Depth Analogy: Data-driven Approach for Single Image Depth Estimation using Gradient Samples

Sunghwan Choi, *Student Member, IEEE,* Dongbo Min, *Member, IEEE,* Bumsub Ham, *Member, IEEE,* Youngjung Kim, *Student Member, IEEE,* Changjae Oh, *Student Member, IEEE,* and Kwanghoon Sohn, *Senior, IEEE*

***Abstract*—Inferring scene depth from a single monocular image is a highly ill-posed problem in computer vision. This paper presents a new gradient-domain approach, called depth analogy, that makes use of analogy as a means for synthesizing a target depth field, when a collection of RGB-D image pairs are given as training data. Specifically, the proposed method employs a non-parametric learning process that creates an analogous depth field by sampling reliable depth gradients using visual correspondence established on training image pairs. Unlike existing data-driven approaches that directly select depth values from training data, our framework transfers depth gradients as reconstruction cues, which are then integrated by Poisson reconstruction. The performance of most conventional approaches relies heavily on the training RGB-D data used in the process, and such a dependency severely degenerates the quality of reconstructed depth maps when the desired depth distribution of an input image is quite different from that of the training data, *e.g.*, outdoor vs. indoor scenes. Our key observation is that using depth gradients in the reconstruction is less sensitive to scene characteristics, providing better cues for depth recovery. Thus, our gradient-domain approach can support a great variety of training range datasets that involve substantial appearance and geometric variations. Experimental results demonstrate that our (depth) gradient-domain approach outperforms existing data-driven approaches directly working on depth domain, even when only uncorrelated training datasets are available.**

***Index Terms*—Depth estimation, 2D-to-3D conversion, gradient transfer, non-parametric sampling, image analogy.**

I. Introduction

UNDERSTANDING 3D structure of a scene undoubtedly plays a fundamental role in perceiving a real world scenery. Indeed, the human visual system (HVS) has no difficulty in understanding its underlying 3D structure by virtue of the ability to perceive a relative depth ordering from pre-learned perceptual experiences as well as to measure an absolute depth value of scenes from the binocular vision system (BVS). In the BVS, two eyes receive slightly different images of the scene, and an associated disparity is subsequently inferred through binocular fusion. This mechanism has widely been adopted in computational stereo approaches that produce a disparity map by seeking two-view correspondence [1]. Interestingly, even in monocular situations, the depth perception still works in the HVS, with the exception of some optical illusions. This is because a prior knowledge needed for understanding a scene depth can be learned from various monocular depth cues such as shading, motion, defocus, or occlusion. In contrast, inferring a 3D structure from a single 2D image using computational approaches remains extremely challenging due to its ill-posed characteristics.

While many depth estimation methods have been developed for extracting plausible depth from a single image based on parallax, motion, or shading cues [2], [3], strict assumptions imposed on their prediction model limit their application up to some restricted environments. To address this limitation, several data-driven approaches have been developed by leveraging the discriminative power of a large scale RGB-D database [4], [5]. They typically attempt to solve a highly ill-posed depth prediction problem by transferring plausible depth labels to an input image from visually similar images retrieved from RGB-D training database. These methods, however, run under a strict assumption that the training RGB-D database contains depth images with geometric characteristics similar to that of an input color image, and thus they work well only when the training data is highly correlated with the input image.

Fig. 1(a) shows the depth distribution computed from two well-known RGB-D datasets (Make3D [6] and NYU Kinect V2 [7]). Interestingly, the Make3D data (for outdoor scenes) shows an exponential-like distribution with a peak near zero, while the NYU Kinect data (for indoor scenes) has a nearly uniform distribution across all depth ranges. It is because the Make3D dataset was generated by a Laser scanner for outdoor scenes, and thus foreground objects (with small depth values) are biased in acquiring depth fields due to the wide coverage of the Laser scanner. Please refer to [6], [8] for more details. This indicates that the depth maps of the retrieved color images do not always provide useful depth cues unless the database is carefully established. Such a dependency severely degenerates the quality of reconstructed depth maps when the desired depth distribution of the input image is quite different from that of the training RGB-D data, *e.g.*, inferring a depth map of an input image taken at outdoor with the NYU Kinect V2 data (mostly capturing indoor scenes).

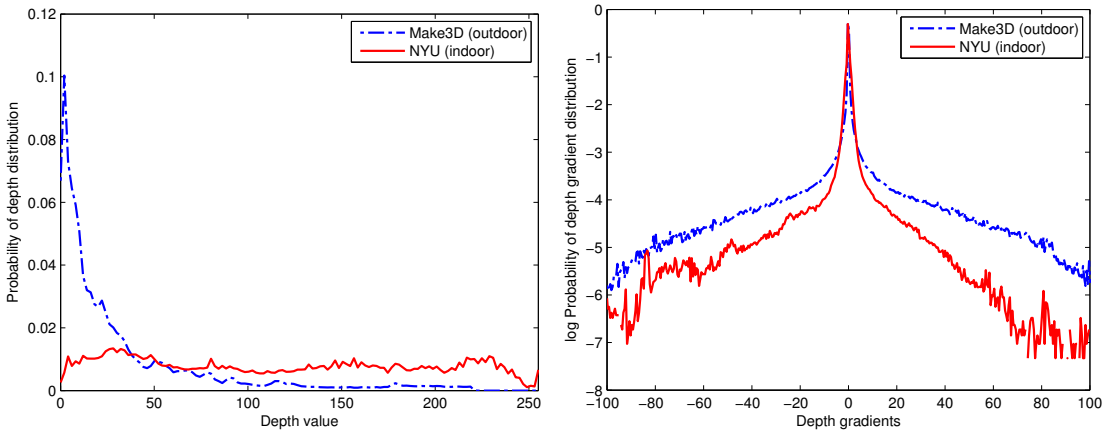
Even when the database with similar scene semantics is used, transferring original depth values from the database may cause a depth ambiguity. Fig. 2 shows two images from

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2013R1A2A2A01068338).

S. Choi, Y. Kim, C. Oh, and K. Sohn are with the School of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, Korea (e-mail: shch@yonsei.ac.kr; read12300@yonsei.ac.kr; ocj1211@yonsei.ac.kr; khsohn@yonsei.ac.kr).

D. Min is with the Advanced Digital Sciences Center, Singapore 100190 (e-mail: dongbo@adsc.com.sg).

B. Ham is with Willow Team, INRIA Grenoble-Rhône-Alpes, Grenoble 250101, France (e-mail: bumsub.ham@inria.fr).



(a) (b)

Fig. 1. Statistics of natural depth images. We measure (a) the probability of depth distribution and (b) the log_{10} probability of depth gradient distribution using publicly available range datasets (Make3D [6] and NYU Kinect V2 [7]). For computing depth distribution, depth images are normalized in the range of [0, 255]. For depth gradient distribution, first order depth gradients along x- and y-axis are measured from the normalized depth images. In (a), the Make3D dataset shows a peak near zero, but the NYU dataset has nearly uniform distribution across all depth ranges. It is because the Make3D dataset is generated by a Laser scanner for outdoor scenes, and that closest objects are more biased in acquiring depth fields due to the wide coverage of the Laser scanner. However, both scenes show similar trends in the depth gradient distribution of (b) (*i.e.* heavy-tailed distribution), indicating that the depth gradient (contrast) is more informative in describing natural depth structures.

the training RGB-D data. These images taken at the same indoor scene exhibit abrupt depth variations for corresponding objects, leading to depth ambiguities when superimposing depth values from two images simultaneously. Although using more sensitive features may help increase the possibility of finding more proper depth candidates thanks to the scale invariant property of the feature descriptors used, two slightly different depth values from Fig. 2(b) and (d) are directly superimposed and averaged to compute a final depth value. This makes the depth reconstruction more challenging, and thus additional constraints are usually introduced such as scene-warping enforcement [4], [9] and/or sophisticated depth interpolation techniques [4], [8]. Additionally, the depth ambiguity becomes more serious when the training data is not sufficiently correlated to an input image. In this case, depth values collected from the training data through a dense scene alignment [4], [9], *e.g.* using SIFT flow [10], may be not so useful for depth recovery of an input image.

To tackle this problem, we propose a new gradient-domain framework for single image depth estimation, called *depth analogy*. Our approach was motivated by the image analogy [11] that explores the coincidence of statistical relations behind training image pairs. By utilizing a statistical similarity with a user-provided training data (*e.g.* two pairs of images or patches), the image analogy can naturally produce various image editing effects such as artistic filtering, texture transfer, and super-resolution. Interestingly, many data-driven approaches [4], [5] for single image depth estimation share similar principles with the image analogy, but they transfer depth values from a large scale RGB-D training data with no user intervention. Similar to existing data-driven approaches, our method also utilizes statistical similarities across a collection of RGB-D image pairs, but we formulate the depth transfer task on the gradient domain. Namely, instead of directly selecting *depth values* from training data, our approach transfers *depth gradients* as reconstruction cues, which are then integrated by Poisson reconstruction. This new formulation enables overcoming several limitations incurred by the strict dependency assumption (between training data and input image) imposed on existing data-driven approaches.

Our key insight is that utilizing depth gradients obtained from nearest neighbor images is less sensitive to scene characteristics than directly transferring depth values. We demonstrate in Fig. 1(b) that the gradient distribution of depth images is independent of scene characteristics. Specifically, it follows a hyper Laplacian distribution, regardless of indoor or outdoor scenes. This is also consistent with the fact that a relative depth contrast (ordering) is one of the most important factors for 3D depth perception of the HVS [12], [13]. With this powerful reconstruction cue (depth gradient), our depth analogy algorithm is thus able to recover a plausible depth field, even when accompanying with challenging training databases exhibiting substantial photometric and geometric variations, which cannot be addressed by existing methods [4], [5].

Our algorithm first synthesizes depth gradient fields by transferring depth gradient values of nearest neighbor training samples extracted from RGB-D database, and then adaptively fusing them based on confidence measures. This gradient field is then integrated with a Poisson surface reconstruction [14], producing an initial depth estimate for an input image. The initial depth map is further refined by smoothing it out with a weighted median filter [15], since it might be often noisy and sparse due to the incomplete reconstruction of the gradient field.

This paper is organized as follows. Section II presents the related work, and Section III explains the statistics of natural depth images. The proposed depth analogy algorithm is presented in Section IV. Then, the performance of the proposed method is demonstrated in Section V. Finally, Section VI discusses limitations and concludes this paper.

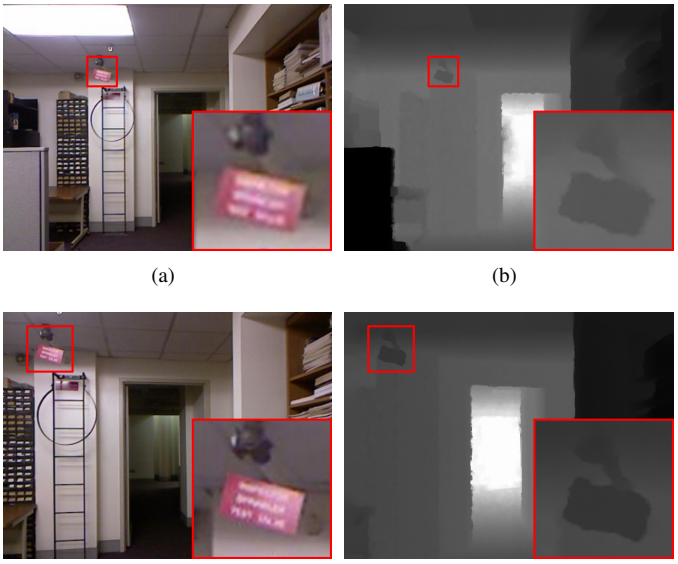


Fig. 2. Examples of depth ambiguity caused by range variations. Both images of (a) and (c) capture the same scene, but capturing positions are slightly different. This incurs a perspective distortion in depth maps as in (b) and (d). The signpost in red box has similar appearance in both images, but the corresponding depths vary according to the perspective distortion. The lower the intensity, the closer the object.

II. Related Work

Early methods for single image depth estimation focus on using user-annotations such as sparse depth scribbles [16]–[18] and/or a priori geometric model [19], [20]. They require the user to manually assign precise annotations to the input image, thus making it difficult to apply the method to automatic vision-related tasks.

In parallel, several automatic methods have been proposed to directly estimate the scene depth from a single image with no user intervention such as parallax-based methods [2] and shape-from-shadings [3]. For instance, parallax-based methods [21] typically require a translational camera motion for a static scene. Some approaches [2] propose to simply convert an object motion to a scene depth, but they do not capture a realistic scene depth when no object motion exists in the scene. Shape-from-shading methods [3] usually require surfaces of an image to have fairly uniform color and texture. In practice, most real images/videos do not always meet such requirements. The depth data we wish to discover is complicatedly coupled with various monocular depth cues other than just a single one. Therefore, these approaches [2], [3] relying on a single cue only do not scale well for general scenes.

Alternatively, data-driven approaches [4], [5] significantly advanced the performance of the single image depth estimation by effectively leveraging the discriminative power of large scale RGB-D databases (consisting of color images and associated ground truth depth maps). They assume that scenes with similar semantics should have roughly similar depth distributions. This assumption on appearance-depth correlation enables using candidate depth samples from RGB-D training data by linking them with the input image through appearance-based correspondence.

As a pioneering work, Konrad *et al.* [5] proposed to use depth maps of K color images that are retrieved by the histograms of oriented gradients (HOG) descriptor [22]. They fuse K depth maps by computing a median depth value for each pixel. This initial depth estimate is then refined by performing a joint bilateral filtering [23]. This method is fast and easy to implement. However, the K depth maps are directly fused with no pixel-level dense alignment, and thus local properties of retrieved K depth maps are not considered.

More recently, Karsch *et al.* devised the depth transfer algorithm that uses a global form of depth fusion to automatically recover scene depth from a single image. In contrast to [5], the retrieved images are densely warped to the input image by making use of a generic dense scene alignment like SIFT Flow [10]. With this, the depth transfer method achieves relatively good depth results even when locally different training samples yet having sufficiently similar global characteristics are provided, although the SIFT flow-based warping of all retrieved images requires a very expensive operation.

There are also several approaches for single image depth estimation. The Make3D [6] algorithm was also proposed to provide realistic depth maps by modeling monocular cues and the relation among multiple regions inside an input image based on a Markov Random Field (MRF). Depth (*e.g.* plane) parameters are trained with a ground truth RGB-D dataset. In [24] and [25], semantic object labels are integrated with monocular depth features to improve depth estimation quality. Eigen *et al.* [26] uses multi scale deep neural networks. The methods [27], [28] proposed to incorporate additional geometry information such as surface normal vectors in order to improve an overall performance of depth map prediction.

Note that the methods in [4] and [6] also used depth gradients as well as depth values in their inference process. Contrarily, we use depth gradients only in order to address an even more challenging scenario: estimating depth using a training dataset with quite different depth distribution from that of an input image. Our approach does not employ any candidate's depth values in the inference procedure, focusing only on the candidate's depth gradients obtained by localized matching. We demonstrate in experiments that this approach can cope with the problems regarding the depth ambiguity and strict dependency on training data.

As already explained in Section I, existing data-driven approaches have two problems; 1) a strict dependency for training data due to the statistical variations of depth distribution as in Fig. 1(a) and 2) a depth ambiguity due to depth range variations in training RGB-D data as in Fig. 2. In practice, finding appropriate training datasets with scene depth semantics similar to that of an input image is non-trivial in most cases. Thus, it is straightforward to expect that existing data-driven approaches [4], [5] often fail when the assumption on appearance-depth correlation is violated. Although Karsch *et al.* [4] introduced a mean depth prior as a constraint in their optimization formulation to relax this limitation, the prior knowledge from the computed mean depths also varies according to the scene categories of training

datasets (*e.g.* indoor or outdoor). Moreover, if the training images having different depth values in regions with similar appearance are retrieved, then existing methods fail to produce correct depth maps due to the depth ambiguity. Therefore, a training database of RGB-D images should be carefully defined due to their reliance on appearance-depth assumption.

Note that one could use information from the image itself or associated metadata to infer the type (*e.g.* indoor or outdoor) or mean depth of the scene [29]. This may help finding appropriate training datasets.

III. Statistics of Natural Depth Images

In this section, we present a detailed description about statistics of natural depth images. By inspecting the statistical characteristics of natural depth images, it is not difficult to see that the depth gradient distribution usually look *statistically invariant* against scenes with different semantics, while the depth distribution varies depending on capturing environments and/or sensing devices. This offers a key insight into our approach, which a depth contrast (gradient) cue is very crucial to recovering a scene geometry.

A. Experimental setup

To verify the statistical characteristics of depth distributions, we conducted experiments using several real-world depth images from two different datasets with distinct semantic characteristics: the Make3D dataset [6] and the NYU Kinect V2 dataset [7] captured by a Laser scanner and a Kinect sensor, respectively. Each depth image was normalized in the range of [0, 255] so as to alleviate a scaling effect. We computed the histogram of depth values and the $\log_{10}$ histogram of depth gradients for each image, respectively. The histograms were averaged and normalized for each dataset. The depth gradients were computed for both x and y coordinates. Fig. 1 plots the statistics of two semantically distinct datasets.

B. Analysis

As shown in Fig. 1(a), depth distributions computed from different semantic scenes exhibit a significant *global* variation. Namely, outdoor scenes (Make3D) show a peak near zero, while indoor scenes (NYU) exhibit a nearly uniform distribution. In contrast, as shown in Fig. 1(b), each depth gradient distribution shows a global consistency in the sense that both distributions have sharp peaks at zero and heavy-tails. Only a small *local* inconsistency are observed. This demonstrates the statistical invariance property (up to a small error bound) of natural depth gradients. Similar investigation can also be found in the seminal work of [30]. In fact, the statistical variation of spatial depth distribution stems from the spatially-varying nature of scene structure. More specifically, depth values may not be equally-comparable due to the diversity that exists in capturing environments as well as in the sensible range of depth sensing devices. While existing approaches [4], [5] implicitly assume the appearance-depth correlation, this is valid only when the depth characteristics of the training data are closely matched to that of an input image. Thus, the statistical variation of spatial depth distribution limits the application of existing approaches to some restricted range database.

C. Verification

To verify this, we conducted additional experiments by measuring the distribution correlation between the ground truth input depth image and its K nearest neighbor training depth images. Given an input image with its ground truth depth map, we retrieved top $K = 40$ nearest neighbor color images for each dataset by using a GIST descriptor [31]. We have measured the distribution correlation by using various scene descriptors such as GIST [31], HOG [22], PHOG [32], and GIST+PHOG (*i.e.* weighted combination of GIST and PHOG). However, there were no considerable changes in the resultant distribution correlation curves. We then computed depth and its gradient histograms in a way similar to Section III-A. We then measured a Bhattacharya similarity [33] between the histograms obtained with depth maps corresponding to the input and retrieved images.

For the input P and its retrieved $Q^{(k)}$ depth histograms with b bins, the correlation score is defined by accumulating similarity for $k = 1 \dots K$ as follows:

$$c = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{b} \left(P_i Q_i^{(k)} \right)^{\frac{1}{2}}, \tag{1}$$

where the space is quantized into $b = 127$ bins.

Fig. 3 plots average correlation curves for depth and its gradient distributions. We evaluated 100 test image pairs randomly selected from the Make3D dataset. A significant drop is visible in case 4) when highly uncorrelated dataset (NYU) is given for the input image (selected from Make3D). This indicates that existing approaches [4], [5] may fail to capture useful depth cues from such an uncorrelated dataset. Thus, they require establishing RGB-D training database very carefully. For instance, indoor (or outdoor) training images should be used accordingly for an indoor (or outdoor) input image, leading to extra difficulties and demands in building the database. Even when using the correlated dataset (Make3D) as in case 3), the correlation of depth distribution decreases as K increases due to depth ambiguity problem (see Fig. 2). In contrast, two correlation curves from case 1) and 2) demonstrate that statistical characteristics of depth gradient are well preserved against depth variations incurred by different training data as well as internal variations (*i.e.* depth ambiguity) within the same training data.

IV. Gradient-domain Depth Analogy

Based on the statistical invariance property inherent in natural depth images, it is intuitive to design a transfer model in the gradient domain. This model allows our algorithm to scale well for various training range data. We synthesize a plausible depth gradient field with a set of depth gradients locally-sampled from the retrieved training pairs. To this end, we leverage recent works on fast dense nearest neighbor field search and efficient edge-aware filter, and combine these

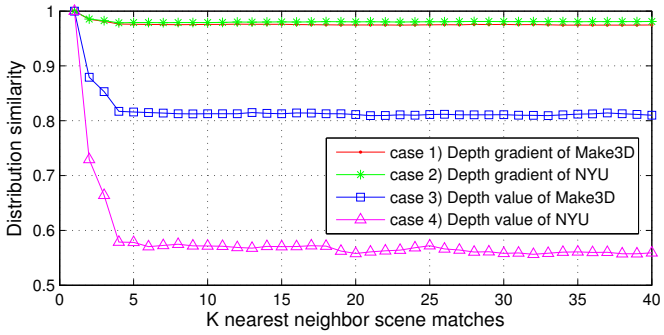


Fig. 3. Distribution correlation curves for Make3D and NYU datasets. Given a randomly selected input image from the Make3D dataset, we measure a correlation of both depth and depth gradient distribution by computing a Bhattacharya distance between histograms of the input ground truth depth map and K nearest neighbor scene matches. Depth gradients are more consistent against appearance variations in training data.

methods smartly in the non-parametric learning framework defined in the gradient domain. This synergetic design makes our approach an order of magnitude faster than existing non-parametric methods [4], while mitigating the strict dependency for RGB-D training data and resolving the depth ambiguity problem.

A. Algorithm Overview

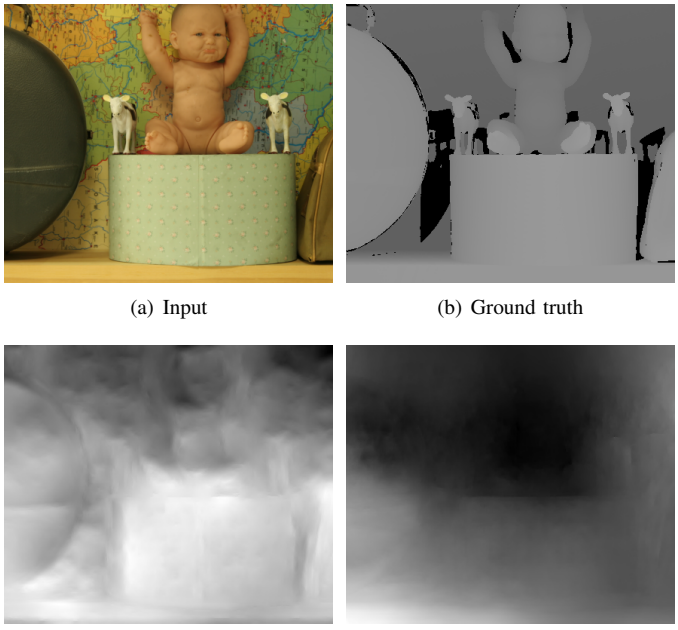
Suppose there is a database of RGB-D images $\mathcal{I} = \{(I_i, D_i)|i = 1, \ldots, N\}$, where I_i and D_i denote respectively a color image and its associated depth map. N is the size of the database. Our objective is to infer a spatially coherent, discontinuity-preserving depth field D^* of an input image I by learning depth from this database.

Our approach consists of four steps. We first retrieve $K \ll N$ training pairs $C = \{(I_k, D_k)|k = 1, \ldots, K\}$ from a database of RGB-D images $\mathcal{I}$ (Sec. IV-B). For simplicity, we re-index the retrieved color and depth maps using $k = 1, \ldots, K$. Then, we sample hypothetical depth gradients by performing a local correspondence search between an input image and each of the retrieved images, resulting in K depth gradient samples for all pixels in the input image. A depth gradient for each pixel is determined from K gradient samples, based on the matching confidence (Sec. IV-C). We then reconstruct the initial depth field D from the gradient field by solving Poisson equations (Sec. IV-D), and finally the spatial smoothness constraint is implicitly enforced on D by applying edge-aware median filter (Sec. IV-E), producing a desired depth field D^*.

B. Retrieval of Training RGB-D Images

To select training pairs $C = \{(I_k, D_k)|k = 1, \ldots, K\}$ from the large scale database $\mathcal{I}$, we retrieve similar images by means of high-level image features. In our approach, visual similarity between two images is measured using the Pyramid of Histograms of Orientation Gradients (PHOG) descriptor [32].

Let us denote $F_I \in \mathbb{R}^n$ as an n-dimensional PHOG feature vector for image I. We adopt the default setting used in [32]: $L = 3$ pyramid levels and $B = 8$ bins for gradient



(a) Input (b) Ground truth

(c) PM-WM (d) SF-WM

Fig. 4. Depth reconstruction using locally- and globally-aligned training pairs (PatchMatch vs. SIFT Flow). (a) Input image and (b) its ground truth depth map. The depth fields of (c) and (d) are reconstructed by solving the Poisson equation [14] (see Sec. IV-D) for depth gradients obtained using (c) PatchMatch-based (PM) and (d) SIFT Flow-based (SF) sampling approaches. Note that post-processing (Sec. IV-E) is not applied to these results for fair comparison. The higher the intensity, the closer the object.

histograms, resulting in a 680-dimensional feature vector (*i.e.* $n = B\sum_{l=0}^{L} 4^l$). The dissimilarity metric between the input image I and the candidate image I_i from the database $\mathcal{I}$ is defined by the sum of squared differences (SSD) between two corresponding feature vectors as follows:

$$dist(I_i, I) = \|F_{I_i} - F_I\|_2^2 . \quad (2)$$

We then extract the lowest K matching pairs with respect to the matching distance (2) and define them as the training pairs C that are relevant for learning depth. Note that the PHOG feature vectors for all images in the database are pre-computed and stored for a fast retrieval.

C. Depth Gradient Reconstruction

As pointed out in Sections I and III, our method designs the transfer model in the depth gradient domain considering the statistical invariance property of depth gradients. In short, we assume that two regions with similar appearances are likely to have similar depth gradients, not similar depth values.

An analogous depth gradient field $\mathbf{g}$ is learned using gradient samples locally aggregated from K training pairs $C = \{(I_k, D_k)|k = 1, \ldots, K\}$. In other words, we want to estimate an analogous depth gradient field $\mathbf{g}(p) = (g_x(p), g_y(p))^T$ for all pixels p that relates to the input image I in the same way that the gradient field of D_k relates to I_k. We do this by establishing a dense visual correspondence between the input image I and all training images I_k, and then aggregate hypothetical depth gradients from corresponding depth fields D_k.

1) Depth Gradient Sampling: Let us define a warping function $m : I \rightarrow \mathbb{R}^2$ over all possible pixel coordinates in image I. Then, the correspondence search from I to I_k can be expressed as finding a warping function $m_k(p)$ for $k = 1 \dots K$:

$$m_k(p) = \arg\min_m \|f_I(p) - f_k(p+m)\|_2^2, \quad (3)$$

where $f_I(p)$ is a feature vector for image I reflecting appearance characteristics around pixel p. Similarly, $f_k(p)$ is a feature vector for training image I_k around pixel p. Regarding appearance variations that exist on natural images, we employ dense SIFT features [34], [35] that properly describe image appearance properties. To efficiently compute the warping function for each training image, we use approximated nearest neighbor (ANN) search algorithm, PatchMatch (PM) [36], [37], as a means for gradient sampling process. The randomized search adopted in the PM enables a fast correspondence estimation over an entire image, thus increasing the likelihood of finding suitable depth gradients.

It should be noted that unlike previous works relying on global correspondence algorithms (*e.g.* SIFT Flow [10]), we impose no smoothness constraints on computing a correspondence field and find the best match independent of neighboring matches. We observed in Fig. 6 that depth gradient fields are sparse where meaningful gradients are mostly located at image boundaries while the rest has a (near-) zero magnitude. This implies that, when aligning candidate training pairs in terms of image appearance, the global smoothness prior commonly adopted in global dense warping algorithms like [10] is not so effective in warping hypothetical depth gradients. Moreover, due to a heavy computational load, they typically constrain the search range to a small local window, not an entire image. Figs. 4(c) and (d) show the depth fields reconstructed using locally- and globally-aligned training pairs (PatchMatch vs. SIFT Flow) in our framework. In Fig. 4(d), the method fails to capture fine details of objects due to spatial regularization of SIFT Flow algorithm. In contrast, the method using the PatchMatch recovers appropriate depth structures as shown in Fig. 4(c).

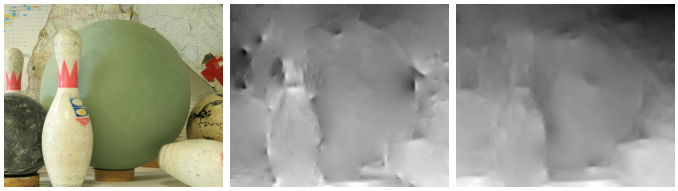
Using the warping functions computed from all training images, we sample depth gradients $\mathbf{g}^{(k)}(p)$ from k^{th} warped training pair (I_k, D_k) as follows:

$$\mathbf{g}^{(k)}(p) = (\nabla_x D_k(p + m_k(p)), \nabla_y D_k(p + m_k(p)))^T, \quad (4)$$

where ∇_q is a gradient operator along q-coordinate. Since some of these estimates may be inaccurate, we also measure the sampling confidence for all pixels based on matching distance. The confidence $w_k(p)$ is defined as the normalized matching distance, which is the form of

$$w_k(p) = 1 - \frac{\|f_I(p) - f_k(p + m_k(p))\|_2^2}{\sum_{t=1}^{K} \|f_I(p) - f_t(p + m_k(p))\|_2^2}. \quad (5)$$

The confidence term $w_k(p)$ gives a higher weight to the pixel p, when two patches centered at p and $p + m_k(p)$ match more closely in the feature space.



(a) Input (b) PM-WTA (c) PM-WM

Fig. 5. Depth reconstruction using two depth gradient voting methods: PatchMatch-based (PM) sampling with (b) winner-takes-all (WTA) and (c) our weighted median (WM) voting. Similar to Fig. 4, the depth fields of (b) and (c) are reconstructed by solving the Poisson equation (see Sec. IV-D). The higher the intensity, the closer the object.

2) Weighted Median Voting: A final gradient is then selected by computing a weighted median value among K gradient samples. For each pixel p, we first sort out confidence values in an ascending order, and then compute an index k^* in which the sum of corresponding weights is approximately a half of the sum of all the weights. Formally, we find the index k^* that satisfies the following process:

$$k^* = \min t \quad s.t. \quad \sum_{q=1}^{t} \bar{w}_q(p) \geq \frac{1}{2} \sum_{q=1}^{K} \bar{w}_q(p), \quad (6)$$

where $\bar{w}$ denotes an ordered confidence value. We create a final gradient field $\mathbf{g}$ by transferring a gradient sample corresponding to the index k^*:

$$\mathbf{g}(p) = \bar{\mathbf{g}}^{(k^*)}(p), \quad (7)$$

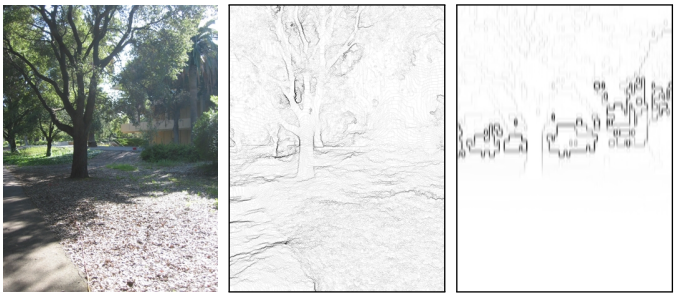
where $\bar{\mathbf{g}}$ is the ordered depth gradient samples of $\mathbf{g}$ in (4).

This weighted median voting process mitigates the outlier that may occur when simply selecting a gradient sample with the highest confidence from the training pairs, since having the highest confidence does not necessarily mean the most correct gradient sample. Fig. 5(b) shows the result estimated using the winner-takes-all (WTA) approach, where the gradient field is synthesized by selecting a gradient sample with the highest confidence. The WTA approach produces inaccurate results, while the proposed weighted median (WM) approach greatly alleviates the outliers as in Fig. 5(c).

Fig. 6 shows the gradient field estimated using $K = 7$ nearest neighbor training pairs. In the training pairs of Figs. 6(b) and (c), most homogeneous regions have almost zero gradients, while useful (strong) gradients are locally distributed around region boundaries that coincide with boundaries in the color image. This sparsity property inherent in the training pairs can also be observed in the estimated gradient field of Fig. 6(a). It should be noted that such a sparsity property of depth gradient fields helps improve a reconstruction performance, even when overall semantics of training data are not tightly correlated with those of an input image.

D. Depth from Gradients

The depth field D can be obtained by integrating the estimated gradient field $\mathbf{g}$. For this, it is required that the gradient field should have zero curl or should be integrable [14]. Numerous sophisticated algorithms have been developed in



(a) I, $\mathbf{g}$, and ground truth

(b) C_1 (c) C_2

Fig. 6. Estimated depth gradient field using $K = 7$ nearest neighbor training pairs: (a) input image and its estimated and ground truth gradient field. (b)-(c) Some of retrieved top $K = 7$ training pairs. For a better visualization, we drew the depth gradient maps in such a way that a lower intensity indicates a higher magnitude.

the field of surface-from-gradients [38] to address integrability enforcement issues. In our work, the least square approach proposed in [14] was employed.

For completeness of algorithm exposition, we briefly review [14]. When $\mathbf{g}(p) = (g_x(p), g_y(p))^T$ is given, a surface (depth field) D can be obtained by minimizing the following objective:

$$\min_D \int\int \left[(D_x - g_x)^2 + (D_y - g_y)^2 \right] dxdy \qquad (8)$$

where D_x and D_y denote respectively the gradient field of D along x- and y-axis. It is well known that the solution of (8) can be obtained by solving the Poisson equation with the Neumann boundary conditions, which is the form of

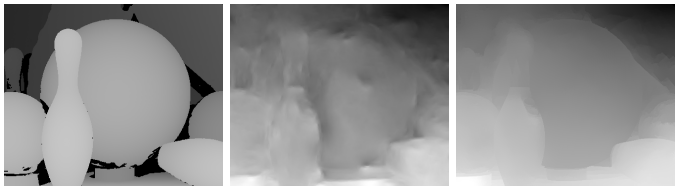
$$\nabla^2 D = div\ \mathbf{g} \quad with \quad \nabla D \cdot \hat{\mathbf{n}} = 0 \qquad (9)$$

where div is a divergence operator and $\hat{\mathbf{n}}$ is a normal vector perpendicular to the surface D. For numerical solution of (9), we refer readers to [14]. The surface D reconstructed using the Poisson solver [14] is shown in Fig. 7(b). The method captures a reasonable, natural depth field D in a global sense.

Note that many depth images D can be induced from the same gradient map $\mathbf{g}(p) = (g_x(p), g_y(p))^T$, since our approach transfers a set of depth gradients only from the training RGB-D database. Thus, we intentionally scale the resulting depth field to be in the range of $[0, 255]$. In this way, a relative depth order is still preserved, while satisfying standard depth encoding.

E. Post-processing Based on Joint Filtering

The depth estimate in Fig. 7(b) is, however, still noisy around some homogeneous regions, and region boundaries



(a) Ground truth (b) Poisson (PM-WM) (c) Refined using WMF

Fig. 7. Depth reconstruction of Fig. 5(a) with post-processing: (a) ground truth depth map, (b) Poisson reconstruction [14], and (c) refined using the weighted median filter (WMF) [15]. The higher the intensity, the closer the object.

are slightly inconsistent with those of the input image due to outliers in the depth gradient field. Here, we enhance it through a simple post-processing based on a joint filtering approach, which has been proved to be effective in improving labeling maps (*e.g.* depth or optical flow) [15], [39]–[41].

We employ modern edge-aware filters for smoothing the reconstructed depth image with the guidance of the input color image. In particular, we adopt a highly efficient weighted median filter (WMF) recently proposed in [15]. It is natural to use the weighted median filter, considering our observation about depth gradient statistics. The derivative statistics shown in Fig. 1(b) has a heavy-tailed distribution. Such a derivative prior can be better modeled by a Laplacian function rather than a Gaussian one [42]. Thus, the weighted median-based refinement makes the gradient distribution of resulting depth fields being matched more closely to that of natural images. Interestingly, similar observations were made by Saxena *et al.* [42] in the context of monocular depth estimation. They proposed to use Laplacian potentials in designing conditional random fields (CRFs) for max-margin parameter learning.

This naturally leads to using the weighted median filter that effectively solves the following L_1 minimization problem [43]:

$$\min_{D^*} \sum_{q \in \mathcal{N}(p)} \psi(p,q) \| D^*(p) - D(q) \|_1 \qquad (10)$$

where D^* is the filtered depth field and $\psi(p,q) = exp\{-\|I(p) - I(q)\|^2 / \sigma\}$ is a weighting function based on the affinity of two pixels p and q in the guide image I. $\mathcal{N}(p)$ is a set of neighboring pixels around p. Fig. 7(c) shows the depth field refined using the WMF [15]. Inconsistent regions are smoothed out, and the overall region boundaries are better aligned to those of the color image.

V. Experimental Results

We validated the performance of the proposed method against two competing algorithms qualitatively and quantitatively: the Depth Transfer (DT) algorithm [4] and the Depth Fusion (DF) algorithm [5]. All methods including ours are data-driven approaches using a large scale RGB-D training database. Note that the work of [8] also utilizes a large-scale RGB-D databases to infer a depth map, and showed that the accuracy of the depth estimation is slightly better than that of the DT method [4]. However, this work still relies on the appearance-depth assumption (by using mean values from RGB-D datasets), and thus it is expected to suffer from

TABLE I
DESCRIPTION OF RGB-D DATASETS USED IN EXPERIMENT

RGB-D Dataset	Abbreviation	Size	Usage
Make3D [6]	M3D134	134	Testing
	M3D	400	Training
NYU Kinect V2 [7]	NYU	1449	Training
Middlebury [44], [45]	MID	31	Testing

similar problems to the DT method, when training database is not tightly correlated with an input image. We obtained the results of DT by using the authors' MATLAB code[1], while the results for DF and ours were obtained using our own MATLAB implementation. All experiments were simulated on a PC with Quad-core CPU 2.93GHz. The codes and more results for various test images will be released at our project page later[2].

Regarding the number of training pairs K, we found that in our method, using more training pairs improves a depth reconstruction accuracy (see Fig. 12), but considering the trade-off between accuracy and runtime efficiency, we retrieved $K = 7$ training pairs in all experiments. For DT [4] and DF [5], we set K with the optimal setting reported in their papers: $K = 45$ in DF [5] and $K = 7$ in DT [4]. For post-processing, we set $\sigma = 20.0$ and the size of window $\mathcal{N}$ as 3×3. With this setting, the WMF post-processing is applied 5 times.

We designed training databases and test images with three publicly available RGB-D datasets consisting of real world color images and depth maps: the Make3D range dataset [6], the NYU Kinect V2 dataset [7], and the Middlebury stereo dataset (MID) [44]–[46]. The Make3D dataset (534 images) was taken from outdoor environment, while the NYU dataset (1449 images) and the Middlebury stereo dataset (31 images) were taken from indoor scenes. Table I summarizes the RGB-D datasets used in the experiment. By following standard practices used in the Make3D dataset [6], we define a set of 400 training images (M3D) as a database of RGB-D images $\mathcal{I}$ and the rest of 134 images (M3D134) as test images. Note that the Make3D color images are of 1704×2272 resolution, but the corresponding depth maps are of 305×55 resolution. We thus resized both color and depth images to the spatial resolution of 345×460 using the bilinear interpolation. For the NYU dataset, we define the entire 1449 images as a database of RGB-D images $\mathcal{I}$. While both color and depth images from the NYU dataset are of the same resolution of 640×480, the depth maps contain hole regions with no valid depth values. We thus neglected such hole pixels during the inference. The Middlebury stereo dataset (MID) are used as test images. Note that the MID dataset provides the disparity map, while the M3D and NYU training databases contain the depth maps. In our MID experiment, the ground truth disparity maps of the Middlebury dataset are acquired under a parallel stereo configuration. Thus, we can simply convert the Middlebury depth maps estimated using our approach in a form of the disparity map with a simple division, and then evaluate the accuracy with the ground truth Middlebury disparity maps. The depth maps of the test images (consisting of indoor and outdoor scenes) are synthesized using the *outdoor* M3D dataset or the *indoor* NYU dataset. This type of evaluation clearly shows the advantage of our approach, which is less sensitive to the training dataset used and resolves the depth ambiguity problem very well.

[1]http://www.kevinkarsch.com/

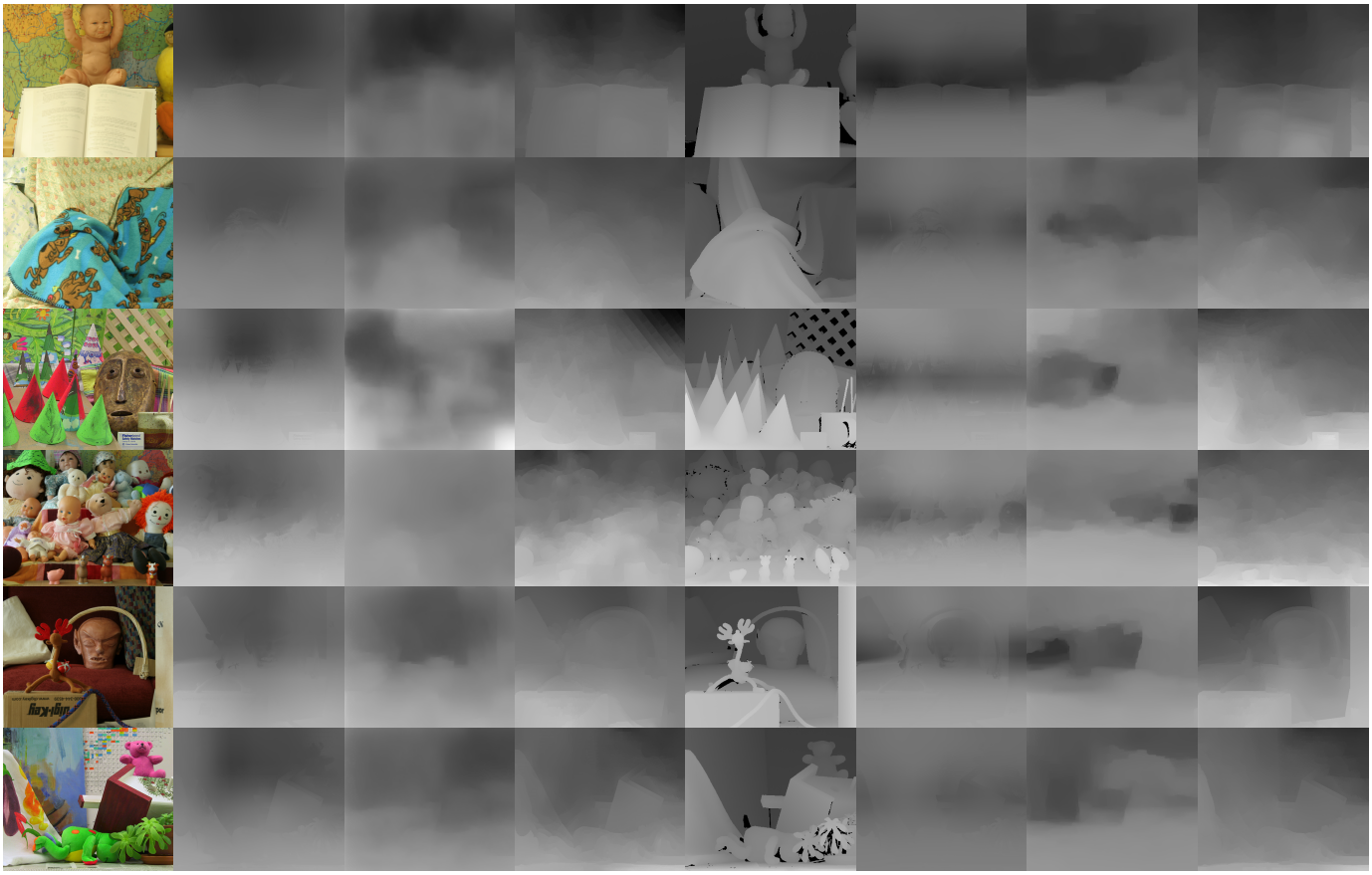
[2]http://sites.google.com/site/depthanalogy/

A. Qualitative Evaluation

We first evaluated the proposed method with 31 test images from the Middlebury indoor dataset (MID). Fig. 8 shows results obtained using two training databases. Figs. 8(b)-(d) present results using indoor training data (NYU), while Figs. 8(f)-(h) show results using outdoor training data (M3D). By using the NYU database, both competing algorithms produce a globally-correct depth map, but the result of DT as in Fig. 8(c) tends to be over-smoothed due to spatial regularization employed in depth interpolation process. In addition, locally inconsistent estimates are observed from the result of DF approach as in Fig. 8(b), due to their strict reliance on the appearance-depth correlation assumption and depth ambiguities in the training data. In contrast, our results better respect the discontinuities in the scene, demonstrating the superiority of locally-sampled depth gradients for reconstruction. When the outdoor M3D database is given, both competing algorithms fail to estimate desired depth structures due to different scene characteristics, as shown in Figs. 8(f) and (g). In contrast, our approach still produces a comparable depth field to the ground truth one.

In Fig. 9, we performed the experiments with test images from M3D134 dataset. Similar to the first experiment, we cross-validated our method by alternating the NYU and the M3D datasets for training. The outperformance of our method is easily confirmed by visually comparing overall structures of reconstructed depth maps. By comparing Figs. 9(d) and (h), it can be verified that our method does not severely depend on the training dataset. This also implies that the local gradient sampling process better aggregates useful depth cues.

Note that our gradient sampling process locally aggregates hypothetical cues by traversing the entire training data. This increases the possibility of finding useful reconstruction cues. The DT method [4] does this in a similar manner using the SIFT Flow [10], but a large 2D displacement vector is penalized in the objective used in the SIFT flow. Additionally, the SIFT flow constrains a search range within a predefined distance due to a heavy computational burden of the global optimization algorithm used. Such a dense alignment may lose the chance of getting more coherent candidates over a whole image (see Fig. 4(d)). Contrarily, our method achieves very convincing depth maps and does not severely depend on the training dataset thanks to 1) the gradient transfer model based on the PatchMatch local search and 2) a synergetic combination of surface-from-gradients methods and joint filtering. For qualitative evaluation, we also report stereo rendering results in Fig. 10 and Fig. 11.



(a) Input (b) DF-NYU (c) DT-NYU (d) Ours-NYU (e) GT (f) DF-M3D (g) DT-M3D (h) Ours-M3D

Fig. 8. Qualitative comparison: (a) input images from the MID dataset, (b)-(d) depth maps obtained by DF, DT, and ours using the NYU training dataset, (e) ground truth depth maps, and (f)-(h) depth maps obtained by DF, DT, and ours using the M3D training dataset. The higher the intensity, the closer the object.

TABLE II
COMPARISON OF DEPTH ESTIMATION ERRORS

Method	K	MID						M3D134					
		M3D			NYU			M3D			NYU		
		MS-SSIM	MGE	RMS*	MS-SSIM	MGE	RMS*	MS-SSIM	MGE	RMS*	MS-SSIM	MGE	RMS*
Baseline	All	0.702	4.17	28.7	0.742	4.14	29.1	0.805	3.19	16.0	0.786	3.19	16.7
DF [5]	7	0.726	4.22	30.9	0.730	4.15	29.8	0.800	3.24	16.5	0.787	3.21	16.5
DT [4]	7	0.721	4.20	30.4	0.734	4.23	33.1	**0.829**	3.21	**13.5**	0.787	**3.19**	16.9
Ours	7	**0.767**	**4.10**	**26.3**	**0.761**	**4.12**	**28.5**	0.803	**3.20**	15.8	**0.791**	3.20	**16.6**
DF [5]	45	0.720	4.22	32.5	0.745	4.12	28.1	0.795	3.23	17.5	0.792	3.20	16.4
DT [4]	45	0.741	4.13	29.1	0.746	4.16	30.2	**0.823**	**3.18**	**14.0**	0.787	**3.19**	17.2
Ours	45	**0.774**	**4.07**	**25.3**	**0.774**	**4.10**	**27.2**	0.805	3.20	15.4	**0.798**	3.20	**16.0**

B. Quantitative Evaluation

For a quantitative evaluation, we computed similarity scores between an estimated depth map and a ground truth depth map. Note that our approach produces a final depth map reconstructed from a relative depth order (gradient), whereas competing algorithms directly estimate absolute depth values, making it difficult to use existing evaluation metrics, *e.g.*, using relative error (REL), log_{10} error, and root mean squared (RMS) error. Interestingly, it was also reported in [4] that these metrics do not fully reflect an estimation quality of data-driven depth reconstruction algorithms.

Thus, we employ three different metrics for a quantitative evaluation. First, for direct comparison with competing algorithms, we report a scaled RMS error of an estimated depth map D^* computed from DF, DT, and ours against a ground truth depth map D^T:

$$RMS^* = \left(\sum_{p=1}^{N} \left(\alpha D^*(p) + \beta - D^T(p) \right)^2 / N \right)^{\frac{1}{2}},$$

where N is the number of pixels and α and β are parameters to reduce the effect of a scaling bias in the estimated depth map. We introduce free scaling parameters of α and β into the RMS metric in order to scale each relative (estimated) depth map to an absolute (ground truth) depth map, which best optimizes the RMS error metric. We determine α and β as the solution

(a) Input (b) DF-M3D (c) DT-M3D (d) Ours-M3D (e) GT (f) DF-NYU (g) DT-NYU (h) Ours-NYU

Fig. 9. Qualitative comparison: (a) input images from the M3D134 dataset, (b)-(d) depth maps obtained by DF, DT, and ours using the M3D training dataset, (e) ground truth depth maps, and (f)-(h) depth maps obtained by DF, DT, and ours using the NYU training dataset. The lower the intensity, the closer the object.

of $\min \sum_p \left(\alpha D^*(p) + \beta - D^T(p)\right)^2$ which can be given by linear regression [47]. In this way, direct comparison with competing algorithms can be reported. Second, we measure a root mean gradient error (MGE) in order to provide how well the estimated gradient field matches the ground truth:

$$MGE = \left(\sum_{p=1}^{N} \left\| \alpha \nabla D^*(p) - \nabla D^T(p) \right\|^2 / N \right)^{\frac{1}{2}},$$

where N is the number of pixels and α is the parameter estimated in RMS^*. Third, instead of measuring residual errors in the depth domain, we measure the multi-scale structural similarity (MS-SSIM) [48] between the scaled version of test depth maps, $\alpha D^* + \beta$, and the ground truth ones. Evaluation scores are averaged over all images in the test dataset. It should be noted that this kind of scaling method has also been employed in DF [5] and DT [4] for ensuring that resultant depth values are in the range of original training depth maps. For a fair comparison, we applied such a scaling method to the results of competing methods as well as ours.

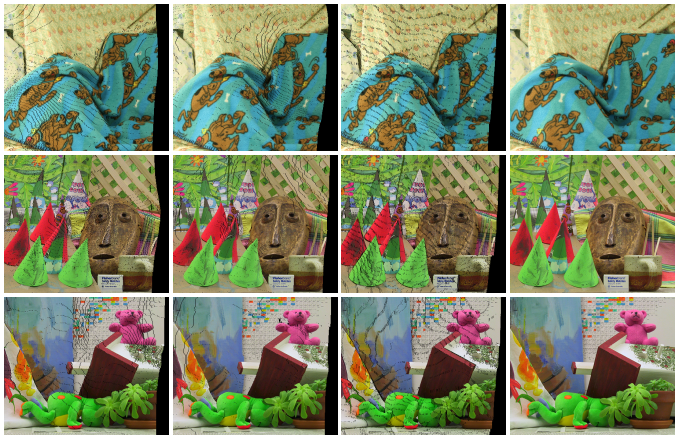
Table II reports correlation scores for the MS-SSIM metric and residual errors for the MGE and RMS* metrics. A higher score indicates a better quality for MS-SSIM, while a lower one is better for MGE and RMS* metrics. Here, the results of a baseline algorithm were obtained by simply averaging all depth maps (from each dataset) pixel-by-pixel. We compared our method with two competing algorithms, DF [5] and DT [4], by varying K. Note that the optimal parameter K reported in their papers is $K = 45$ in DF [5] and $K = 7$ in DT [4], respectively. In all experiments except the M3D134 test using the M3D dataset, the proposed method achieves a higher MS-SSIM value than competing algorithms. Also, our method is as good as or better than the competing algorithms for both MGE and RMS* metrics.

In the MID experiment, it should be noted that the scene semantics of the test images can be regarded as those of indoor scenes (NYU). However, the depth distribution of both test and training images does not match closely, causing depth ambiguities in competing algorithms. As a result, existing methods yield evaluation scores around 0.73, 4.18, 31 for MS-SSIM, MGE, and RMS* in all training datasets, while our method achieves 0.77, 4.10, 27 for these metrics. Our method outperforms existing methods in both $K = 7$ and $K = 45$.

In the M3D134 experiment, the scene characteristic of both the test image and the M3D training data matches closely. This satisfies the assumption of appearance-depth correlation

(a) Ground truth (b) Ours-M3D (c) Ours-NYU

Fig. 10. Stereo rendering results on the MID dataset. The anaglyph images are produced using the depth image-based rendering algorithm proposed in [4]. (a) Rendering results using the ground truth depth maps, (b)-(c) rendering results using depth maps obtained by ours with M3D and NYU training datasets.



(a) DF (b) DT (c) Ours (d) GT

Fig. 11. Results of depth image-based rendering on the MID dataset. Right images are synthesized using left color images and their estimated depth maps. (a)-(c) Synthesized results using depth maps obtained by (a) DF, (b) DT, and (c) ours with the M3D training dataset. (d) The ground truth right image. Note that hole regions are displayed in black.

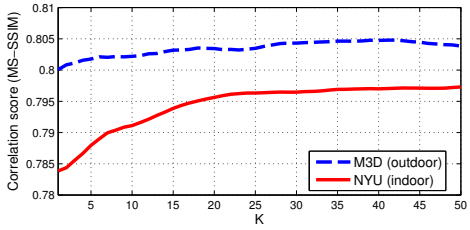


Fig. 12. Correlation curves with varying the number of training images K. Correlation scores are measured by using the MS-SSIM metric [48]. Given the M3D134 outdoor test images as input, scores are reported respectively on the training database as the M3D (dashed line) dataset and the NYU (solid line) dataset.

used in DF [5] and DT [4]. When $K = 7$, the DT provides the best results. However, our gradient-domain approach also produces comparable results to the state-of-the-art methods. Interestingly, it was reported in the DT [4] that $K = 7$ is optimal for the M3D134 dataset, and thus using more than $K = 7$ does not significantly improve results. In our approach, however, slightly better results can be achieved when using $K = 45$. When using an obviously uncorrelated training database (NYU), the proposed method shows slightly better performance than the existing methods, but no significant gains are observed when compared to the MID cases. It is because the ground truth depth maps of M3D134 dataset are of low resolution and coarsely quantized (see Fig. 9(e)), making an objective evaluation hard. But, we can find a significant improvement in Fig. 9 in terms of visual quality. We reserve a new metric better measuring the quality of depth maps as a future work. Although all the methods including ours perform worse than the previous case using the M3D training database, our method is less sensitive to the training dataset used than existing methods, and produces realistic depth maps (see Fig. 9(h)). This demonstrates that the depth gradient is a more informative cue than the depth value.

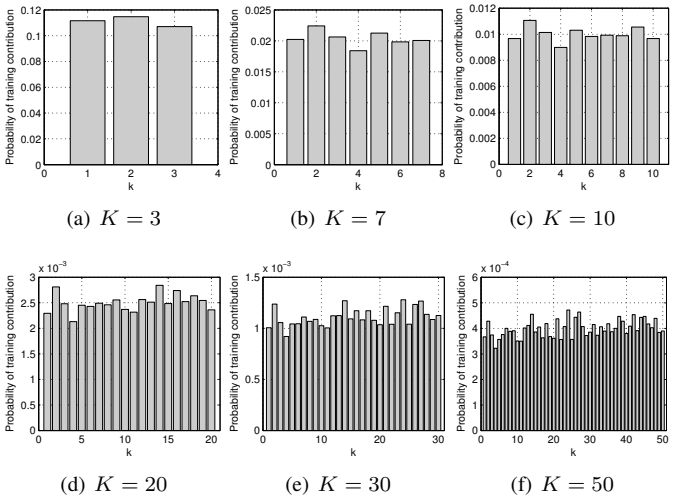
C. Analysis on Varying K

We evaluated the performance of our method according to varying the number of training images K. We used the M3D134 dataset as inputs, and alternated the M3D and NYU datasets for training.

Using the M3D134 outdoor test images as inputs, we evaluated the correlation score based on the MS-SSIM metric [48] averaged over all test images when using different values of $K = 1 \ldots 50$. Fig. 12 shows the correlation curves with varying K. As we can see the blue dashed line in Fig. 12, the correlation score slowly increases, when using correlated training data (*i.e.* outdoor vs. outdoor). In the case of using an uncorrelated training data as in the red solid line of Fig. 12, the correlation score abruptly increases by around $K = 20$ and slowly increases beyond $K = 20$. It is natural to expect that using more *relevant* images likely increases the possibility of finding useful reconstruction cues. Considering trade-off between accuracy and runtime efficiency, we set K to 7 for all experiments, but more accurate results are achievable with $K \geq 7$.

In DT [4], however, using more images from RGB-D database does not necessarily increase an accuracy of depth estimation. It is because all depth values aggregated from the RGB-D database do not always provide useful cues for depth reconstruction due to the strict assumption on appearance-depth correlation and the depth ambiguity problem.

We also studied how depth gradients obtained from each training depth map contribute to a final result. For instance, for $K = 3$, an occurrence histogram bin k $(= 1 \sim 3)$ is incremented by one, when the depth gradient of k^{th} training depth image is chosen in the final resultant depth map. Fig. 13 plots contribution histograms regarding $K = 3, 7, 10, 20, 30, 50$. Interestingly, the top matched $(k = 1)$ training sample does



(a) $K = 3$ (b) $K = 7$ (c) $K = 10$

(d) $K = 20$ (e) $K = 30$ (f) $K = 50$

Fig. 13. Analysis of the contribution of training samples used in single image depth estimation. All the training samples similarly contribute to the inference, indicating that using more training samples likely increases the possibility of aggregating proper depth cues.

TABLE III
ANALYSIS ON RUNNING TIME IN SECONDS (N.A.: NOT APPLICABLE)

Method ($K = 7$)	DF [5]	DT [4]	Ours
Scene retrieval	0.33		
Scene warping	N.A	121.09	6.84
Reconstruction	0.17	8.34	0.92
Total	0.5	129.76	8.09

not contribute to the majority of inference for all experiments. Instead, all the training samples similarly contribute to the depth inference. This analysis is also consistent with the intuition that using more training samples likely increases the possibility of aggregating proper depth cues.

D. Computational Complexity

In Table III, we measure the runtime of the proposed method and competing algorithms. The input image and training database are of 345×460 resolution, and $K = 7$ training image pairs are selected using the PHOG descriptor. The proposed method is about 16x faster than the DT method, while achieving a superior estimation quality. It is because our gradient sampling process is performed locally without using a global optimization. Most time-consuming part of the DT algorithm is the scene warping via SIFT Flow [10]. Our approach benefits from the computational efficiency of the ANN search algorithm [37] and the fast filtering technique [15]. This makes our approach a more efficient and practical solution to estimating depth from a single image. Note that the DF method does not employ a scene warping process. Instead, it computes the median of candidate depth maps and applies a fast bilateral filter from [49], taking 0.5 seconds in our own implementation. However, this method does not take into account local properties of retrieved K depth maps, producing unsatisfactory results (see Fig. 8(b) and (f)).

(a) Repeated textures (b) Gradient-like depth field

Fig. 14. Failure case of our method. When the input image (a) has repeated textures such as grass in background, our method is not able to estimate a correct depth gradient field due to the duplication of similar gradients. This results in a gradient-like linear structure as in (b).

VI. CONCLUSION

This paper has presented the gradient-domain single image depth estimation method using large scale RGB-D images, where depth gradients are sampled based on the visual correspondence mechanism. The locally-aligned depth gradient sampling strategy allows one to synthesize a plausible depth gradient field by accurately identifying good gradient samples from nearest neighbor training images. We have showed that the depth gradient serves as a better cue for data-driven depth inference. Our depth gradient aggregation approach is beneficial to dealing with various training range images involving substantial appearance and geometric variations. More importantly, our method is less dependent on the appearance-depth correlation assumption strictly imposed on previous methods, and thus is capable of estimating scene depth from limited training data in terms of a variety of scene semantics. Also, the synergetic combination of the reconstruction method based on Poisson solver and the edge-aware filter simplifies the depth inference framework, leading to a faster runtime efficiency, when compared to previous methods that require solving complex optimization problems [9], [50].

There are some limitations in our approach, though. First, it produces the relative depth order only, different from previous methods. The depth map is obtained by solving the Poisson equations. Thus, the reconstructed depth map has the scale ambiguity. However, we would like to point out that our primary goal is to estimate relative depth orderings, rather than absolute depth values. Actually, the effect of the scale ambiguity may vary depending on what applications the depth map is used for. For example, in virtual view synthesis [51], [52], the depth ordering is a more important factor than the depth value itself. Second, our method is unable to produce convincing results when an input image contains repeated textures as in Fig. 14. Our transfer model is established on the locally-aligned sampling strategy. Specifically, the patches from repeated textures in the input image are likely to be matched to a single patch in one of the training dataset. Then, the same depth gradient value obtained from the matched training data is assigned to these repetitive textures. This may cause gradient-like depth structures as in Fig. 14(b), since depth gradients in these regions monotonically increase or decrease.

REFERENCES

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1-3, pp. 7–42, 2002.

[2] L. Zhang, C. Vázquez, and S. Knorr, "3d-tv content creation: automatic 2d-to-3d video conversion," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 372–383, 2011.

[3] R. Zhang, P.-S. Tsai, J. Cryer, and M. Shah, "Shape-from-shading: a survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 8, pp. 690–706, Aug 1999.

[4] K. Karsch, C. Liu, and S. B. Kang, "Depth transfer: Depth extraction from video using non-parametric sampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2144–2158, Nov 2014.

[5] J. Konrad, M. Wang, P. Ishwar, C. Wu, and D. Mukherjee, "Learning-based, automatic 2d-to-3d image and video conversion," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3485–3496, Sept 2013.

[6] A. Saxena, M. Sun, and A. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, May 2009.

[7] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Proc. European Conf. Comput. Vis.*, 2012, pp. 746–760.

[8] M. Liu, M. Salzmann, and X. He, "Discrete-continuous depth estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 716–723.

[9] J. Konrad, G. Brown, M. Wang, P. Ishwar, C. Wu, and D. Mukherjee, "Automatic 2d-to-3d image conversion using 3d examples from the internet," in *Proc. SPIE (Electronic Imaging)*, 2012, pp. 82 880F–82 880F.

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

[11] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin, "Image analogies," in *Proc. Conf. Comput. Graph. Inter. Tech.*, 2001, pp. 327–340.

[12] P. Burt and B. Julesz, "A disparity gradient limit for binocular fusion," *Science*, vol. 208, no. 4444, pp. 615–617, 1980.

[13] H. BulthoffT, M. Fahle, and M. Wegmann, "Perceived depth scales with disparity gradientt," *Perception*, vol. 20, pp. 145–1, 1991.

[14] A. Agrawal, R. Raskar, and R. Chellappa, "What is the range of surface reconstructions from a gradient field?" in *Proc. European Conf. Comput. Vis.*, 2006, pp. 578–591.

[15] Q. Zhang, L. Xu, and J. Jia, "100+ times faster weighted median filter (wmf)," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014.

[16] O. Wang, M. Lang, M. Frei, A. Hornung, A. Smolic, and M. Gross, "Stereobrush: interactive 2d to 3d conversion using discontinuous warps," in *Proc. Euro. Symp. Sketch-Based Inter. Model.*, 2011, pp. 47–54.

[17] M. Liao, J. Gao, R. Yang, and M. Gong, "Video stereolization: Combining motion analysis with user interaction," *IEEE Trans. Vis. and Comput. Graph.*, vol. 18, no. 7, pp. 1079–1088, 2012.

[18] X. Cao, Z. Li, and Q. Dai, "Semi-automatic 2d-to-3d conversion using disparity propagation," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 491–499, 2011.

[19] Y. Horry, K.-I. Anjyo, and K. Arai, "Tour into the picture: using a spidery mesh interface to make animation from a single image," in *Proc. Conf. Comput. Graph. Inter. Tech.*, 1997, pp. 225–232.

[20] L. Zhang, G. Dugas-Phocion, J.-S. Samson, and S. M. Seitz, "Single-view modelling of free-form scenes," *J. Vis. Comput. Ani.*, vol. 13, no. 4, pp. 225–235, 2002.

[21] D. Kim, D. Min, and K. Sohn, "A stereoscopic video generation method using stereoscopic display characterization and motion analysis," *IEEE Trans. Broadcast.*, vol. 54, no. 2, pp. 188–197, 2008.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[23] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 257–266, 2002.

[24] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1253–1260.

[25] L. Ladicky, J. Shi, and M. Pollefeys, "Pulling things out of perspective," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 89–96.

[26] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Info. Process. Sys.*, 2014, pp. 2366–2374.

[27] D. F. Fouhey, A. Gupta, and M. Hebert, "Data-driven 3d primitives for single image understanding," in *Proc. Int'l Conf. Comput. Vis.*, 2013, pp. 3392–3399.

[28] B. Zeisl, M. Pollefeys *et al.*, "Discriminatively trained dense surface normal estimation," in *Proc. European Conf. Comput. Vis.*, 2014, pp. 468–484.

[29] A. Torralba and A. Oliva, "Depth estimation from image structure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1226–1238, 2002.

[30] J. Huang, A. Lee, and D. Mumford, "Statistics of range images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, 2000, pp. 324–331.

[31] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[32] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proc. ACM Int. Conf. Image and Video Retrieval*, 2007, pp. 401–408.

[33] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, 2000, pp. 142–149.

[34] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[35] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1469–1472.

[36] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman, "Patchmatch: a randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009.

[37] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein, "The generalized patchmatch correspondence algorithm," in *Proc. European Conf. Comput. Vis.*, 2010.

[38] H.-S. Ng, T.-P. Wu, and C.-K. Tang, "Surface-from-gradients without discrete integrability enforcement: a gaussian kernel approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2085–2099, 2010.

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

[40] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2432–2439.

[41] 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, 2012.

[42] D. Batra and A. Saxena, "Learning the right model: Efficient max-margin learning in laplacian crfs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2136–2143.

[43] Y. Li, S. Osher *et al.*, "A new median formula with applications to pde based denoising," *Commun. Math. Sci*, vol. 7, no. 3, pp. 741–753, 2009.

[44] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, 2003, pp. I–195.

[45] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[46] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[47] N. R. Draper and H. Smith, *Applied regression analysis*. John Wiley & Sons, 2014.

[48] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. IEEE Conf. Signals, Systems and Computers*, vol. 2, 2003, pp. 1398–1402.

[49] S. Paris and F. Durand, "A fast approximation of the bilateral filter using a signal processing approach," in *Proc. European Conf. Comput. Vis.*, 2006, pp. 568–580.

[50] K. Karsch, C. Liu, and S. B. Kang, "Depth extraction from video using non-parametric sampling," in *Proc. European Conf. Comput. Vis.*, 2012.

[51] S. Choi, B. Ham, and K. Sohn, "Space-time hole filling with random walks in view extrapolation for 3d video," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2429–2441, 2013.

[52] B. Ham, D. Min, C. Oh, M. N. Do, and K. Sohn, "Probability-based rendering for view synthesis," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 870–884, 2014.