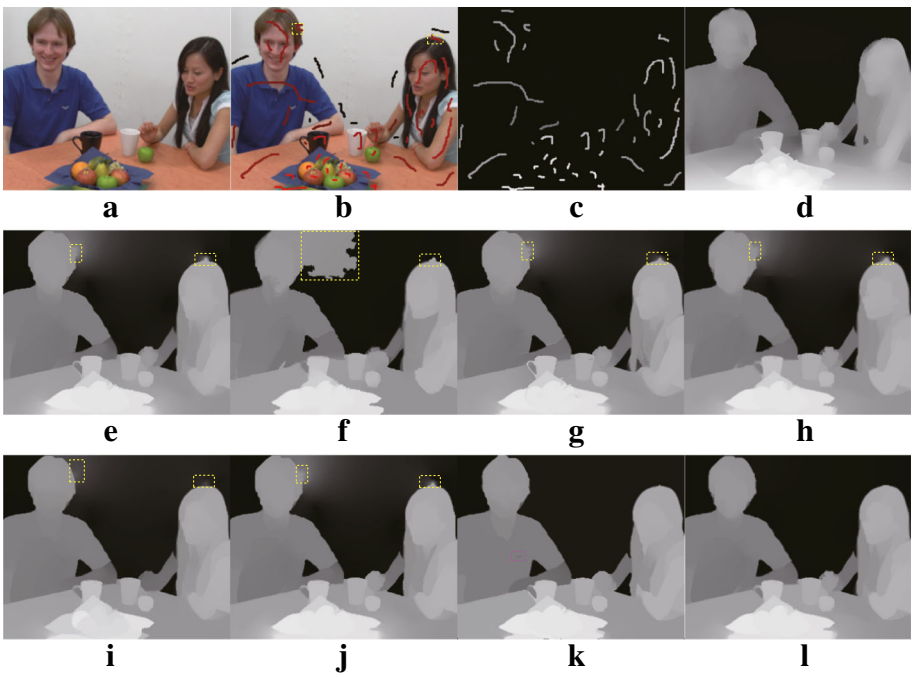

Fig. 5 Results of RGBZ_02 with cross-boundary input. **a** Input image. **b** User-labeled image. **c** Sparse depth. **d** Groundtruth depth. **e** Depth of RW. **f** Depth of HGR. **g** Depth of NRW. **h** Depth of OPT. **i** Depth of OCP. **j** Depth of SDF. **k** Depth of ℓ_1. **l** Depth of the proposed method. Please zoom in to see details

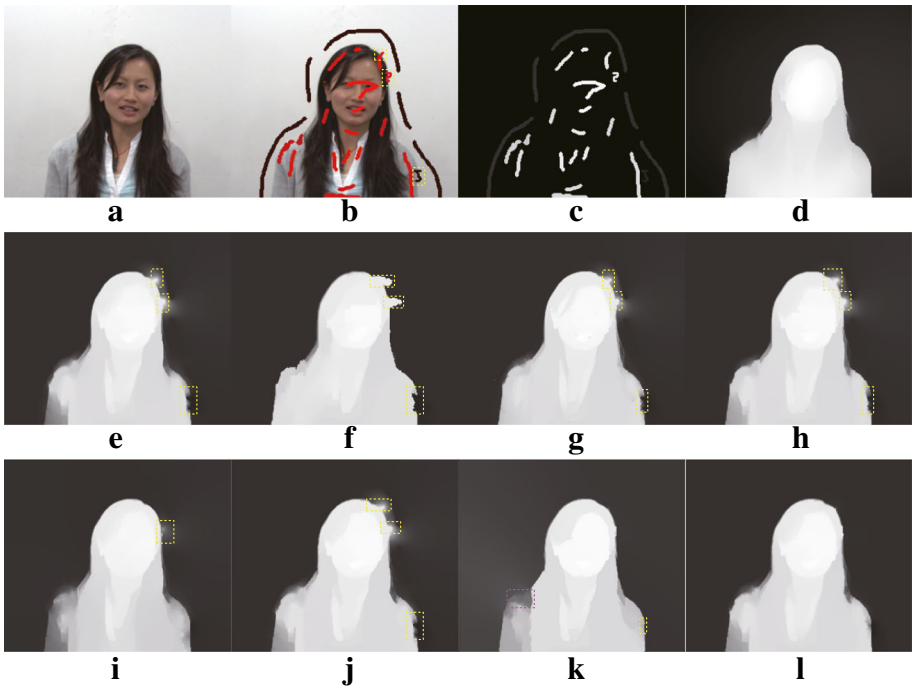

Fig. 6 Results of RGBZ_03 with cross-boundary input. **a** Input image. **b** User-labeled image. **c** Sparse depth. **d** Groundtruth depth. **e** Depth of RW. **f** Depth of HGR. **g** Depth of NRW. **h** Depth of OPT. **i** Depth of OCP. **j** Depth of SDF. **k** Depth of ℓ_1. **l** Depth of the proposed method. Please zoom in to see details

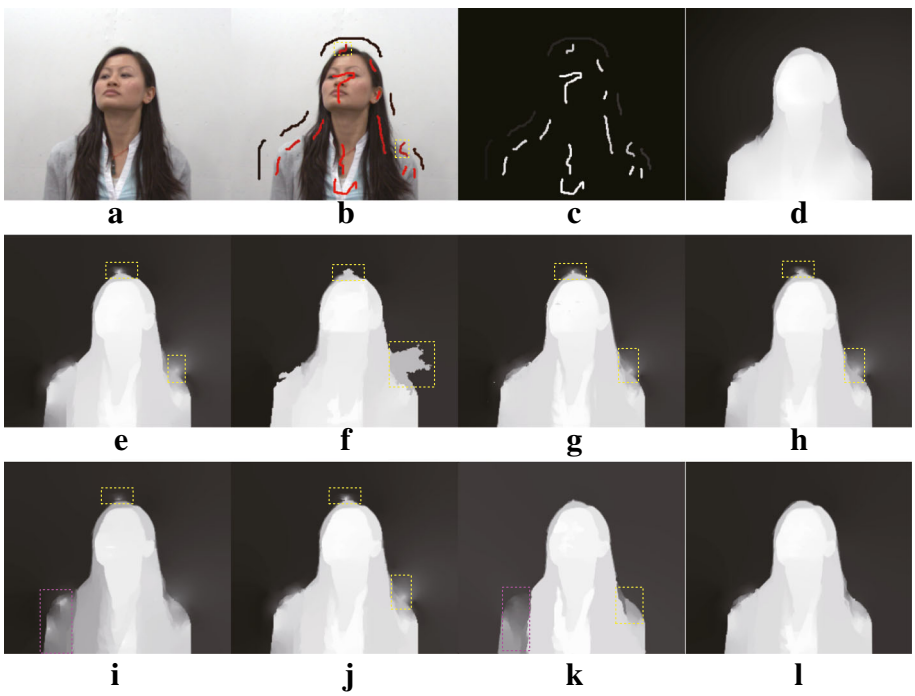

Fig. 7 Results of RGBZ_04 with cross-boundary input. **a** Input image. **b** User-labeled image. **c** Sparse depth. **d** Groundtruth depth. **e** Depth of RW. **f** Depth of HGR. **g** Depth of NRW. **h** Depth of OPT. **i** Depth of OCP. **j** Depth of SDF. **k** Depth of ℓ_1. **l** Depth of the proposed method. Please zoom in to see details

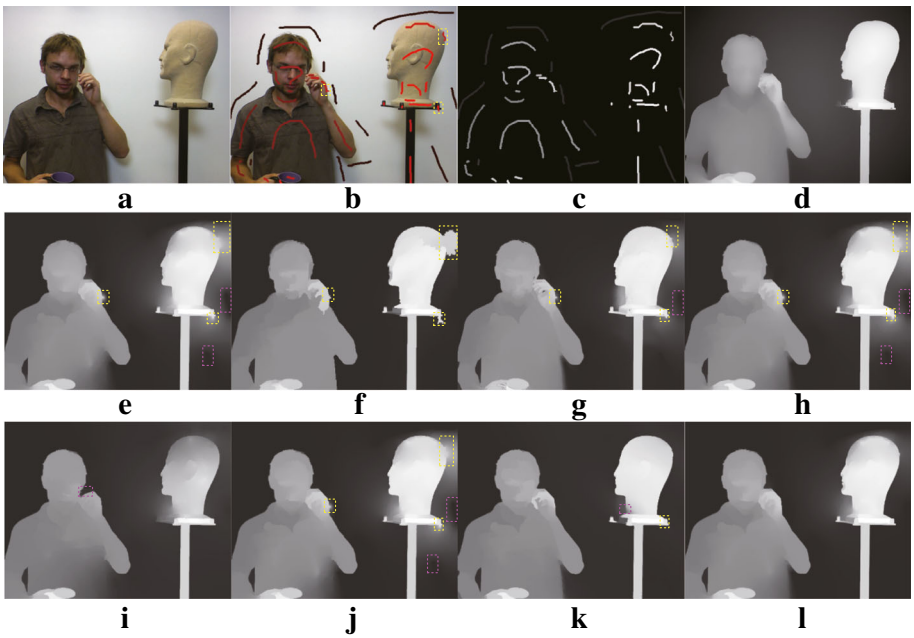

Fig. 8 Results of RGBZ_05 with cross-boundary input. **a** Input image. **b** User-labeled image. **c** Sparse depth. **d** Groundtruth depth. **e** Depth of RW. **f** Depth of HGR. **g** Depth of NRW. **h** Depth of OPT. **i** Depth of OCP. **j** Depth of SDF. **k** Depth of ℓ_1. **l** Depth of the proposed method. Please zoom in to see details

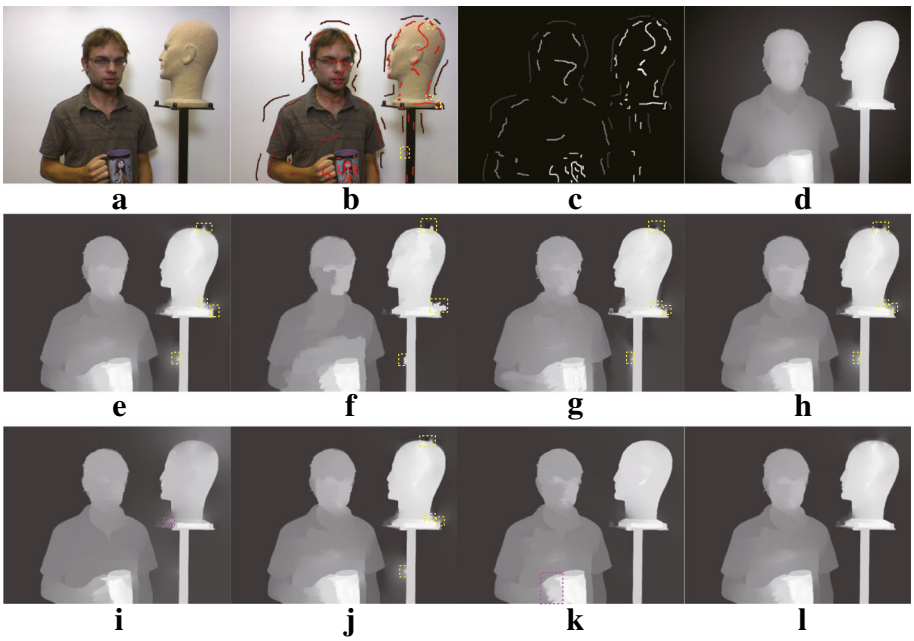

Fig. 9 Results of RGBZ_06 with cross-boundary input. **a** Input image. **b** User-labeled image. **c** Sparse depth. **d** Groundtruth depth. **e** Depth of RW. **f** Depth of HGR. **g** Depth of NRW. **h** Depth of OPT. **i** Depth of OCP. **j** Depth of SDF. **k** Depth of ℓ_1. **l** Depth of the proposed method. Please zoom in to see details

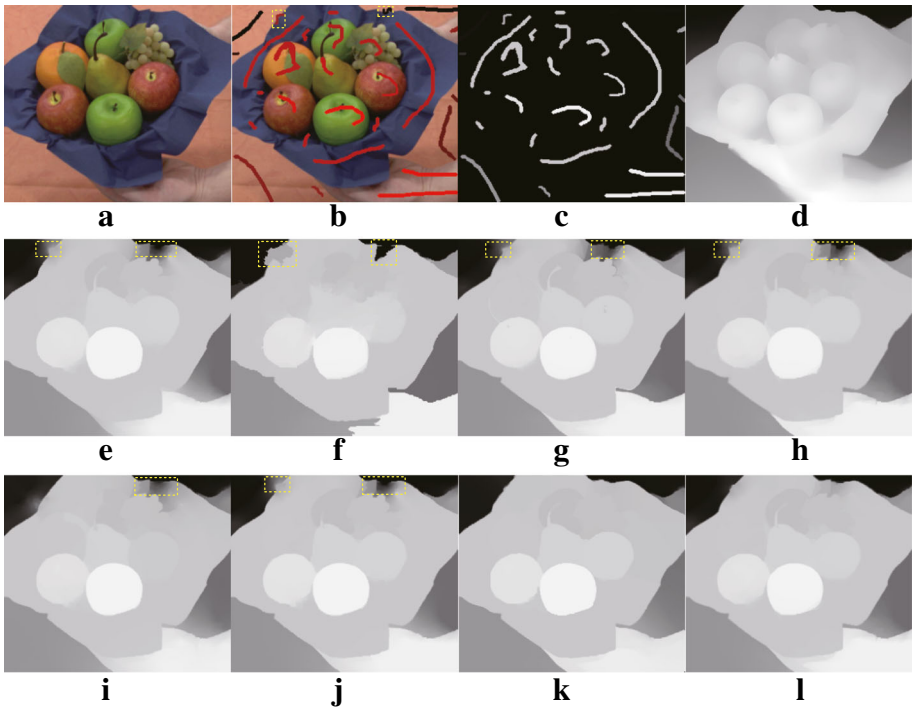

Fig. 10 Results of RGBZ_07 with cross-boundary input. **a** Input image. **b** User-labeled image. **c** Sparse depth. **d** Groundtruth depth. **e** Depth of RW. **f** Depth of HGR. **g** Depth of NRW. **h** Depth of OPT. **i** Depth of OCP. **j** Depth of SDF. **k** Depth of ℓ_1. **l** Depth of the proposed method. Please zoom in to see details

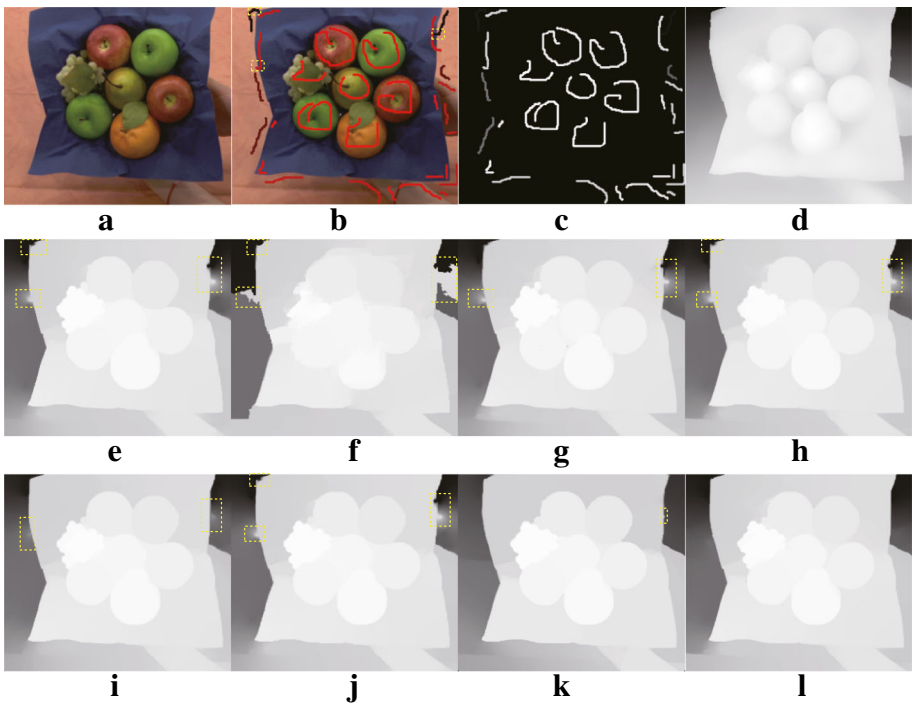

Fig. 11 Results of RGBZ_08 with cross-boundary input. **a** Input image. **b** User-labeled image. **c** Sparse depth. **d** Groundtruth depth. **e** Depth of RW. **f** Depth of HGR. **g** Depth of NRW. **h** Depth of OPT. **i** Depth of OCP. **j** Depth of SDF. **k** Depth of ℓ_1. **l** Depth of the proposed method. Please zoom in to see details

minimize the energy function in Formula (4), taking its derivatives on $\mathbf{d}^k$, Formula (5) can be obtained.

$$\frac{\partial E\left(\mathbf{d}^k\right)}{\partial \mathbf{d}^k} = 2\mathbf{R}^{k-1}\left(\mathbf{d}^k - \mathbf{u}\right) + 2\lambda \mathbf{L}\mathbf{d}^k. \tag{5}$$

The energy function in Formula (4) can be minimized by setting $\frac{\partial E\left(\mathbf{d}^k\right)}{\partial \mathbf{d}^k}$ in Formula (5) equal to zero, and Formula (6) is obtained.

$$\left(\mathbf{R}^{k-1} + \lambda \mathbf{L}\right)\mathbf{d}^k = \mathbf{R}^{k-1}\mathbf{u}. \tag{6}$$

The linear system in Formula (6) is sparse, and thus, it can be solved using standard methods such as preconditioned conjugate gradient.

2.3 Analysis

It can be seen from Formula (4) that in each iteration, user-specified depth values can only be preserved if the residuals between estimated and user-specified depth values are small.

Specifically, the unwanted user input introduced by cross-boundary scribbles will make the depth values of labeled pixels differ from their neighbors. Meanwhile, the regularization term will enforce the estimation to be consistent with their neighbors, and thus make the estimated depth to deviate from the user input. As a result, the residual between the estimated and user-specified depth values of the unwantedly labeled pixel will be increased, and the confidence computed from the residual in Formula (3) will be decreased to zero during the iterative solution process. Therefore, the proposed method can remove unwanted user input introduced by cross-boundary scribbles.

As for user-expected input, the specified values of labeled pixels are consistent with their neighbors; thus, the estimation mainly depends on the data fidelity term which enforces the estimated depth to approximate the user input. Therefore, the residuals of expectedly labeled pixels are almost 0, and their confidence will be remained at 1 with the proper setting of η in Formula (3). For this reason, the proposed method can preserve the expected user input.

Figure 3 shows the change curve of confidence from user scribbles in an input image. It can be seen that confidence of the unwanted input rapidly drops to 0 while confidence of the expected input remains at 1.

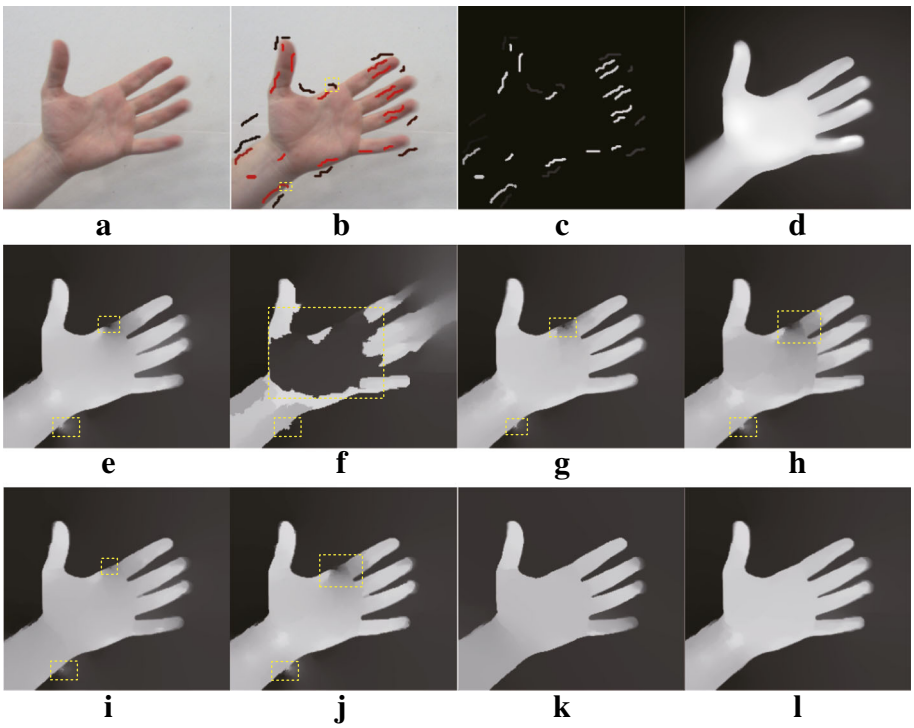

Fig. 12 Results of RGBZ_09 with cross-boundary input. **a** Input image. **b** User-labeled image. **c** Sparse depth. **d** Groundtruth depth. **e** Depth of RW. **f** Depth of HGR. **g** Depth of NRW. **h** Depth of OPT. **i** Depth of OCP. **j** Depth of SDF. **k** Depth of ℓ_1. **l** Depth of the proposed method. Please zoom in to see details

3 Experimental results and discussion

3.1 Experimental details

RGBZ (red, green, blue plus z-axis depth) datasets [36] are used for comparison which include objects, human figures, and multiple human interaction. Performance are also evaluated on four Middlebury stereo datasets, Tsukuba, Venus, Teddy, and Cones [37]. The source code and more experimental results can be downloaded from https://github.com/tcyhx/rdopt.

In the proposed method, the bandwidth parameters, η, are empirically set to 9000. A maximum number of five iterations is used to solve Formula (1). β is set to 100 for RGBZ datasets and 50 for Middlebury datasets. Results of the proposed method are compared to the state-of-the-art: RW [17], hybrid GC and RW (HGR) [18], nonlocal RW (NRW) [22], optimization (OPT) [24], OCP [31], SDF [33], and ℓ_1 [34]. Note that OCP originally aims for interactive segmentation, and this paper applies it to 2D-to-3D conversion by replacing the confidence in Formula (3) with the aggregation of the OCPs in a local neighborhood. Structural similarity (SSIM) [38] is used for performance evaluation since it can predict human perception of image quality. The standard deviation of SSIM in the experiments is set to 4 so as to evaluate the similarity of semi-global structure [39].

In the experiments, a trained user is asked to draw scribbles with a standard brush by referring to the groundtruth depth values, where higher intensities indicate the labeled

Fig. 13 Results of Tsukuba with cross-boundary input. **a** User-labeled image. **b** Sparse depth. **c** Groundtruth depth. **d** Synthesized view using c. **e** Depth of RW. **f** Synthesized view using **e**. **g** Depth of HGR. **h** Synthesized view using **g**. **i** Depth of NRW. **j** Synthesized view using **i**. **k** Depth of OPT. **l** Synthesized view using **k**. **m** Depth of OCP. **n** Synthesized view using **m**. **o** Depth of SDF. **p** Synthesized view using **o**. **q** Depth of ℓ_1. **r** Synthesized view using **q**. **s** Depth of the proposed method. **t** Synthesized view using **s**. Please zoom in to see details

pixels are closer to the camera. Since depth propagation from user scribbles relies on color or intensity similarity between neighboring pixels, more scribbles are drawn in high textured areas. To make the comparison as fair as possible, a sparse depth-map is extracted from user scribbles, and each algorithm estimates a dense depth-map from the sparse depth-map.

3.2 Experiments with cross-boundary user scribbles

In this section, a user is asked to assign the initial depth values manually by drawing some scribbles across object boundaries. Tables 1 and 2 show the SSIM values of the proposed algorithm in comparison with other methods on the RGBZ and Middlebury datasets, respectively. As shown in Tables 1 and 2, the proposed method achieves the highest average SSIM among all of the competing methods for both datasets. Except for the comparison with ℓ_1 in RGBZ_05 and Teddy, the SSIM values of the proposed method are higher than those of the other methods.

For RGBZ datasets, qualitative comparisons are shown in Figs. 4, 5, 6, 7, 8, 9, 10, 11 and 12. Qualitative

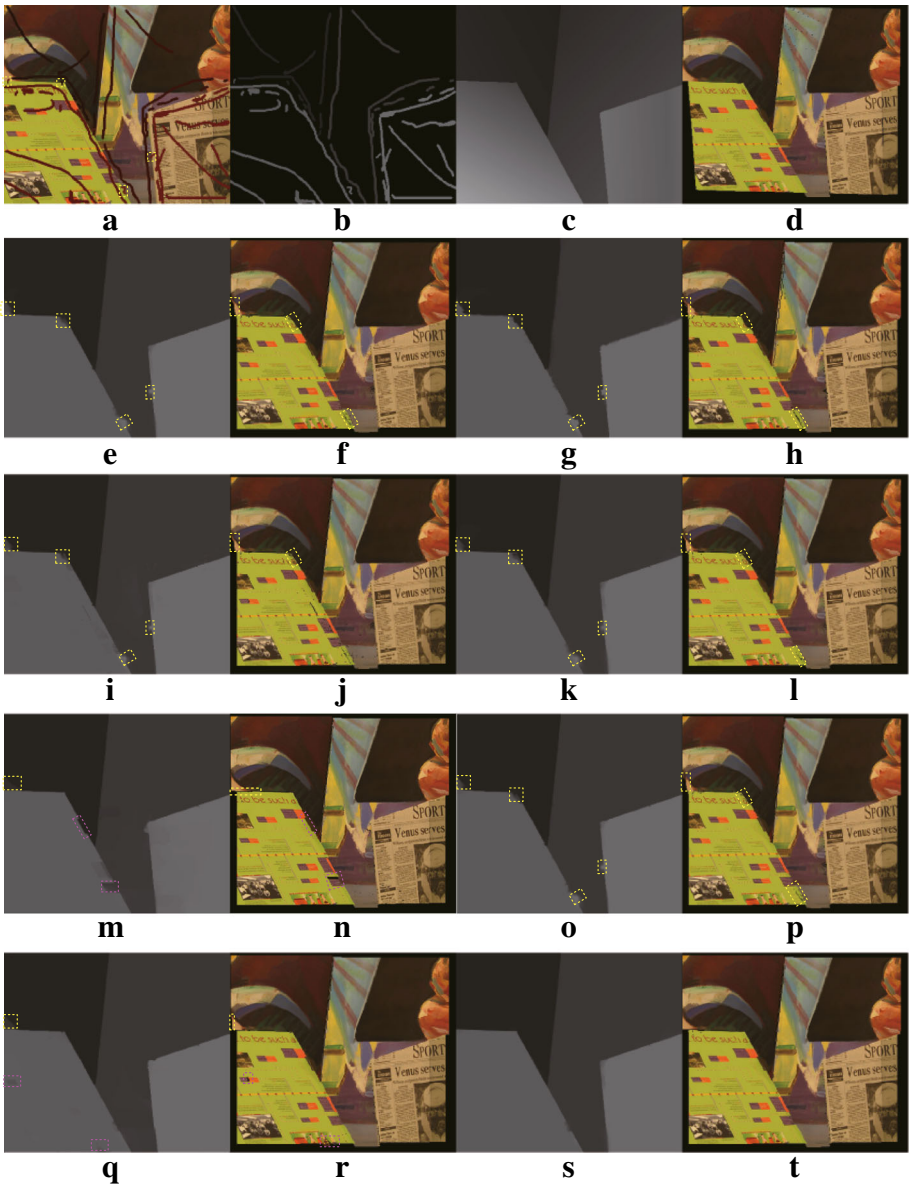

Fig. 14 Results of Venus with cross-boundary input. **a** User-labeled image. **b** Sparse depth. **c** Groundtruth depth. **d** Synthesized view using **c**. **e** Depth of RW. **f** Synthesized view using **e**. **g** Depth of HGR. **h** Synthesized view using **g**. **i** Depth of NRW. **j** Synthesized view using **i**. **k** Depth of OPT. **l** Synthesized view using k. **m** Depth of OCP. **n** Synthesized view using m. **o** Depth of SDF. **p** Synthesized view using o. **q** Depth of ℓ_1. **r** Synthesized view using q. **s** Depth of the proposed method. **t** Synthesized view using s. Please zoom in to see details

comparisons on Middlebury datasets are given in Figs. 13, 14, 15, and 16. The rendered images based on depth are only shown for Middlebury datasets in order to avoid making the lengthy paper. In each figure, the yellow rectangles on depth-maps or synthesized views represent artifacts caused by cross-boundary scribbles while the purple ones denote artifacts caused by other issues. The cross-boundary scribbles of user-labeled images are marked by the yellow rectangles (Figs. 4, 5, 6, 7, 8, 9, 10, 11, 12b, 13, 14, 15, and 16a).

RW [17] assumes that user scribbles should not cross object boundaries and thus generates depth artifacts around cross-boundary labeled regions (see Figs. 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, and 16e). These artifacts cause distortions when a new view is synthesized from the depth as shown in Figs. 13, 14, 15, and 16f. HGR [18] relies on GC to preserve depth boundaries. However, GC is sensitive to the outliers. The quality of depth-maps produced from HGR thus degrades significantly when user scribbles cross object boundaries (see Figs. 4, 5, 6, 7, 8, 9, 10, 11, 12f, 13, 14, 15, and 16g), which leads to significant degradation of quality in synthesized views (see Figs. 13, 14, 15, and 16h). Although introducing non-local constraints, NRW [22] is difficult

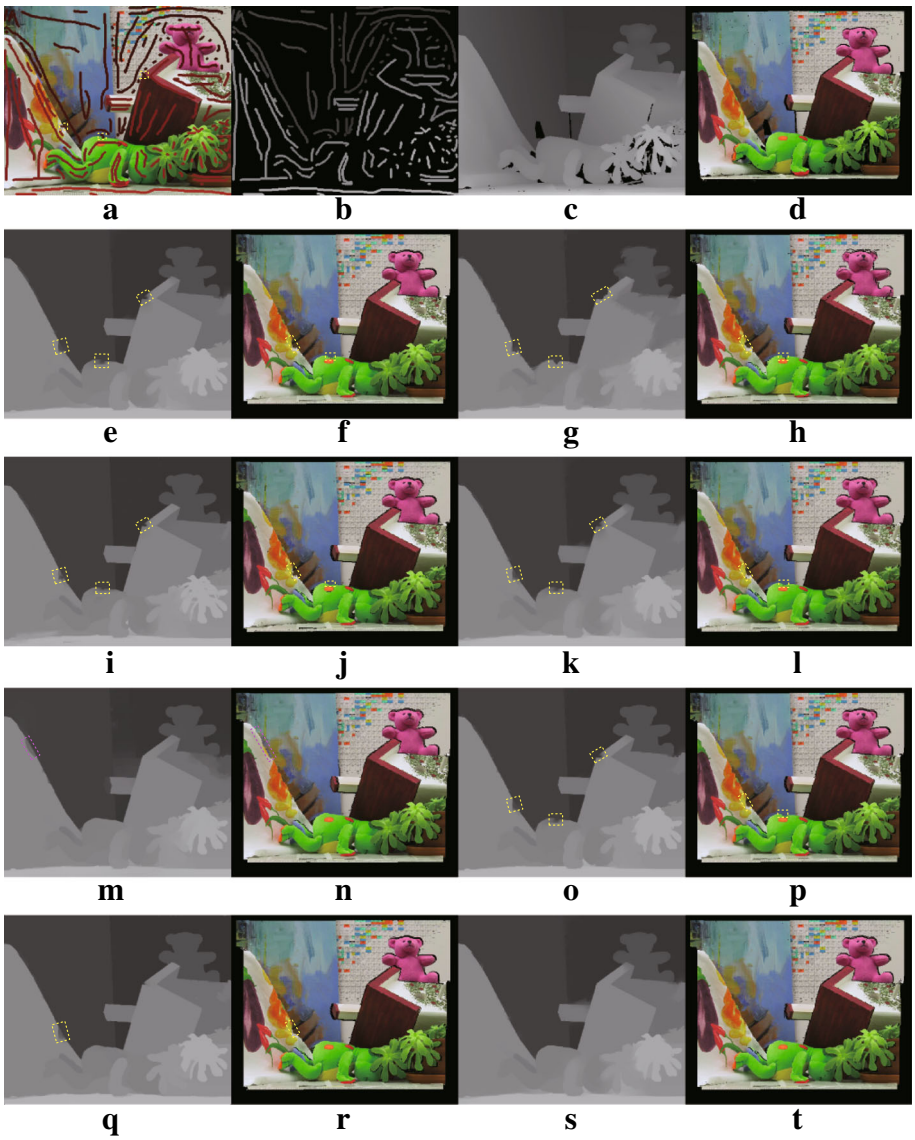

Fig. 15 Results of Teddy with cross-boundary input. **a** User labeled image. **b** Sparse depth. **c** Groundtruth depth. **d** synthesized view using **c**. **e** Depth of RW. **f** Synthesized view using e. **g** Depth of HGR. **h** Synthesized view using **g**. **i** Depth of NRW. **j** Synthesized view using **i**. **k** Depth of OPT. **l** Synthesized view using **k**. **m** Depth of OCP. **n** Synthesized view using **m**. **o** Depth of SDF. **p** Synthesized view using **o**. **q** Depth of ℓ_1. **r** Synthesized view using **q**. **s** Depth of the proposed method. **t** Synthesized view using **s**. Please zoom in to see details

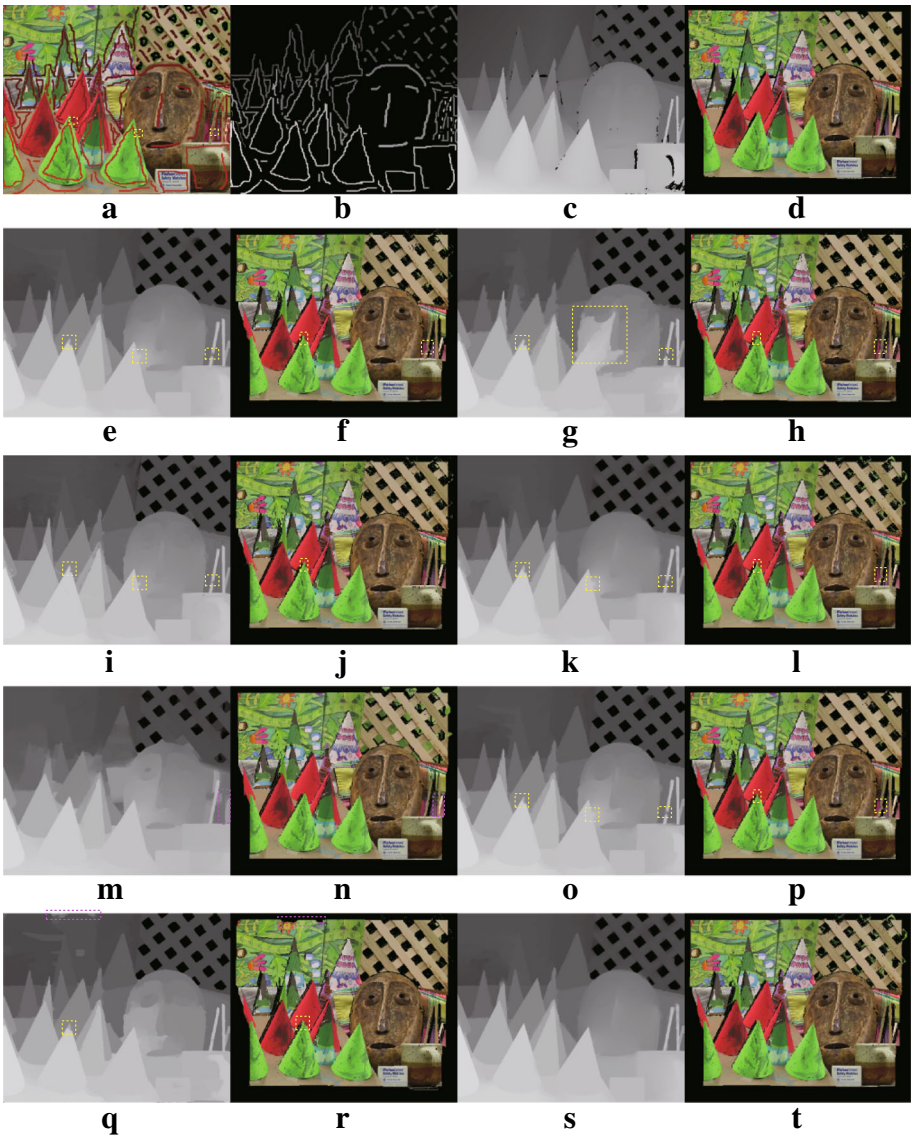

Fig. 16 Results of Cones with cross-boundary input. **a** User-labeled image. **b** Sparse depth. **c** Groundtruth depth. **d** synthesized view using **c**. **e** Depth of RW. **f** Synthesized view using **e**. **g** Depth of HGR. **h** Synthesized view using **g**. **i** Depth of NRW. **j** Synthesized view using **i**. **k** Depth of OPT. **l** Synthesized view using **k**. **m** Depth of OCP. **n** Synthesized view using **m**. **o** Depth of SDF. **p** Synthesized view using **o**. **q** Depth of ℓ_1. **r** Synthesized view using **q**. **s** Depth of the proposed method. **t** Synthesized view using **s**. Please zoom in to see details

to remove depth artifacts caused by cross-boundary user scribbles (see Figs. 4, 5, 6, 7, 8, 9, 10, 11, 12g, 13, 14, 15, and 16i), which results in distortions in synthesized views (see Figs. 13, 14, 15, and 16j). OPT [24] constrains the estimated depth values of labeled pixels to be consistent with the user input; thus, unwanted information propagates to the neighbors (see Figs. 4, 5, 6, 7, 8, 9, 10, 11, 12h, 13, 14, 15, and 16k). Distortions in synthesized views caused by input errors are shown in yellow rectangles of Figs. 13, 14, 15, and 16l. OCP [31] can remove some depth artifacts caused by cross-boundary user input, but it fails when the cross-boundary-labeled pixels have similar color distributions; thus, residual artifacts are still visible (see Figs. 4, 5, 6, 7i, 10, 11, 12i, 13, and 14m). OCP may also consider some expected scribbles as unwanted ones [31], which yields distortions as shown in the purple rectangles of Figs. 7, 8, 9i, 14, 15, and 16m. SDF [33] can reduce depth artifacts caused by structural differences between color and depth images by using the Welsch function as a regularizer. However, SDF is hard to

Table 3 SSIM of estimated depth on RGBZ datasets when cross-boundary scribbles are absent

	RW	HGR	NRW	OPT	OCP	SDF	ℓ_1	Proposed
RGBZ_01	0.925	0.913	**0.936**	0.932	0.924	0.933	0.863	*0.934*
RGBZ_02	0.932	0.917	**0.934**	*0.928*	0.919	0.880	0.920	*0.928*
RGBZ_03	0.876	**0.892**	*0.890*	0.888	0.872	0.881	0.817	*0.890*
RGBZ_04	0.928	0.936	0.938	*0.944*	0.907	0.934	0.821	**0.945**
RGBZ_05	0.925	0.919	0.930	0.928	0.911	0.927	**0.934**	*0.932*
RGBZ_06	**0.923**	0.895	*0.921*	0.908	0.890	0.908	0.913	*0.921*
RGBZ_07	*0.910*	0.892	0.905	*0.910*	*0.910*	0.907	0.904	**0.915**
RGBZ_08	0.949	0.920	**0.951**	0.949	*0.950*	0.948	0.939	**0.951**
RGBZ_09	0.946	0.877	**0.953**	0.943	0.938	0.918	0.902	*0.951*
Average	0.924	0.907	*0.929*	0.926	0.914	0.915	0.890	**0.930**

The first and second best SSIM at each row are shown in bold and italics, respectively

handle artifacts introduced by the cross-boundary scribbles (see Figs. 4, 5, 6, 7, 8, 9, 10, 11, 12j, 13, 14, 15, and 16o), which leads to distortions in synthesized views as shown in Figs. 13, 14, 15, and 16p. ℓ_1 [34] tends to produce a nearly piecewise constant depth-map with sparse structures. Therefore, it generates artifacts when depth discontinuities do not coincide with object boundaries (see purple rectangles of Figs. 4, 5, 6, 7, 8, 9k, 14q, and 16q), which causes distortions in synthesized views (see purple rectangles of Figs. 14r and 16r). The proposed method alleviates the influence of cross-boundary user scribbles successfully and produces high-quality depth-maps (see Figs. 4, 5, 6, 7, 8, 9, 10, 11, and 12l, and 13, 14, 15 and 16s). Therefore, the proposed method can reduce distortions in synthesized views caused by cross-boundary input as shown in Figs. 13, 14, 15 and 16t.

3.3 Experiments without cross-boundary user scribbles

In this section, the user carefully draws on an input image, ensuring that scribbles do not cross object boundaries. In this case, unwanted scribbles are usually inside objects when depth discontinuity occurs. Tables 3 and 4 show the SSIM obtained from different methods on RGBZ and Middlebury datasets, respectively. It can be seen from Table 3 that the proposed method gives the highest average SSIM on RGBZ datasets. As shown in Table 4, both the proposed method and OPT [24] obtain the highest average SSIM on Middlebury datasets. Therefore, the proposed method has comparable performance to the state-of-the-art methods when user scribbles do not cross object boundaries.

4 Conclusion

To remove unwanted input from cross-boundary scribbles in semi-automatic 2D-to-3D conversion, this paper proposes a residual-driven energy function for depth estimation from user input. The residual between the estimation and user-specified depth value will be large at the unwantedly labeled pixel due to inconsistency with its neighbors and be small at expectedly labeled pixel due to consistency with the neighbors. Therefore, the residual can differentiate unwanted scribbles from the user input. The experimental results demonstrate that the proposed method eliminates the depth artifacts caused by cross-boundary scribbles effectively and outperforms existing methods when cross-boundary input is present.

Table 4 SSIM of estimated depth on Middlebury datasets when cross-boundary scribbles are absent

	RW	HGR	NRW	OPT	OCP	SDF	ℓ_1	Proposed
Tsukuba	*0.731*	0.725	0.729	**0.733**	0.711	**0.733**	0.726	*0.731*
Venus	**0.975**	0.970	0.972	*0.974*	0.966	0.972	0.972	**0.975**
Teddy	0.865	0.860	*0.867*	0.866	0.860	0.862	**0.869**	0.865
Cones	*0.904*	0.888	0.902	**0.905**	0.858	0.902	0.892	**0.905**
Average	*0.869*	0.861	0.868	**0.870**	0.849	0.868	0.865	**0.870**

The first and second best SSIM at each row are shown in bold and italics, respectively

Abbreviations
RGBZ: Red, green, blue plus *z*-axis depth; SVM: Support vector machines; RW: Random-walks; GC: Graph-cuts; OCP: Co-occurrence probability; DIBR: Depth image-based rendering; HGR: Hybrid GC and RW; NRW: Nonlocal RW; OPT: Optimization; SSIM: Structural similarity

Acknowledgements
The author would like to thank the editors and anonymous reviewers for their valuable comments.

Funding
This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LY16F010014, and Ningbo Natural Science Foundation under Grant No. 2017A610109.

Availability of data and materials
The author can provide the data and source code.

Authors' contributions
HY designed the research, analyzed the data, then wrote and edited the manuscript. The author read and approved the final manuscript.

Authors' information
Hongxing Yuan is currently an Associate Professor at the School of Electronics and Information Engineering, Ningbo University of Technology, China. He received doctor's degree from University of Science and Technology of China, in 2010. His current research interests include computer vision, 3D video processing, and 2D-to-3D conversion.

Competing interests
The author declares that he has no competing interests.

Publisher's Note

Received: 31 January 2018 Accepted: 17 July 2018
Published online: 02 August 2018

References
1. W Huang, X Cao, K Lu, Q Dai, AC Bovik, Toward naturalistic 2D-to-3D conversion. IEEE Trans. Image Process. **24**(2), 724–733 (2015)
2. T-Y Kuo, Y-C Lo, C-C Lin, in *Proceedings of the IEEE Intl. Conf. on Acoustics, Speech and Signal Process*. 2D-to-3D conversion for single-view image based on camera projection model and dark channel model (IEEE, Piscataway, 2012), pp. 1433–1436
3. Y-K Lai, Y-F Lai, Y-C Chen, An effective hybrid depth-generation algorithm for 2D-to-3D conversion in 3D displays. J. Disp. Technol. **9**(3), 154–161 (2013)
4. H Han, G Lee, J Lee, J Kim, S Lee, A new method to create depth information based on lighting analysis for 2D/3D conversion. J. Cent. South Univ. **20**(10), 2715–2719 (2013)
5. J Lin, X Ji, W Xu, Q Dai, Absolute depth estimation from a single defocused image. IEEE Trans. Image Process. **22**(11), 4545–4550 (2013)
6. C-C Han, H-F Hsiao, Depth estimation and video synthesis for 2D to 3D video conversion. J. Sign. Process. Syst. **76**(1), 33–46 (2014)
7. T-T Tsai, T-W Huang, R-Z Wang, A novel method for 2D-to-3D video conversion based on boundary information. EURSIP J. Image Video Process. **2** (2018). https://link.springer.com/article/10.1186%2Fs13640-017-0239-5
8. AH Somaiya, RK Kulkarni, in *Proceedings of the Intl. Conf. on Signal Process. Image Process. Pattern Recognition (ICSIPR)*. Depth cue selection for 3D television (IEEE, Piscataway, 2013), pp. 14–19
9. F Liu, C Shen, G Lin, I Reid, Learning depth from single monocular images using deep convolutional neural fields. IEEE Trans. on Pattern Anal. Mach. Intell. **38**(10), 2024–2039 (2016)
10. C Godard, OM Aodha, GJ Brostow, in *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. Unsupervised monocular depth estimation with left-right consistency (IEEE, Piscataway, 2017), pp. 6602–6611
11. I Laina, C Rupprecht, V Belagiannis, F Tombari, N Navab, in *Proceedings of the Intl. Conf. on 3D Vision (3DV)*. Deeper depth prediction with fully convolutional residual networks (IEEE, Piscataway, 2016), pp. 239–248
12. J Xie, R Girshick, A Farhadi, in *Proceedings of the European Conf. on Computer Vision (ECCV)*. Deep3D: fully automatic 2D-to-3D video conversion with deep convolutional neural networks (Springer, Berlin, 2016), pp. 842–857
13. A Lopez, E Garces, D Gutierre, in *Proceedings of the Spanish Computer Graphics Conference*. Depth from a single image through user interaction (Wiley, Hoboken, 2014), pp. 1–10
14. R Rzeszutek, R Phan, D Androutsos, in *Proceedings of the ACM Intl. Conf. on Multimedia*. Depth estimation for semi-automatic 2D to 3D conversion (ACM, New York, 2012), pp. 817–820
15. M Guttmann, L Wolf, D Cohen-Or, in *Proceedings of the IEEE Intl. Conf. on Computer Vision (ICCV)*. Semi-automatic stereo extraction from video footage (IEEE, Piscataway, 2009), pp. 136–142
16. D S'ykora, D Sedlacek, S Jinchao, J Dingliana, S Collins, Adding depth to cartoons using sparse depth (in)equalities. Comput. Graph. Forum. **29**(2), 615–623 (2010)
17. R Rzeszutek, R Phan, D Androutsos, in *Proceedings of the IEEE Intl. Conf. on Multimedia & Expo*. Semi-automatic synthetic depth map generation for video using random walks (IEEE, Piscataway, 2011), pp. 1–6
18. R Phan, D Androutsos, Robust semi-automatic depth map generation in unconstrained images and video sequences for 2D to stereoscopic 3D conversion. IEEE Trans. Multimedia. **16**(1), 122–136 (2014)
19. X Xu, L-M Po, K-W Cheung, K-H Ng, in *Proceedings of the IEEE Intl. Conf. on Signal Processing, Communication and Computing (ICSPCC)*. Watershed and random walks based depth estimation for semi-automatic 2D to 3D image conversion (IEEE, Piscataway, 2012), pp. 84–87
20. Z Zhang, C Zhou, Y Wang, W Gao, Interactive stereoscopic video conversion. IEEE Trans. Circuits Syst. Video Technol. **23**(10), 1795–1807 (2013)
21. Q Zeng, W Chen, H Wang, C Tu, D Cohen-or, D Lischinski, B Chen, Hallucinating stereoscopy from a single image. Comput. Graph. Forum. **34**(2), 1–12 (2015)
22. H Yuan, S Wu, P Cheng, P An, S Bao, Nonlocal random walks algorithm for semi-automatic 2D-to-3D image conversion. IEEE Signal Proc. Let. **22**(3), 371–374 (2015)
23. Z Liang, J Shen, in *Proceedings of the IEEE Intl. Conf. on Digital Signal Processing*. Consistent 2D-to-3D video conversion using spatial-temporal nonlocal random walks (IEEE, Piscataway, 2016), pp. 672–675
24. O Wang, M Lang, M Frei, A Hornung, A Smolic, M Gross, in *Proceedings of the Eur. Symp. Sketch-Based Interfaces and Modeling*. StereoBrush: interactive 2D to 3D conversion using discontinous warps (Springer, Berlin, 2011), pp. 47–54
25. A Levin, D Lischinski, Y Weiss, Colorization using optimization. ACM Trans. Graph. **23**(3), 689–694 (2004)
26. S Wu, H Yuan, P An, P Cheng, Semi-automatic 2D-to-3D conversion using soft segmentation constrained edge-aware interpolation. ACTA Electron. Sin. **43**(11), 2218–2224 (2015)
27. J Liao, S Shen, E Eisemann, in *Graph. Interface Conf.* Depth Map Design and Depth-based Effects With a Single Image (ACM, New York, 2017), pp. 57–63
28. M Calemme, P Zanuttigh, S Miiani, M Cagnazzo, B Pesquet-Popescu, in *Proceedings of the IEEE Intl. Conf. on Image Processing*. Depth map coding with elastic contours and 3D surface prediction (IEEE, Piscataway, 2016), pp. 1106–1110
29. K Subr, S Paris, C Soler, J Kautz, Accurate binary image selection from inaccurate user input. Comput. Graph. Forum. **32**(2pt1), 41–50 (2013)
30. J Bai, X Wu, in *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. Error-tolerant scribbles based interactive image segmentation (IEEE, Piscataway, 2014), pp. 392–399
31. C Oh, B Ham, K Sohn, Robust interactive image segmentation using structure-aware labeling. Expert Syst. Appl. **79**, 90–100 (2017)
32. B-W Hong, J-K Koo, H Dirks, M Nurger, in *Proceedings of the German Conf. on Pattern Recognition (GCPR)*. Adaptive regularization in convex composite optimization for variational imaging problems (Springer, Berlin, 2017), pp. 268–280
33. B Ham, M Cho, J Ponce, Robust guided image filtering using nonconvex potentials. IEEE Trans. Pattern. Anal. Mach. Intell. **40**(1), 192–207 (2018)

34. H Yuan, P An, S Wu, Y Zheng, Error-tolerant semi-automatic 2D-to-3D conversion via l1 optimization. Acta Electron. Sin. **46**(2), 447–455 (2018)
35. M Jung, Piecewise-smooth image segmentation models with l1 data-fidelity terms. J. Sci. Comput. **70**(3), 1229–1261 (2017)
36. C Richardt, C Stoll, NA Dodgson, H-P Seidel, C Theobalt, Coherent spatiotemporal filtering, upsampling and rendering of RGBZ videos. Comput. Graph. Forum. **31**(2), 247–256 (2012)
37. D Scharstein, R Szeliski, in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings*. High-accuracy stereo depth maps using structured light (IEEE, Piscataway, 2003), pp. 195–2021
38. Z Wang, AC Bovik, HR Sheikh, EP Simoncelli, Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
39. Y Konno, M Tanaka, M Okutomi, Y Yanagawa, K Kinoshita, M Kawade, in *2016 23rd International Conference on Pattern Recognition (ICPR)*. Depth map upsampling by self-guided residual interpolation (IEEE, Piscataway, 2016), pp. 1394–1399