Depth Superresolution by Transduction

Bumsub Ham, *Member, IEEE*, Dongbo Min, *Member, IEEE*, and Kwanghoon Sohn, *Senior Member, IEEE*

***Abstract*— This paper presents a depth superresolution (SR) method that uses both of a low-resolution (LR) depth image and a high-resolution (HR) intensity image. We formulate depth SR as a graph-based transduction problem. In particular, the HR intensity image is represented as an undirected graph, in which pixels are characterized as vertices, and their relations are encoded as an affinity function. When the vertices initially labeled with certain depth hypotheses (from the LR depth image) are regarded as input queries, all the vertices are scored with respect to the relevances to these queries by a classifying function. Each vertex is then labeled with the depth hypothesis that receives the highest relevance score. We design the classifying function by considering the local and global structures of the HR intensity image. This approach enables us to address a depth bleeding problem that typically appears in current depth SR methods. Furthermore, input queries are assigned in a probabilistic manner, making depth SR robust to noisy depth measurements. We also analyze existing depth SR methods in the context of transduction, and discuss their theoretic relations. Intensive experiments demonstrate the superiority of the proposed method over state-of-the-art methods both qualitatively and quantitatively.**

***Index Terms*— Depth super-resolution, active range sensor, transduction, graph regularization.**

I. Introduction

SENSING an accurate depth image is a fundamental task in computer vision and image processing, and recently its importance has been on the rise in numerous applications, including image-based rendering and 3D object modeling. Besides, the depth image can be used in various ways, to tackle challenging problems, e.g., scene labeling [1], object detection [2], tracking [3], and visual saliency [4]. It is because depth information provides a decisive clue, making several tasks much easier to address, while facilitating more accurate and efficient solutions.

A number of depth sensing methods have been introduced. In a static scene, a laser range scanner or an active illumination with structured light is the best way for capturing an accurate depth image [5], but they are applicable at controlled environments only. Alternatively, the depth image is estimated through a computational approach, i.e., by finding correspondences between images. In the initial phase, its performance was very far from practical usage due to an enormous computational complexity and an unstable estimation quality. The performance of correspondence matching algorithms has been significantly improved, but many challenges still remain, e.g., sensitivity to lighting conditions. An active range sensor such as a time-of-flight (ToF) camera is an alternative to acquiring depth information. The ToF camera captures the distance between a sensor and an object by measuring a phase delay between emitted and received light waves. It provides dense depth measurements at video rate, and can be used in dynamic environments. Nevertheless, an actual application has been impeded by inherent physical limitations of the sensor – the acquired depth image is of relatively low-resolution (LR), and is corrupted by a huge amount of noise.

There have been many attempts to enhance the resolution and accuracy of the depth image obtained from the ToF camera. Previous works on depth super-solution (SR) can be generally categorized into two groups (depth guided methods and intensity guided methods), according to whether or not a high-resolution (HR) intensity image is used. The first approach stems from classical SR techniques: a spatial resolution is enhanced by combining multiple LR depth images [6]. To avoid using multiple images, one can use a dictionary-based approach using registered and paired HR/LR depth images [7]. Although this approach works well even in case that multiple LR depth images are not available, tens of thousands of ground truth depth images are required to build a training data set. The second approach exploits a HR intensity image as a depth cue. It assumes that there exist co-occurrence statistics between depth and intensity discontinuities [8]–[23]. The additional intensity image helps align depth boundaries to intensity edges, but this may cause the textures in the intensity image to be transferred to the depth image. More importantly, this type of method using the intensity image prevents a sharp depth transition at depth discontinuities [7]. Such a drawback, *a depth bleeding problem*, can result in jagged artifacts in scene reconstruction [7].

This paper presents a sensor fusion approach to enhancing the spatial resolution of the depth image. Our approach belongs to the second category: LR depth and HR intensity images are jointly used to reconstruct a fine quality depth image. The main contributions of this work are as follows:

- We view depth SR as a labeling task, and formulate it as a graph-based transduction problem. This approach avoids the depth bleeding problem by considering the local and global structures of the intensity image.

Manuscript received August 4, 2014; revised December 11, 2014; accepted February 7, 2015. Date of publication February 24, 2015; date of current version March 6, 2015. This work was supported in part by the National Research Foundation of Korea within the Korea Government through the Ministry of Science, ICT and Future Planning under Grant NRF-2013R1A2A2A01068338 and in part by the European Research Council, VideoWorld. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chang-Su Kim.

B. Ham is with the School of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, Korea, and also with the WILLOW Team, Institut National de Recherche en Informatique et en Automatique, École Normale Supérieure, (CNRS/ENS/INRIA UMR 8548), Paris 75013, France (e-mail: bumsub.ham@inria.fr).

D. Min is with the Department of Computer Science and Engineering, Chungnam National University, Daejeon 305-764, Korea (e-mail: dbmin99@gmail.com).

K. Sohn is with the School of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, Korea (e-mail: khsohn@yonsei.ac.kr).

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

Digital Object Identifier 10.1109/TIP.2015.2405342

- We design an indication matrix that leverages the property of a noisy depth image, making depth SR robust to incorrect depth measurements.
- We present a comprehensive review of existing depth SR methods, analyze them in the context of transduction, and discuss their theoretic relations.

The remainder of this paper is organized as follows. Section II describes related works for depth SR and transduction. Section III analyzes current depth SR methods and describes the proposed method in detail. An extensive analysis of experimental results is then presented in Section IV. Finally, conclusion and discussion are given in Section V.

II. Related Work

In this section, we review recent works related to the proposed method: intensity guided depth SR and transduction.

A. Intensity Guided Depth SR

Intensity guided depth SR methods can be further classified into three categories: filtering-based methods, optimization-based methods, and cost aggregation-based methods.

1) Filtering-Based Methods: This type of method applies a filtering operation to the depth image under guidance of the HR intensity image. That is, the LR depth image is regularized with a weight function that is calculated from the HR intensity image. This indicates that conventional filtering schemes can be utilized to depth SR. It is worth mentioning that this joint filtering scheme has been employed in a variety of image processing tasks, e.g., colorization and tone mapping [8]. One of the seminal works is the joint bilateral upsampling (JBU) [8] that applies the bilateral filter [24] to a target signal under guidance of another signal. Since then, many filtering-based methods have been proposed, e.g., using fast edge-preserving filters such as guided filter (GF) [9] and geodesic filter [20].

2) Optimization-Based Methods: Optimization-based methods enhance the spatial resolution of the depth image by minimizing an objective function that commonly consists of data and regularization terms. Diebel and Thrun used a MRF framework, where the regularization term is penalized in accordance with texture derivatives [13]. The weighted least square (WLS) framework has been popularly utilized by formulating the regularization term accordingly [16], [17]. For examples, Park *et al.* modeled the regularization term based on segmentation and edge saliency. A nonlocal means regularization term was additionally involved, which results in protecting thin structures. Ham *et al.* regarded depth SR as a label propagation task, and solved it with a random walk with restart (RWR) framework [22]. Ferstl *et al.* formulated depth SR as a convex optimization problem using a higher order regularization term, enforcing a piecewise affine solution [18]. Note that these optimization-based methods can be thought of as one type of filtering-based methods where the global structure of the HR intensity image is considered [9].

3) Cost Aggregation-Based Methods: Cost aggregation-based methods perform depth SR in a label domain, not an image domain itself. That is, these methods construct a 3D cost volume using a current depth value, and then perform an adaptive smoothing independently for each 2D cost layer, similar to the cost aggregation step in a local correspondence matching algorithm [19]. Yang *et al.* [11], [12] aggregated each of 2D cost slice by using the bilateral filter [24]. Min *et al.* proposed the weighted mode filter (WModF) that computes a mode from a joint histogram computed with a set of input measurements within a local window, improving the depth accuracy significantly [14]. The weighted median filter (WMedF) [19] and nonlocal weighted median filter (NMedF) [23] were recently employed to depth SR, which addresses a limitation of using a local mean filter, e.g., the bilateral filter [24].

B. Transduction

Transduction aims to infer a classifying function for labeled and unlabeled data, from which the labels of the unlabeled data are predicted. It is highly related to manifold ranking in that manifold ranking is an extreme case of transduction, where only a single label is available [25]. Transduction or manifold ranking has been employed to numerous applications, e.g., web page ranking [26], clustering [27], image segmentation [28], [29], image retrieval [30], and visual saliency [31]. Recently, transduction was also applied to an image interpolation task [32]. In this work, intensity values from a LR input image are considered as labeled data, and the interpolation is performed using a transductive *regression*. Similar to WLS [16], [17] and RWR [22], an objective function defined on an intensity domain is minimized to estimate *continuous* labels (intensity values). In contrast, we regard depth SR as a transductive *classification*, i.e., a discrete labeling task. The objective function is defined for each depth hypothesis, and is minimized to find the classifying function. This function labels each pixel in the HR intensity image with a *discrete* label among a set of depth hypotheses. Moreover, we introduce a novel indication matrix – a confidence matrix, which makes transduction robust to incorrect initial label assignments. In [32], the initial labels of LR samples are fixed, and thus transduction is not robust to noisy data. To the best of our knowledge, our approach first attempts to explicitly employ the transductive classification in depth SR.

III. Depth SR by Transduction

In this section, we first analyze the problems of conventional intensity guided depth SR methods, and then describe the proposed method in detail.

A. Analysis of Current Depth SR Methods

1) Limitations: Fig. 1 shows a visual comparison of intensity guided depth SR methods: 1) filtering-based methods, (c) GF [9], 2) optimization-based methods, (d) RWR [22] and (e) anisotropic total generalized variation (ATGV) [18], and 3) cost aggregation-based methods, (f) NMedF [23], (g) WModF [14], and (h) the proposed method. Although they provide visually pleasing results, there still remain some matters to iron out, so-called depth blurring and depth bleeding artifacts.

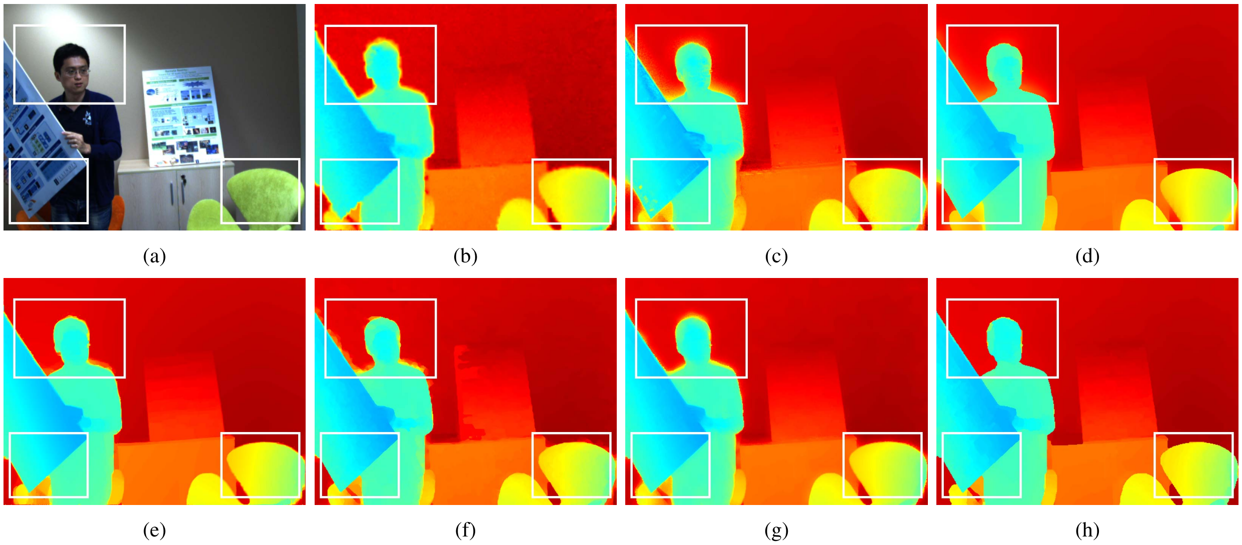


Fig. 1. Visual comparison of intensity guided depth SR methods: given (a) the HR intensity image, (b) the LR depth image is upsampled by 1) filtering-based methods, (c) GF [9], 2) optimization-based methods, (d) RWR [22] and (e) ATGV [18], and 3) cost aggregation-based methods, (f) NMedF [23], (g) WModF [14], and (h) the proposed method. We formulate depth SR as a discrete labeling problem similar to cost aggregation-based methods. In contrast to conventional cost aggregation-based methods (e.g., WModF), both local and global structures of the intensity image are taken into account in the proposed method, providing convincing results without depth bleeding artifacts at depth boundaries. Note that optimization-based methods (e.g., RWR and ATGV) also consider the global structure of the intensity image, but they regularize the LR depth image itself; thus, depth bleeding artifacts are still observed around object boundaries.

We differentiate the depth bleeding from the depth blurring, although they show similar behaviors. The depth blurring refers to over-smoothing of depth discontinuities. It corresponds to halo artifacts that appear in local filtering methods [9], and most filtering-based depth SR methods suffer from depth blurring artifacts as shown in Fig. 1 (c). This problem could be alleviated by choosing an appropriate kernel function. Alternatively, we can reduce depth blurring artifacts by using a global optimization scheme as in optimization-based methods [9], or by performing depth SR in a label domain as in cost aggregation-based methods. The depth bleeding refers to the leak of depth information, and it is caused by blurry boundaries in the intensity image.[1] The depth bleeding artifacts may be reduced by considering depth values as discrete variables, not continuous variables (as in filtering-based and optimization-based methods) and formulating depth SR as a discrete labeling problem, like cost aggregation-based methods. Cost aggregation-based methods give relatively good results, but depth bleeding artifacts are still observed around object boundaries as shown in Fig. 1 (f) and (g).

2) Motivation: Let us consider a toy example for depth SR as shown in Fig. 2 (a). The vertices correspond to 5D feature vectors of the HR intensity image consisting of spatial location and color. Let us assume that the vertices in the shapes of circle and moon belong to two different depth hypotheses, i.e., a red circle and a blue cross. The task is to label each vertex in the intensity image with given sparse depth hypotheses. Most cost aggregation-based methods infer the depth hypothesis by considering the local structure of the intensity image only, through a weighted averaging filter [11], a weighted median filter [19], or a weighted mode filter [14].

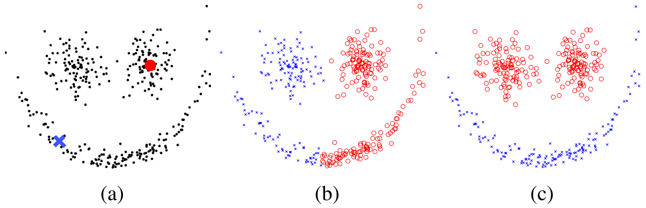


Fig. 2. Toy example for depth SR. (a) 5D feature vectors of the HR intensity image and sparse depth hypotheses, described by a red circle and a blue cross, respectively, (b) labeling guided by a local structure only, and (c) labeling guided by local and global structures. Let us assume that the vertices in the shapes of circle and moon belong to two different depth hypotheses. The task is to infer the depth hypothesis of each vertex in the intensity image with given sparse depth hypotheses. The current cost aggregation-based methods consider the local structure of the intensity image only. Such a local aggregation strategy is insufficient to capture the underlying structure of the intensity image. This causes depth bleeding problems, i.e., an inaccurate depth value is penetrated into neighboring regions with a different depth hypothesis. In contrast to this, intrinsic structures can be captured by exploiting the global structure of the intensity image. (Best viewed in color).

Such a local aggregation strategy is insufficient to capture the underlying structure of the intensity image, resulting in depth bleeding artifacts[2]; an inaccurate depth value is thus penetrated into neighboring regions with a different depth hypothesis, as shown in Fig. 2 (b). This type of problem has been well studied in the context of transduction [33], and has been addressed by designing a suitable classifying function with reference to the underlying structure.

Inspired by this literature, we formulate depth SR as a graph-based transduction problem. Namely, it is formulated as a discrete labeling task similar to cost aggregation-based methods, but we exploit both local and global structures of

[1] The edges in the intensity image may be smoothed when captured due to several factors, e.g., aerial scattering.

[2] An efficient nonlocal cost aggregation method was recently proposed using a tree structure, and it was applied to NMedF [23]. NMedF uses a global aggregation, but the tree structure may fail to fully capture local image attributes. This causes depth leakage artifacts as shown in 1 (f).

the intensity image, mitigating depth bleeding artifacts, as in Fig. 2 (c). The vertices (pixels) in the HR intensity image are labeled with an assumption that nearby vertices are likely to have similar depth hypotheses, and vertices on the same structure are likely to have similar depth hypotheses [33]. It is worth pointing out that although optimization-based methods also consider the global structure of the intensity image, there still exist depth bleeding artifacts, as shown in Fig. 1 (d) and (e). In contrast, our labeling approach, guided by both local and global structures of the intensity image, achieves the best upsampling results, as shown in Fig. 1 (h).

B. Transduction for Depth SR

1) Problem Statements: Let us assume that HR intensity and LR depth images are registered with a size of n and p $(< n)$, respectively. Specifically, the LR depth image is initially warped to the HR intensity image, such that p points in the HR intensity image are labeled with reference to the depth values of the LR depth image, and $(n - p)$ points in the HR intensity image remain invalid.

Concretely, given a set of vertices in the HR intensity image, $\mathcal{X} = (\mathcal{X}_l, \mathcal{X}_u) = \{x_1, ..., x_p, x_{p+1}, ..., x_n\}$, some vertices x_i $(i \leq p)$ in $\mathcal{X}_l$ are labeled as $y_i \in \mathcal{L}$ where $\mathcal{L} = \{1, ..., c\}$, i.e., a set of depth hypotheses with a maximum depth value c. We aim to predict the depth hypotheses (labels) of unlabeled vertices x_j $(p + 1 \leq j \leq n)$ in $\mathcal{X}_u$ with labeled ones in $\mathcal{X}_l$. This can be achieved by measuring the relevance of each unlabeled vertex to the labeled vertices, and by labeling it with the depth hypothesis that receives the highest relevance score [25]. Depth SR then boils down to designing a scoring function (or a classifying function).

2) Graph Construction: Let us represent the HR intensity image as an undirect graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of vertices on the data set $\mathcal{X}$ and $\mathcal{E}$ is a set of links. An undirected edge $E_{ij} \in \mathcal{E}$ exists if two vertices $x_i, x_j \in \mathcal{V}$ are adjacent. In our work, a local 4-neighborhood system is used. The local structure is encoded as an affinity function $W_{ij} \in \mathcal{W}$, where $\mathcal{W}$ is a set of $n \times n$ matrices with nonnegative elements, and it is computed based on the distance between adjacent vertices as follows.

$$W_{ij} = \exp\left\{-d^2(g_i, g_j)\right\} \quad x_i, x_j \in \mathcal{V}, \tag{1}$$

where $d(g_i, g_j) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a metric that measures the similarity of features between vertices. Various features including spatial location, intensities, and textures can be used to represent the distinctiveness of vertices. For simplicity, we use Euclidean distance between color values as follows.

$$d(g_i, g_j) = \left\|g_i - g_j\right\|_2 / \sqrt{2}\sigma, \tag{2}$$

where σ is a range bandwidth. g_i represents the color values of a vertex x_i in the *Lab* space, and $\|\cdot\|_2$ denotes l_2 norm. Then, affinity and corresponding degree matrices of the graph $\mathcal{G}$ can be described as $\mathbf{W} = [W_{ij}]_{n\times n}$ and $\mathbf{D} = diag\{D_1, ..., D_n\}$ where $D_i = \sum_{j=1}^{n} W_{ij}$.

3) Confidence Matrix: The $n \times c$ indication matrix $\mathbf{Y} \in \mathcal{F}$ is introduced to identify initial label assignments where $\mathcal{F}$ represents a set of $n \times c$ matrices with nonnegative elements. The simplest way of constructing the indication matrix $\mathbf{Y} = [Y_1^{\mathrm{T}}, ..., Y_n^{\mathrm{T}}]^{\mathrm{T}}$ is to set $Y_{ij} = 1$ if a vertex x_i is labeled as $y_i = j$ where $j \in \mathcal{L}$, and $Y_{ij} = 0$ otherwise [33]. That is, the indication matrix has valid values only when the vertex x_i is labeled as a certain depth hypothesis $j \in \mathcal{L}$ from the initial sparse depth measurements. This works well in case that all the initial label assignments are correct, but the depth image captured by the active range sensor is typically contaminated by sensory noise.

Considering the potential errors of the initial depth image, we design the indication matrix in a probabilistic way, such that each element in the matrix can be varied according to the noisy level in the initial depth image, not being fixed to the binary value (1 or 0). We call this matrix as a *confidence* matrix. When a vertex x_i is labeled with a certain depth hypothesis $y_i = j$ from the initial sparse depth measurements, the confidence matrix is defined by assigning probability values to remaining *probable* depth hypotheses as follows:

$$Y_{ik} = \begin{cases} \max(1 - \delta\,|j - k|\,, 0) & y_{i\leq p} = j,\ j - \eta c \leq k \leq j + \eta c \\ 0 \quad otherwise, & \end{cases} \tag{3}$$

where a constant δ controls the shape of the function, and η is a parameter for the noise variance. It can be seen that when η is zero, the confidence matrix in (3) becomes conventional indication matrix, i.e., $Y_{ij} = 1$ only when a vertex x_i is labeled as $y_i = j$.

4) Classifying Function: Let $\mathbf{F} = [F_1^{\mathrm{T}}, ..., F_n^{\mathrm{T}}]^{\mathrm{T}} \in \mathcal{F}$ be a $n \times c$ matrix that plays a role of classifying a vertex x_i on the data set $\mathcal{X}$, by which each vertex x_i is labeled as y_i, in such a fashion $y_i = \arg\max_{j\leq c} F_{ij}$. That is, the hypothesis of the highest rank among a set of depth hypotheses is selected as a depth value at the vertex x_i.

The classifying function $\mathbf{F}$ generally meets the following conditions [33]: first, a good classifying function refrains from alternating labels from initially assigned ones. Second, it should not change too much between nearby vertices, and sufficiently smooth with accordance to the underlying structure captured by the intensity image. Finding the classification function can be formalized by energy minimization, and the first and second constraints are modeled as data and regularization terms, respectively, as below:

$$\mathcal{Q}_1(\mathbf{F}) = \mu \sum_{i=1}^{n} \|F_i - Y_i\|^2 + \sum_{i,j=1}^{n} \frac{W_{ij}}{D_i} \left\|F_i - F_j\right\|^2, \tag{4}$$

where μ $(0 < \mu < 1)$ is a regularization parameter that balances data and regularization terms. There exist many ways to design an energy function for transduction, and a comprehensive review can be found at [34] in the context of image retrieval. It is worth noting that this approach is different from optimization-based methods (see [22]), in that energy minimization is done with respect to the classifying function $\mathbf{F}$ in a label domain, whereas optimization-based methods minimize an energy function with respect to the depth value itself in an image domain.

In (4), it is assumed that all the initial labels equally contribute to the classifying function. Recall that, in the previous section, the initial labels are identified by the confidence matrix to authorize the robustness to noisy data. Similarly, let us suppose that we have prior knowledge about the confidences of initial labels. The different scores can then be assigned to the initial labels by the analogy of their respective confidences [25]. We assume that a distinct region such as edges and corners give relatively more reliable information than a homogeneous region. This assumption is modeled by using a degree value D_i, since it implicitly describes a property of the neighborhood of a vertex x_i, e.g., the degree value of a distinct region is smaller than that of a homogeneous region. This observation is encoded in (4), leading to a new energy function:

$$\mathcal{Q}_2(\mathbf{F}) = \mu \sum_{i=1}^{n} \left\| F_i - D_i^m Y_i \right\|^2 + \sum_{i,j=1}^{n} \frac{W_{ij}}{D_i} \left\| F_i - F_j \right\|^2. \quad (5)$$

The constant m determines the degree of confidences on initial labels, and it is set to $-1/2$ in our work, enabling a distinct region to have a relatively higher confidence than a homogeneous region.

5) Numerical Solution: The classifying function can be found by minimizing the energy function $\mathcal{Q}_2(\mathbf{F})$ within a domain $\mathcal{F}$:

$$\mathbf{F}^\star = \arg\min_{\mathbf{F}\in\mathcal{F}} \mathcal{Q}_2(\mathbf{F}). \quad (6)$$

In matrix/vector form, the energy function as in (5) can be rewritten as

$$\mathcal{Q}_2(\mathbf{F}) = \mu(\mathbf{F} - \mathbf{D}^{-\frac{1}{2}}\mathbf{Y})^{\mathrm{T}}(\mathbf{F} - \mathbf{D}^{-\frac{1}{2}}\mathbf{Y}) + \mathbf{F}^{\mathrm{T}}(\mathbf{I} - \mathbf{D}^{-1}\mathbf{W})\mathbf{F}, \quad (7)$$

where $\mathbf{I}$ represent a $n \times n$ identity matrix. Since (7) is linear and strictly convex, a global minimum is guaranteed, and it can be obtained by differentiating $\mathcal{Q}_2(\mathbf{F})$ with respect to $\mathbf{F}$ as follows.

$$\left.\frac{\partial \mathcal{Q}_2(\mathbf{F})}{\partial \mathbf{F}}\right|_{\mathbf{F}=\mathbf{F}^\star} = \mu\mathbf{F} - \mu\mathbf{D}^{-\frac{1}{2}}\mathbf{Y} + \mathbf{F} - \mathbf{D}^{-1}\mathbf{W}\mathbf{F} = 0. \quad (8)$$

It is rewritten as,

$$\mathbf{F}^\star + \frac{\mu}{1+\mu}\mathbf{D}^{-\frac{1}{2}}\mathbf{Y} - \frac{1}{1+\mu}\mathbf{D}^{-1}\mathbf{W}\mathbf{F}^\star = 0. \quad (9)$$

When α is substituted with $1/(1+\mu)$ in (9), then

$$(\mathbf{I} - \alpha\mathbf{D}^{-1}\mathbf{W})\mathbf{F}^\star = (1-\alpha)\mathbf{D}^{-\frac{1}{2}}\mathbf{Y}. \quad (10)$$

The following relation can then be derived, from which the relevance scores between labeled and unlabeled vertices are obtained.

$$\mathbf{F}^\star \propto (\mathbf{I} - \alpha\mathbf{D}^{-1}\mathbf{W})^{-1}\mathbf{D}^{-\frac{1}{2}}\mathbf{Y}. \quad (11)$$

Note that the inverse matrix $(\mathbf{I} - \alpha\mathbf{D}^{-1}\mathbf{W})^{-1}\mathbf{D}^{-\frac{1}{2}}$ encodes the local and global structures of the intensity image [9], [33]. Note also that this matrix can be considered as an implicit affinity (kernel) function [9]. Namely, (11) is equivalent to implicitly filtering each column of the confidence matrix $\mathbf{Y}$ by the inverse matrix $(\mathbf{I} - \alpha\mathbf{D}^{-1}\mathbf{W})^{-1}\mathbf{D}^{-\frac{1}{2}}$.

Algorithm 1 Depth SR by Transduction

```
  Input : HR intensity and LR depth images
  Output : HR depth image
1 Construct the affinity matrix $\mathbf{W} = [W_{ij}]_{n\times n}$ on the graph $\mathcal{G}$,
  according to (1) and (2).
2 Calculate the stochastic matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$.
  $\mathbf{D} = diag\{D_1, ..., D_n\}$ where $D_i = \sum_{j=1}^{n} W_{ij}$.
3 Construct the confidence matrix $\mathbf{Y} = [Y_1^{\mathrm{T}}, ..., Y_n^{\mathrm{T}}]^{\mathrm{T}}$ as in (3).
4 Compute the classifying function $\mathbf{F}^\star = [F_1^{\star\mathrm{T}}, ..., F_n^{\star\mathrm{T}}]^{\mathrm{T}}$ as in (11).
5 Label each vertex $x_i$ as $y_i \in \mathcal{L}$ by $y_i = \arg\max_{j\leq c} F_{ij}$.
```

The proposed depth SR method is summarized in Algorithm 1.

6) Computational Complexity: The main computational cost of the proposed method comes from solving the linear system of (11), which involves matrix inversion and multiplication. The $n \times c$ confidence matrix $\mathbf{Y}$ can be decomposed into independent $n \times 1$ vectors, and this enables splitting the linear system of (11) into c linear equations. For solving each linear equation, general dense direct solvers such as Gaussian elimination and LU factorization accompany $\mathcal{O}(n^3)$ complexity. Since we use a 4-neighborhood system, several sparse direct solvers can be used to solve our sparse liner equations. We use the direct solver in MATLAB, by which each sparse linear equation can be solved with $\mathcal{O}(n^{1.5})$ complexity [35]. Accordingly, the total cost of solving the linear system of (11) is $\mathcal{O}(cn^{1.5})$.[3] It is computationally expensive compared to other methods, e.g., $\mathcal{O}(n)$ in GF [9] and $\mathcal{O}(cn)$ in WMedF [19]. Several sparse solvers have been recently proposed [35], [36], which are highly efficient and have a linear $\mathcal{O}(n)$ complexity; thus, the complexity of solving the linear system of (11) can be reduced to $\mathcal{O}(cn)$ by using these solvers. Since the c linear equations can be solved in parallel, the linear system of (11) can be further efficiently solved by using a parallel processing unit.

7) Links With Graph Cut and Gaussian Random Fields:

a) Graph cut: When the classifying function takes discrete values, e.g., $\mathcal{L}^n = \{1, ..., c\}^n$, not continuous (probability) values, minimizing the energy function of (4) becomes the following combinatorial problem:

$$\mathbf{F}^\star = \mathop{\arg\min}_{\substack{\mathbf{F}\in\mathcal{L}^n \\ F_i = Y_i \,\mathrm{on}\, X_l}} \mathbf{F}^{\mathrm{T}}\Delta\mathbf{F}. \quad (12)$$

Δ is the random walk Laplacian:

$$\Delta = \mathbf{D}^{-\frac{1}{2}}(\mathbf{I} - \mathbf{S})\mathbf{D}^{\frac{1}{2}}, \quad (13)$$

where $\mathbf{S} = \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$. This combinatorial optimization problem can be efficiently solved by multi-label graph-cut algorithm [37].

b) Gaussian random fields: When the regularization parameter μ approaches to infinity, the objective function of (4) reduces to the Gaussian random fields [38],

[3] Current un-optimized MATLAB implementation on 2.5 GHz CPU takes about 3.4 seconds to upsample a depth image of size 384 × 288 when the number of depth hypotheses is 16.

TABLE I

TRANSDUCTIVE VIEWPOINT OF DEPTH SR METHODS

Confidence matrix	Robustness	Methods
$Y_{ij} = \begin{cases} 1 & y_{i \le p} = j \\ 0 & otherwise \end{cases}$	✗	JBU [8], GF [9], RWR [22], WMedF [19], NMedF [23]
$Y_{ik} = \begin{cases} \max\left(1 - \delta \lvert j-k \rvert, 0\right) & y_{i \le p} = j, 1 \le k \le c \\ 0 & otherwise \end{cases}$	✓	3D JBU [11]
$Y_{ik} = \begin{cases} \exp\left\{-\delta \lvert j-k \rvert\right\} & y_{i \le p} = j, 1 \le k \le c \\ 0 & otherwise \end{cases}$	✓	WModF [14]
$Y_{ik} = \begin{cases} \max(1 - \delta \lvert j-k \rvert, 0) & y_{i \le p} = j, j - \eta c \le k \le j + \eta c \\ 0 & otherwise \end{cases}$	✓	Proposed method
Kernel matrix	**Global structure**	**Methods**
$H_{ij} = \exp\left\{-d^2(g_i, g_j) - d_s^2(i, j)\right\}$	✗	JBU [8], 3D JBU [11], WModF [14]
$H_{ij} = \frac{1}{\lvert\mathcal{N}\rvert^2} \sum_{l:(i,j)\in\mathcal{N}} \Big(1 + (g_i - \mu_l)(\Sigma_l + \tau\mathbf{I})^{-1}(g_j - \mu_l)\Big)$	✗	GF [9], WMedF [19]
$H_{ij} = \exp\left\{-d_{L1}(g_i, g_j)\right\}$	✓	NMedF [23]
$H_{ij} = (\mathbf{I} - \alpha\mathbf{D}^{-1}\mathbf{W})^{-1}$	✓	RWR [22]
$H_{ij} = (\mathbf{I} - \alpha\mathbf{D}^{-1}\mathbf{W})^{-1}\mathbf{D}^{-\frac{1}{2}}$	✓	Proposed method
Decision rule	**Type**	**Methods**
$y_i = \frac{\sum_{\mathcal{L}} jF_{ij}}{\sum_{\mathcal{L}} F_{ij}}$	Mean	JBU [8], GF [9], RWR [22]
$y_i = median_{j \le c} F_{ij}$	Median	WMedF [19], NMedF [23]
$y_i = \arg\max_{j \le c} F_{ij}$	Mode	3D JBU [11], WModF [14], Proposed method

keeping the smoothness term only. The classifying function can be found by forcing $F_i = Y_i$ for $x_i \in \mathcal{X}_l$ as follows:

$$\mathbf{F}^{\star} = \underset{\substack{\mathbf{F}\in\mathcal{F} \\ F_i = Y_i \text{on} \mathcal{X}_l}}{\arg\min} \ \mathcal{Q}_1(\mathbf{F}). \tag{14}$$

Note that the optimal classifying function $\mathbf{F}^{\star}$ is a harmonic function.

C. Transductive Viewpoint of Depth SR

The proposed method casts depth SR as a discrete labeling problem, which is closely related to the cost aggregation-based methods. It is also associated with the filtering-based and optimization-based methods, in that kernel-based regularization is employed. In the context of transduction, we investigate intensity guided depth SR methods with three steps: designing confidence and kernel matrices, and determining labels. For example, our method labels each vertex x_i as $y_i = \arg\max_{j \le c} F_{ij}$, with the kernel matrix $\mathbf{H} = (\mathbf{I} - \alpha\mathbf{D}^{-1}\mathbf{W})^{-1}\mathbf{D}^{-\frac{1}{2}}$ and the confidence matrix $\mathbf{Y}$ of (3). Here, the classifying function can be represented as a product of kernel and confidence matrices, i.e., $\mathbf{F} = \mathbf{HY}$. Table I shows a classification of depth SR methods from the viewpoint of transduction.

1) 3D JBU: The 3D JBU [11] defines the confidence matrix in a manner similar to our method as

$$Y_{ik} = \begin{cases} \max\left(1 - \delta\,|j-k|, 0\right) & y_{i \le p} = j, 1 \le k \le c \\ 0 & otherwise, \end{cases} \tag{15}$$

and decides each depth hypothesis by $y_i = \arg\max_{j \le c} F_{ij}$. The kernel matrix is designed by the affinity function used in the bilateral filter [24] as in (16). This matrix considers the local structure of the intensity image only, which leads to depth bleeding artifacts.

$$H_{ij} = \exp\left\{-d^2(g_i, g_j) - d_s^2(i, j)\right\} \quad x_i, x_j \in \mathcal{V}, \tag{16}$$

where

$$d_s(i, j) = \frac{\|i - j\|_2}{\sqrt{2}\sigma_s}. \tag{17}$$

σ_S is a spatial bandwidth.

2) NMedF and WMedF: The confidence matrix of NMedF [23] and WMedF [19] are allocated as in (18).

$$Y_{ij} = \begin{cases} 1 & y_{i \le p} = j \\ 0 & otherwise. \end{cases} \tag{18}$$

Although the performance of these methods may be altered according to the quality of the initial depth image, NMedF [23] and WMedF [19] are robust to noisy data. The main reason is that they determine the depth hypothesis by selecting a median value, i.e., $y_i = median_{j \le c} F_{ij}$, and the median value – optimal with respect to the L_1 norm minimization – is generally robust to outliers. The kernel matrix of NMedF [23] is defined with the local affinity function as in (19), but it considers a nonlocal neighborhood by connecting distant pixels.

$$H_{ij} = \exp\left\{-d_{L1}(g_i, g_j)\right\} \quad x_i, x_j \in \mathcal{V}, \tag{19}$$

where

$$d_{L1}(g_i, g_j) = \frac{\left\|g_i - g_j\right\|_1}{\sigma}. \tag{20}$$

The kernel matrix of WMedF [19] is defined with the affinity function from the local filter, e.g., GF [9] as in (21), which causes depth bleeding artifacts.

$$H_{ij} = \frac{1}{|\mathcal{N}|^2} \sum_{l:(i,j)\in\mathcal{N}} \Big(1 + (g_i - \mu_l)(\Sigma_l + \tau\mathbf{I})^{-1}(g_j - \mu_l)\Big), \tag{21}$$

where $\mathcal{N}$ is a neighborhood of the vertex l, and $|\mathcal{N}|$ is the number of vertices in $\mathcal{N}$. Σ_l is a 3×3 covariance matrix, and μ_l is a 3×1 mean vector of g_l in $\mathcal{N}$. The constant τ is

TABLE II

OBJECTIVE EVALUATION OF DEPTH SR METHODS ON THE MIDDLEBURY TEST BED [39]

Sequence	*Tsukuba*		*Venus*		*Teddy*		*Cones*		*Art*	*Books*	*Dolls*	*Laundry*	*Moebius*	*Reindeer*
Method	*all*	*disc*	*all*	*disc*	*all*	*disc*	*all*	*disc*	*all*	*all*	*all*	*all*	*all*	*all*
Bilinear interpolation	10.40	46.30	3.29	37.10	11.80	35.30	14.60	35.80	34.59	13.07	14.59	22.85	17.15	17.48
(**F**) GF [9]	9.87	43.20	2.74	26.50	15.50	37.50	15.50	34.40	45.31	19.14	19.64	26.89	22.22	21.70
(**O**) RWR [22]	9.63	39.50	1.36	12.30	11.40	27.50	11.70	25.40	41.07	17.67	18.36	21.74	21.28	25.71
(**O**) ATGV [18]	5.40	23.90	1.31	14.40	9.82	29.00	10.20	23.60	23.59	13.14	16.14	15.12	18.26	10.42
(**C**) WMedF [19]	6.14	28.00	1.03	10.10	7.88	22.20	8.10	19.20	22.13	8.67	9.58	14.26	10.52	10.04
(**C**) WModF [14]	4.35	20.20	0.61	5.73	9.51	23.70	9.43	19.20	12.39	10.68	11.48	9.91	8.60	12.63
(**C**) NMedF [23]	3.18	10.80	0.47	3.26	6.65	16.20	4.27	7.90	9.71	9.80	9.21	7.42	7.31	11.07
(**C**) Proposed method	1.85	8.80	0.42	4.49	5.61	14.40	3.58	7.85	8.60	8.15	8.07	6.77	6.47	6.55

a regularization parameter that determines the smoothness of the kernel matrix. Please refer to [9] for more details.

3) WModF: The WModF [14] defines the confidence matrix as in (22), endowing with the robustness to noisy data.

$$Y_{ik} = \begin{cases}\exp\{-\delta|j-k|\} & y_{i\le p} = j, 1 \le k \le c \\ 0 & otherwise.\end{cases} \tag{22}$$

The depth hypothesis is determined in such a fashion $y_i = \arg\max_{j\le c} F_{ij}$; thus, the resulting depth image is more sharper than that determined by the mean operation that is widely used in the filtering-based and optimization-based methods as in (24). However, depth bleeding artifacts are still observed as the kernel matrix used in WModF as in (16) considers the local structure of the intensity image only.

4) RWR: The RWR [22] constructs the confidence matrix as in (18).[4] Nevertheless, its performance does not vary depending on the noise level due to the kernel matrix used in RWR:

$$H_{ij} = (\mathbf{I} - \alpha\mathbf{D}^{-1}\mathbf{W})^{-1}. \tag{23}$$

This method determines a depth value through an adaptive summation over depth hypotheses with the corresponding classifying function as follows.

$$y_i = \frac{\sum_{\mathcal{L}} jF_{ij}}{\sum_{\mathcal{L}} F_{ij}}. \tag{24}$$

This type of decision rule based on L_2 norm minimization alleviates the influence of noisy data.

5) JBU and GF: The JBU [8] and GF [9] determine depth values by an adaptive summation as in (24), and thus they are robust against noisy data. Both methods use the confidence matrix as in (18), and differ only in designing kernel matrices; (16) for JBU and (21) for GF. Thus, the global structure of the intensity image is not taken into account, resulting in depth bleeding artifacts.

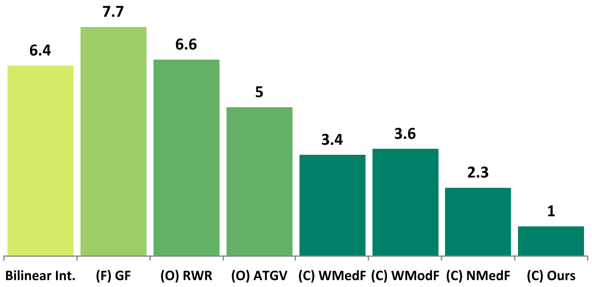


Fig. 3. Average rank of each method according to the error rate in Table II. Cost aggregation-based methods tend to have a better performance than other methods, among which the proposed method gives the superior performance.

IV. EVALUATION

We analyze the performance of our method through various experiments with both synthetic and real-world examples. The Middlebury benchmark data set [39] provides HR intensity images and ground truth depth images. We can synthesize LR depth images by downsampling ground truth depth images, and evaluate the proposed method quantitatively. The experiment has also been conducted with depth images acquired from the ToF camera.

A. Experimental Environments

The proposed method requires four parameters – σ for the graph construction, δ and η for the confidence matrix, and α for the classifying function, and they are set to 60, 0.01, $10/c$, and 0.999, respectively. All the parameters are fixed in experiments, unless otherwise specified. To demonstrate the performance, the bad matching error (%) [40] is measured, i.e., the percentage of erroneous pixels, as follows.

$$\mathcal{B} = \frac{1}{n}\sum(|\mathcal{D}_C - \mathcal{D}_T| > \varepsilon), \tag{25}$$

where ε is a depth error tolerance. $\mathcal{D}_C$ and $\mathcal{D}_T$ represent upsampled depth and ground truth depth images, respectively.

We compare the proposed method with state-of-the-art methods including GF [9], RWR [22], ATGV [18], WMedF [19], WModF [14], and NMedF [23]. As described earlier, these methods are classified into three categories: filtering-based methods, optimization-based methods, and cost aggregation-based methods. Hereafter, we call them (**F**), (**O**), and (**C**), respectively, e.g., (**F**) GF. All the results have been obtained with the source codes provided by

[4] The filtering-based methods including implicit approaches, such as an optimization-based filtering, can be formalized with a joint histogram [14]. Each bin in the joint histogram is filled with the value from a given kernel function, which corresponds to the classifying function for transduction. Thus, computing the joint histogram can also be modeled as a product of kernel and confidence matrices. The RWR, JBU, and GF construct each bin with a delta function $\delta(\cdot)$, which leads to the confidence matrix as in (18). Please refer to [14] for more details.

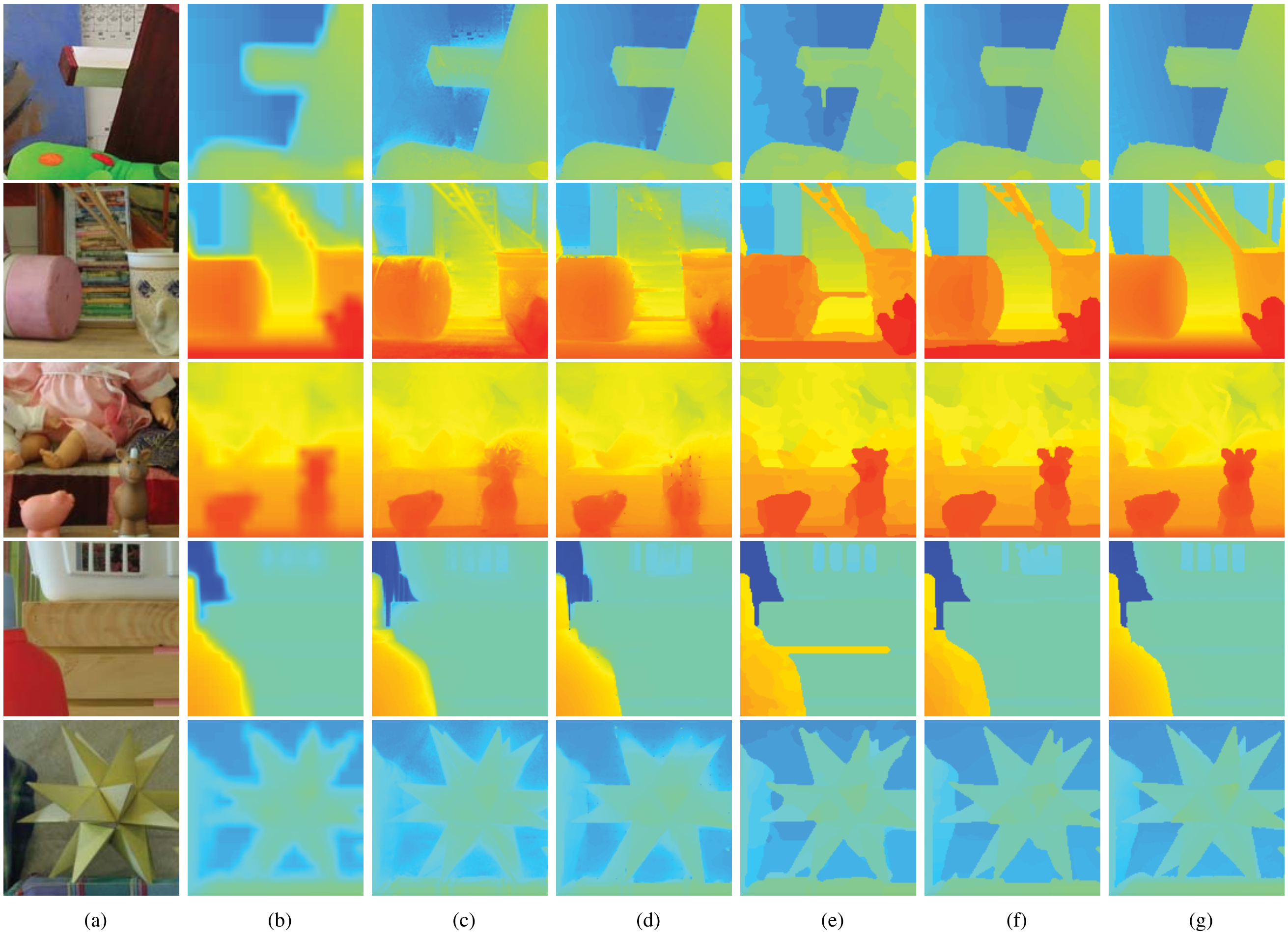


Fig. 4. Visual comparison of upsampled depth images with the Middlebury benchmark data set [39]. Snippets of (from top to bottom) *Teddy*, *Art*, *Dolls*, *Laundry*, and *Moebius*: (a) the HR intensity image, (b) bilinear interpolation, (c) GF [9], (d) ATGV [18], (e) NMedF [23], (f) the proposed method, and (g) the ground truth. Depth bleeding artifacts appear in all types of methods, expect for the proposed method.

the authors, and parameters for these methods have been carefully set through intensive experiments for achieving the best performance. In WModF [14], the multiscale color measure (MCM) has been employed to reduce aliasing artifacts. It is shown in [14], [18], [22], and [23] that the upsampling methods that use the bilateral filter [24] have exhibited worse performance than the above-mentioned methods, and thus the results obtained with such methods as JBU [8] and 3D JBU [11] are not shown here. All the results can be found at the project webpage.[5]

B. Performance Evaluation With Synthetic Examples

We demonstrate the performance of the proposed method with synthetic examples from the Middlebury benchmark data set [39]: *Tsukuba*, *Venus*, *Teddy*, *Cones*, *Art*, *Books*, *Dolls*, *Laundry*, *Moebius*, and *Reindeer*. The LR depth image is synthesized by downsampling a ground truth image with a factor of 8 in each dimension, and the corresponding color image is used as the HR intensity image.

Table II summarizes the bad matching errors with a tolerance $\varepsilon = 1$. They have been measured at all regions (*all*) only, except when the ground truth index map for discontinuous regions (*disc*) is available. The top three methods yielding low depth errors among all methods are marked with a shadow; the less the method gives error, the darker the cell gets. The proposed method outperforms other state-of-the-art methods, especially around depth discontinuities. This is consistent with our analysis that the proposed method successfully preserves depth boundaries, and it is free from depth bleeding artifacts.

Fig. 3 shows an average rank of each method with respect to the error rate in Table II. As expected, the proposed method yields the best performance. Cost aggregation-based methods tend to have a better performance than other types of methods, and optimization-based methods outperform filtering-based methods. Interestingly, a simple bilinear interpolation shows a better performance than some filtering-based and optimization-based methods, although its subjective quality looks worse than these methods (see depth discontinuities in Fig. 4(b)). Filtering-based and optimization-based methods use regularization operators explicitly or implicitly under guidance of the intensity image, and this often causes texture-copying artifacts.[6] Bilinear interpolation, however, does not suffer from

[5] http://www.di.ens.fr/%7Ebham/depthsr/

[6] In intensity guided depth SR methods, the texture in the intensity image may be transferred to the depth image. It is because discontinuities in the intensity image include both of depth and texture edges, but most of the existing methods cannot differentiate depth discontinuities from texture edges.

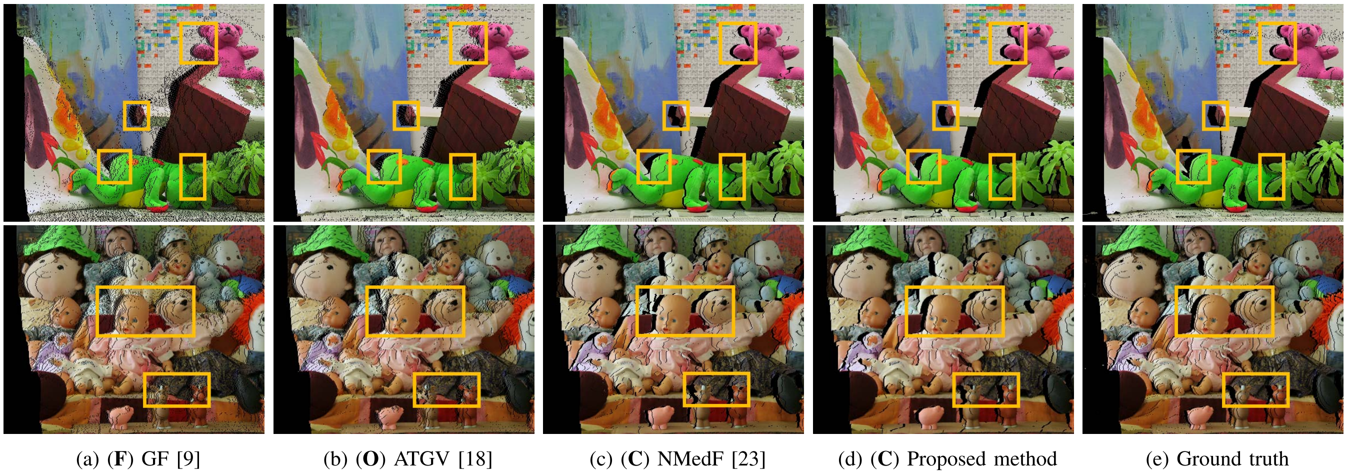
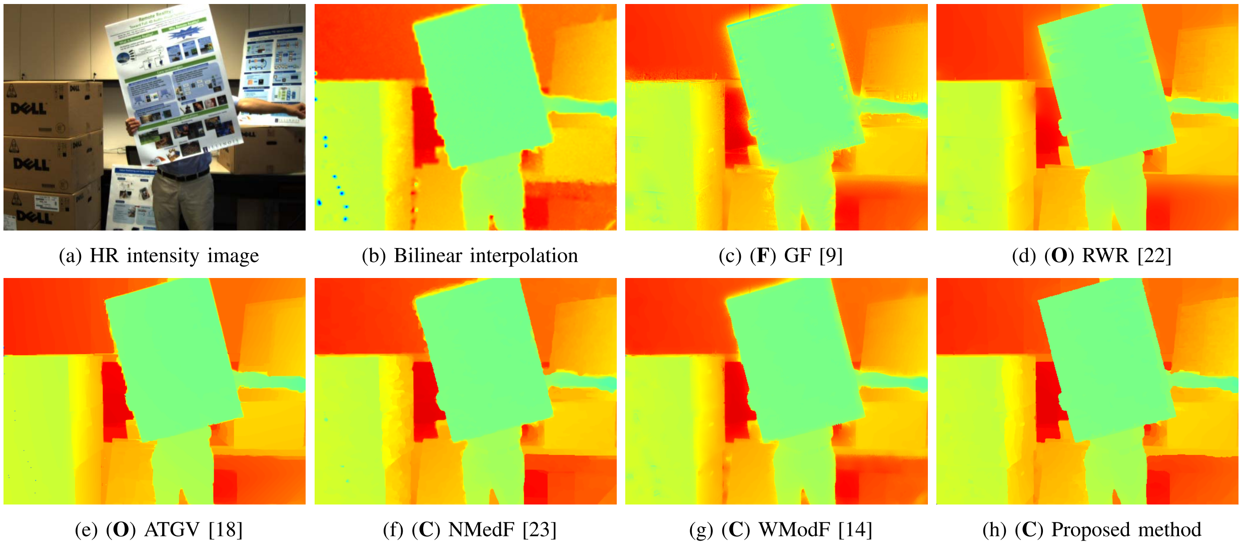

(a) (**F**) GF [9] (b) (**O**) ATGV [18] (c) (**C**) NMedF [23] (d) (**C**) Proposed method (e) Ground truth

Fig. 5. Synthesized views of (from top to bottom) *Teddy* and *Dolls*. They are rendered with the HR intensity image and corresponding upsampled depth images from (a) GF [9], (b) ATGV [18], (c) NMedF [23], (d) the proposed method, and (e) the ground truth. The rendering results of the proposed method are superior to those of other methods. Note that depth bleeding artifacts introduced by (a) GF [9], (b) ATGV [18], and (c) NMedF [23] cause warping several pixels in the synthesized view to the wrong location, e.g., the left side of *teddy bear*.


(a) HR intensity image (b) Bilinear interpolation (c) (**F**) GF [9] (d) (**O**) RWR [22]

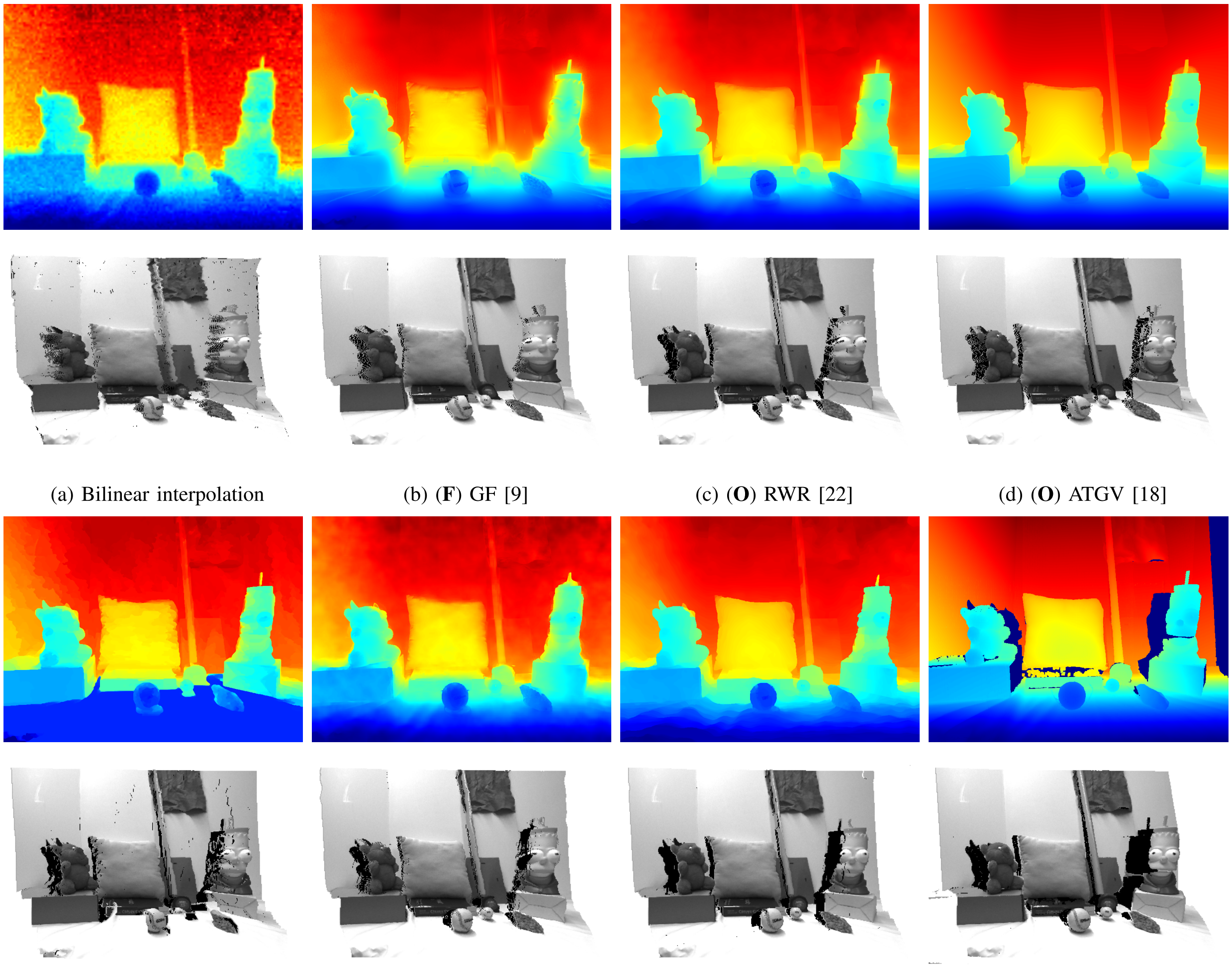
(e) (**O**) ATGV [18] (f) (**C**) NMedF [23] (g) (**C**) WModF [14] (h) (**C**) Proposed method

Fig. 6. Examples of upsampled depth images with the ToF camera: (a) the HR intensity image, (b) bilinear interpolation, (c) GF [9], (d) RWR [22], (e) ATGV [18], (f) NMedF [23], (g) WModF [14], and (h) the proposed method. Most methods alleviate sensor noise by propagating it to other regions. Note that conventional cost aggregation-based methods, including (f) NMedF [23], show similar performances to other types of methods in noisy environments. In contrast, the proposed method shows the superior performance, solving the most critical problems in conventional methods.

these artifacts, since the HR depth image is estimated by using the LR depth image only, not utilizing the intensity image.

Fig. 4 shows the results of three types of methods, (**F**) GF [9], (**O**) ATGV [18], (**C**) NMedF [23], and (**C**) the proposed method. This figure clearly shows the behavior of these methods: filtering-based methods undergo depth bleeding artifacts at depth discontinuities. Optimization-based methods consider the global structure of the intensity image, leading to better results, but depth bleeding artifacts are still witnessed. Cost aggregation-based methods yield sharper depth boundaries than other types of methods, but still suffer from depth bleeding artifacts. NMedF [23] aggregates nonlocal information very efficiently by performing edge-aware smoothing in a tree structure approximated from 2D image grid. Such an approximation of an image grid using a tree structure, however, may fail to fully capture local image attributes, thus leading to depth leakage artifacts. For example, the tree may connect distant pixels along a line when they have similar color values (e.g., *Teddy*, *Art*, and *Laundry*). The proposed method can be regarded as one of cost aggregation-based methods, and both local and global structures of the intensity image are also taken into account, providing very convincing results. It is worth noting that the cost aggregation-based methods, including our approach, can be seen as a discrete labeling approach, so the resulting depth image is inherently quantized. This artifacts can be alleviated by the quadratic fitting [11].

To visualize the influence of depth errors, virtual views are synthesized with the HR intensity image and corresponding upsampled depth images as shown in Fig. 5. As stated earlier, all methods except the proposed method suffer from depth bleeding artifacts, warping several pixels in the virtual view to the wrong location, e.g., the left side of *teddy bear*. In contrast, the synthesized view of the proposed method



(a) Bilinear interpolation (b) (**F**) GF [9] (c) (**O**) RWR [22] (d) (**O**) ATGV [18]

(e) (**C**) NMedF [23] (f) (**C**) WModF [14] (g) (**C**) Proposed method (h) Ground truth

Fig. 7. Examples of (top) upsampled depth images and (bottom) point cloud scene reconstructions of *Devil* in the Graz data set [18]. They are rendered with the HR intensity image and corresponding depth images from (a) bilinear interpolation, (b) GF [9], (c) RWR [22], (d) ATGV [18], (e) NMedF [23], (f) WModF [14], (g) the proposed method, and (h) the ground truth.

TABLE III

AVERAGE PSNR AND SSIM VALUES OF THE VIRTUAL VIEWS SYNTHESIZED WITH UPSAMPLED DEPTH IMAGES IN TABLE II

Method	*PSNR*	*SSIM*
Bilinear interpolation	17.82	0.68
(**F**) GF [9]	17.10	0.59
(**O**) RWR [22]	18.67	0.68
(**O**) ATGV [18]	18.49	0.69
(**C**) WMedF [19]	18.85	0.71
(**C**) WModF [14]	19.67	0.74
(**C**) NMedF [23]	19.77	0.72
(**C**) Proposed method	20.03	0.76

shows the sharp depth transition. Table III shows average PSNR and SSIM values [41] of the virtual views synthesized with the upsampled depth images in Table II.

C. Performance Evaluation With Real-World Examples

We have simulated upsampled depth images with the ToF camera. For this configuration, we have used *Mesa Imaging SR4000* [42] and *Point Grey Flea camera* [43] to capture LR depth images (176 × 144) and corresponding HR intensity images (512 × 384). They have been registered by warping depth values to the coordinate of the intensity image with calibration parameters. As shown in Fig. 6 (b), the depth image is degraded by various types of sensor noise. Fig. 6 shows the results of filtering-based methods, (c) GF [9], optimization-based methods, (d) RWR [22] and (e) ATGV [18], and cost aggregation-based methods, (f) NMedF [23], (g) WModF [14], and (h) the proposed method. Although most methods reduce sensor noise relatively well, some severe outliers in the LR depth image are propagated, e.g., blue dots in Fig. 6 (b), degrading the quality of the upsampled depth image. On the contrary, the proposed method shows the superior performance, addressing the most critical problems in conventional methods – it shows sharp depth discontinuities without depth bleeding artifacts. Note that our method and RWR [22] use a similar optimization formulation. The difference is that the proposed method minimizes the energy function with respect to relevance scores (in a label domain), while RWR [22] does with respect to depth values (in a depth domain).

Recently, Ferstl *et al.* have introduced a benchmark data set consisting of LR depth images captured by the ToF camera and ground truth depth images acquired from

TABLE IV

OBJECTIVE COMPARISON OF DEPTH SR METHODS ON THE GRAZ DATA SET [18]

$\varepsilon = 256 \times 0.01$	*Books*	*Devil*	*Shark*
Bilinear interpolation	62.81	64.28	61.60
(F) GF [9]	55.62	66.58	60.08
(O) RWR [22]	51.21	66.37	55.11
(O) ATGV [18]	46.70	64.26	52.73
(C) WMedF [19]	47.56	61.85	50.20
(C) WModF [14]	49.95	60.38	51.23
(C) NMedF [23]	49.39	63.25	58.02
(C) Proposed method	44.47	57.87	49.22
$\varepsilon = 256 \times 0.05$	*Books*	*Devil*	*Shark*
Bilinear interpolation	16.21	13.68	17.60
(F) GF [9]	19.65	13.12	20.68
(O) RWR [22]	14.23	9.44	16.63
(O) ATGV [18]	13.03	9.01	14.55
(C) WMedF [19]	13.33	9.81	15.77
(C) WModF [14]	11.84	9.91	13.29
(C) NMedF [23]	13.69	21.71	22.86
(C) Proposed method	10.46	8.31	12.92

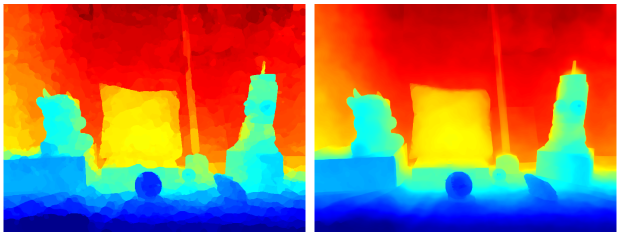


Fig. 8. Influence of the confidence matrix on smoothness of the resulting depth image: (a) $\eta = 0$, (b) $\eta = 100/c$. The parameter η determines the degree of smoothness in the upsampled depth image. Specifically, the resulting depth image is more regularized as η increases, indicating that there exists a trade off between the degrees of smoothness and sharpness.

structured light [18]. Table IV shows the quantitative evaluation of depth SR methods on the Graz data set [18]. This table shows that the proposed method always outperforms other state-of-the-art methods. Existing cost aggregation-based methods show similar performances to other types of methods in noisy environments. In NMedF [23], an aggregation is performed on the whole surface with a fronto-parallel assumption. This is problematic on textureless slanted surfaces (e.g., *Devil* and *Shark*), and every pixel in these surfaces has the same disparity value (see Fig. 7. (e)). The subjective comparison can be found in Fig. 7. As expected, the proposed method gives convincing results without depth bleeding artifacts. As mentioned earlier, cost aggregation-based methods including our method consider depth values as discrete labels, and thus they are inherently quantized in contrast to filtering- or optimization-based methods. The smoothed depth images of Fig. 7 (b), (c) and (d) might look better especially at slanted surfaces, but such a smoothing also causes depth bleeding artifacts and even depth blurring artifacts at object boundaries. These artifacts produce undesirable results with highly noticeable jagged artifacts in scene reconstruction.

D. Influence of Confidence Matrix on Depth Smoothness

As previously stated, the proposed method defines the confidence matrix Y_{ij} with the probability that the vertex x_i is labeled as $y_i = j$. The parameter η of $\mathbf{Y}$ in (3) determines a degree of smoothness in the upsampled depth image, as shown in Fig. 8. As the parameter η increases, the relevance score between vertices is more regularized, making neighboring vertices have similar scores. This indicates the parameter η can be adjusted to deal with noise that may exist in the initial depth image, and should be set properly to balance the degrees of smoothness and sharpness in the upsampled depth image.

TABLE V

CLASSIFICATION OF STATE-OF-THE-ART DEPTH SR METHODS

Type	Local structure	Global structure
F	JBU [8], GF [9]	✗
O	✗	RWR [22], ATGV [18]
C	WMedF [19], WModF [14] 3D JBU [11]	WMedF [23], ours

V. DISCUSSION AND CONCLUSION

We have described the intensity guided depth SR method. We have principally resolved the depth bleeding problem in current depth SR methods. Depth SR was regarded as a discrete labeling task, and was then formulated as a graph-based transduction problem. We have compared our approach with a large number of state-of-the-art methods through synthetic and real-world examples. Experimental results have shown that cost aggregation-based methods tend to achieve a better performance than filtering-based and optimization-based methods. The proposed method upsamples the LR depth image in a way similar to the cost aggregation-based method, yet the global structure of the intensity image is considered. Our method outperforms existing approaches in terms of both qualitative and quantitative evaluations. The comparison of several depth SR methods is summarized in Table V.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

[1] X. Ren, L. Bo, and D. Fox, "RGB-(D) scene labeling: Features and algorithms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2759–2766.

[2] D. Lin, S. Fidler, and R. Urtasun, "Holistic scene understanding for 3D object detection with RGBD cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1417–1424.

[3] S. Song and J. Xiao, "Tracking revisited using RGBD camera: Unified benchmark and baselines," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 233–240.

[4] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 101–115.

[5] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2003, pp. I-195–I-202.

[6] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "LidarBoost: Depth superresolution for ToF 3D shape scanning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 343–350.

[7] O. M. Aodha, N. D. F. Campbell, A. Nair, and G. J. Brostow, "Patch based synthesis for single depth image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 71–84.

[8] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, 2007, Art. ID 96.

[9] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 1–14.

[10] D. Chan, H. Buisman, C. Theobalt, and S. Thrun, "A noise-aware filter for real-time depth upsampling," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2008, pp. 1–12.

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

[12] Q. Yang *et al.*, "Fusion of median and bilateral filtering for range image upsampling," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4841–4852, Dec. 2013.

[13] J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," in *Proc. NIPS*, vol. 18. 2005, pp. 291–298.

[14] D. Min, J. Lu, and M. N. Do, "Depth video enhancement based on weighted mode filtering," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1176–1190, Mar. 2012.

[15] B. Huhle, T. Schairer, P. Jenke, and W. Straßer, "Fusion of range and color images for denoising and resolution enhancement with a non-local filter," *Comput. Vis. Image Understand.*, vol. 114, no. 12, pp. 1336–1345, 2010.

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

[17] S. Schwarz, M. Sjostrom, and R. Olsson, "A weighted optimization approach to time-of-flight sensor fusion," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 214–225, Jan. 2014.

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

[19] Z. Ma, K. He, Y. Wei, J. Sun, and E. Wu, "Constant time weighted median filtering for stereo matching and beyond," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2013, pp. 49–56.

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

[21] J. Choi, D. Min, and K. Sohn, "Reliability-based multiview depth enhancement considering interview coherence," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 4, pp. 603–616, Apr. 2014.

[22] B. Ham, D. Min, and K. Sohn, "A generalized random walk with restart and its application in depth up-sampling and interactive segmentation," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2574–2588, Jul. 2013.

[23] Q. Yang, "Stereo matching using tree filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[24] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. 6th Int. Conf. Comput. Vis.*, Jan. 1998, pp. 839–846.

[25] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Scholkopf, "Ranking on data manifolds," in *Proc. NIPS*, vol. 17. 2004, pp. 169–176.

[26] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[27] X.-M. Wu, Z. Li, A. M. So, J. Wright, and S.-F. Chang, "Learning with partially absorbing random walks," in *Proc. NIPS*, 2012, pp. 3086–3094.

[28] T. H. Kim, K. M. Lee, and S. U. Lee, "Learning full pairwise affinities for spectral segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1690–1703, Jul. 2013.

[29] O. Duchenne, J.-Y. Audibert, R. Keriven, J. Ponce, and F. Ségonne, "Segmentation by transduction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[30] J. He, M. Li, H.-J. Zhang, H. Tong, and C. Zhang, "Generalized manifold-ranking-based image retrieval," *IEEE Trans. Image Process.*, vol. 15, no. 10, pp. 3170–3177, Oct. 2006.

[31] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3166–3173.

[32] X. Liu, D. Zhao, J. Zhou, W. Gao, and H. Sun, "Image interpolation via graph-based Bayesian label propagation," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1084–1096, Mar. 2014.

[33] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. NIPS*, vol. 16. 2003, pp. 321–328.

[34] M. Donoser and H. Bischof, "Diffusion processes for retrieval revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1320–1327.

[35] D. Krishnan, R. Fattal, and R. Szeliski, "Efficient preconditioning of Laplacian matrices for computer graphics," *ACM Trans. Graph.*, vol. 32, no. 4, 2013, Art. ID 142.

[36] D. Min, S. Choi, J. Lu, B. Ham, K. Sohn, and M. N. Do, "Fast global image smoothing based on weighted least squares," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5638–5653, Dec. 2014.

[37] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[38] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. ICML*, vol. 3. 2003, pp. 912–919.

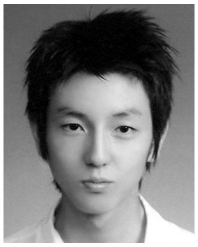
[39] *Middlebury Benchmark*. [Online]. Available: http://vision.middlebury.edu/stereo/

[40] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, nos. 1–3, pp. 7–42, 2002.

[41] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[42] *Mesa Imaging SR 4000*. [Online]. Available: http://www.mesa-imaging.ch/

[43] *Point Grey Flea Camera*. [Online]. Available: http://www.ptgrey.com/



Bumsub Ham (M'14) received the B.S. and Ph.D. degrees from the School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea, in 2008 and 2013, respectively. He is currently a Post-Doctoral Research Fellow with Willow Team, INRIA Rocquencourt, École Normale Supérieure, and the Centre National de la Recherche Scientifique. His current research interests include computer vision, computational photography, and machine learning, in particular, regularization, graph matching, and superresolution, both in theory and applications. He was a recipient of the Honor Prize in the 17th Samsung Human-Tech Prize in 2011 and the Grand Prize in Qualcomm Innovation Fellowship in 2012.

Dongbo Min (M'09) received the B.S., M.S., and Ph.D. degrees from the School of Electrical and Electronic Engineering, Yonsei University, in 2003, 2005, and 2009, respectively. From 2009 to 2010, he was with Mitsubishi Electric Research Laboratories as a Post-Doctoral Researcher, where he developed a prototype of 3D video system. From 2010 to 2015, he was with the Advanced Digital Sciences Center, Singapore, which was jointly founded by the University of Illinois at Urbana-Champaign and the Agency for Science, Technology and Research, a Singapore Government Agency. Since 2015, he has been an Assistant Professor with the Department of Computer Science and Engineering, Chungnam National University, Daejeon, Korea. His research interests include computer vision, 2D/3D video processing, computational photography, augmented reality, and continuous/discrete optimization.

Kwanghoon Sohn (M'92–SM'12) received the B.E. degree in electronic engineering from Yonsei University, Seoul, Korea, in 1983, the M.S.E.E. degree in electrical engineering from the University of Minnesota, Minneapolis, MN, USA, in 1985, and the Ph.D. degree in electrical and computer engineering from North Carolina State University, Raleigh, NC, USA, in 1992. He was a Senior Member of the Research Staff with the Satellite Communication Division, Electronics and Telecommunications Research Institute, Daejeon, Korea, from 1992 to 1993, and a Post-Doctoral Fellow with the MRI Center, Medical School of Georgetown University, Washington, DC, USA, in 1994. He was a Visiting Professor with Nanyang Technological University, Singapore, from 2002 to 2003. He is currently a Professor with the School of Electrical and Electronic Engineering, Yonsei University. His research interests include 3D image processing, computer vision, and image communication. He is a member of the International Society for Optics and Photonics.