SuperDepthTransfer: Depth Extraction from Image Using Instance-Based Learning with Superpixels

Yuesheng Zhu, Yifeng Jiang, Zhuandi Huang, and Guibo Luo

Shenzhen Key Lab of Information Theory & Future Network Arch, Communication & Information Security Lab, Institute of Big Data Technologies
Shenzhen Graduate School, Peking University
Shenzhen, Guangdong 518055 - China
[e-mail: zhuys@pkusz.edu.cn, jiangyifeng@sz.pku.edu.cn, 1101213414@sz.pku.edu.cn, luoguibo@sz.pku.edu.cn]
*Corresponding author: Yuesheng Zhu

Received November 18, 2016; revised March 26, 2017; revised May 5, 2017; accepted May 28, 2017; published October 31, 2017

Abstract

In this paper, we primarily address the difficulty of automatic generation of a plausible depth map from a single image in an unstructured environment. The aim is to extrapolate a depth map with a more correct, rich, and distinct depth order, which is both quantitatively accurate as well as visually pleasing. Our technique, which is fundamentally based on a preexisting DepthTransfer algorithm, transfers depth information at the level of superpixels. This occurs within a framework that replaces a pixel basis with one of instance-based learning. A vital superpixels feature enhancing matching precision is posterior incorporation of predictive semantic labels into the depth extraction procedure. Finally, a modified Cross Bilateral Filter is leveraged to augment the final depth field. For training and evaluation, experiments were conducted using the Make3D Range Image Dataset and vividly demonstrate that this depth estimation method outperforms state-of-the-art methods for the correlation coefficient metric, mean log10 error and root mean squared error, and achieves comparable performance for the average relative error metric in both efficacy and computational efficiency. This approach can be utilized to automatically convert 2D images into stereo for 3D visualization, producing anaglyph images that are visually superior in realism and simultaneously more immersive.

Keywords: Depth estimation, instance-based learning, superpixels, semantic label, 2D-to-3D conversion

1. Introduction

In sharp contrast with direct stereo shooting, a process that can extinguish 3D viability for promotion in film and television, 2D-to-3D conversion mandates efficient conversion of massive amounts of existing 2D content into 3D. Thus, 2D-to-3D conversion is gaining momentum as a pursuit of great significance.

Normally, the 2D-to-3D conversion process is distributed into two basic steps: depth estimation for a given 2D image and subsequent Depth Image Based Rendering (DIBR) of a query image to form a stereopair. The rendering step is well understood, and algorithms exist that produce satisfactory results. The fundamental remaining challenge lies in the means to extract or infer accurate depth from an image. Notably, with increasing depth information comes the critically important concomitant capacity to reproduce parallax for 3D display technology. An ultimate aim is development to the point of implementation in applications such as 3D cinema, advertising, TV and desktop displays, among others [1]. Depth extraction, therefore, is the focus and concentration of this paper.

A novel solution is presented in this report that generates a depth map that excels ordinary 2D image based DepthTransfer algorithms by increasing fidelity, richness, and obvious depth order [2] [3]. Improvements are principally four-fold:

- *The first proposal is to transfer depth information at the level of Superpixel, which can provide superior spatial support for aggregating features that could belong to the same object and region. This also improves the accuracy of mapping, helps keep edge information of objects in the scene, and additionally, is computationally efficient in contrast with pixel-based approaches.*
- *Second, depth information is incorporated into graph based image segmentation to define a novel pre-processing candidate set.*
- *Next, there is simultaneous estimation of the semantic label and depth value for the same region. With the benefit of a known semantic label, depth transfer can be simpler and more precise.*
- *Finally, employment of a cross bilateral filter, rather than sophisticated global optimization for depth map smoothing, is proposed. A cross bilateral filter not only can effectively eliminate a blocking effect; It can also help align the depth edge with the query image edges, all the while preserving a globally-consistent depth of the preliminary estimate.*

As an improved version of the Depth Transfer algorithm, this method is applicable to arbitrary images, and works well in cases where conventional depth recovery methods fail. It is clearly demonstrated that the depth map extracted from this proposed algorithm possesses a more accurate and richer depth level than state-of-the-art methods, and moreover, is sufficient to generate compelling 3D images when applied in 2D-to-3D conversion, as illustrated in **Fig. 1**.

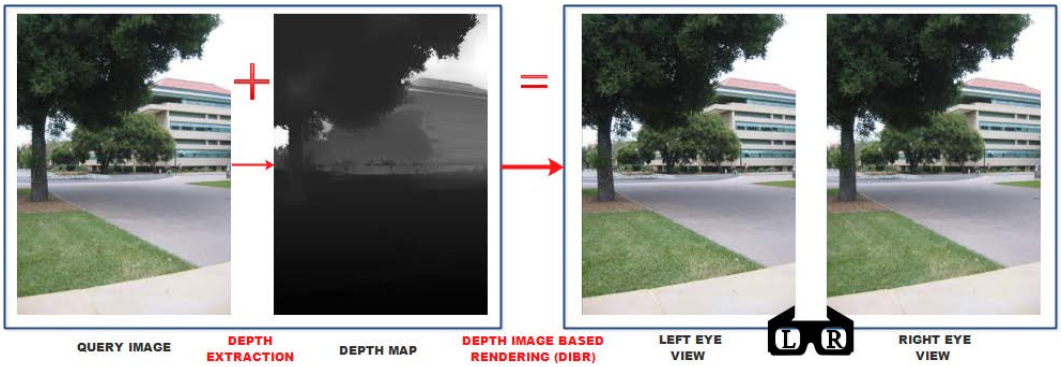


Fig. 1. Example of 2D-to-3D conversion resulting from application of the improved version of the DepthTransfer method. The left end shows the input query image and its corresponding extracted depth generated by this novel method. The right end is the result of the performance of DIBR and its respective yield of binocular vision.

2. Related Work

Depth map contains spatial information about the distance of the surfaces of scene objects from a viewpoint. This better addresses appearance, viewing angle, and complex lighting condition variations. Additionally, the depth map can provide complementary visual spatial information for a conventional planar RGB image, which only has three channels of color information. Therefore, to date, the depth map has been regarded as a rather vital source of visual data, and employed in various research fields to attain more promising results, such as 2D-to-3D conversion [4], 3D reconstruction [5], augmented reality & virtual reality, pattern recognition (e.g. human action recognition [6], 3D object detection [7], and subspace learning [8]), automatic driving [9], and other applications. Apparently, depth extraction (or depth estimation) is increasingly becoming a subject of active research by the computer vision community.

Depth Extraction can be roughly categorized into semiautomatic and automatic approaches. In the semiautomatic method, a skilled operator assigns depths to various parts of an image or video, which with error correction, can successfully yield a plausible depth map [10]. Depth assignment is also included in this category [11]. The latter automatic depth estimation method requires no operator intervention, as an elegant algorithm automatically estimates the depth—a more cost efficient approach.

Typically, conventional automatic depth extraction methods rely on robust assumptions including shape from shading [12], structure from motion [13], depth from defocus [14], depth from visual saliency [15], depth from perspective geometry [16] and so on. Despite that such methods have proved to be applicable to a restricted set of scenarios, they are suboptimal for arbitrary scenes.

Machine-learning-inspired methods have recently been proposed to automatically estimate the depth map of a single monocular image. Compared to conventional methods, machine-learning-based methods make no assumptions on the use of scenarios, and lead to more accurate depth mapping. Hoiem et al. [17] created a convincing reconstruction of outdoor images by assuming an image could be broken into a few planar surfaces, and pixels could then be classified into limited labels, e.g., ground, sky, and vertical walls. Similarly, Delage et al. [18] developed a Bayesian framework for reconstructing indoor scenes. This framework would comprise a 3D algorithm [19, 20], and devise a supervised learning strategy

to infer the absolute depth of each pixel in the monocular image. This would assume that most 3D scenes are made up of numerous small, approximately planar surfaces. With the use of Markov Random Field (MRF), Saxena et al. model both the monocular depth cues as well as the relationship between different parts of image. Make3D has been further improved to create realistic reconstructions for general scenes. The Semantic-Label algorithm [21] achieves better depth estimates by incorporating semantic labels to guide the 3D reconstruction process. Notably, though, such additional cues are not generally available. To address this issue, nonparametric methods [22 - 25] have been introduced. To produce the most likely depth map, these approaches proceed by firstly matching the high-level image features. This enables a search for candidates from a repository of 3D images (RGB plus corresponding Depth or stereopairs) that are photometrically similar. Then the depth of these candidates are fused in a variety of ways. Taking inspiration from this, Miaomiao Liu et al. implemented a nonparametric approach for the retrieval of candidate depth maps. They avoided the immediate fusion process by formulating depth estimation as an inference in a discrete-continuous graphical model [26].

Accelerated interest has appeared more recently in employing deep learning methods [27 - 29]. There have been attempts to regress depth immediately from the image [27]. An associated serious implication is over-fitting, as such a procedure might mandate hundreds of thousands of examples to train its model. Notably, when in short supply of training data, substandard performance is the result. Learning deep features for inferring superpixels depths directly, and subsequently trying to enforce coherence with a CRF to capture scene structure is examined by [28]. In a distinct departure from the above two methods, [29] regresses on a small set of depth reconstruction weights, and harnesses statistical regularities to the problem. This is the first application of couple dictionary learning and regression to depth estimation.

In addition, many more probabilistic modeling approaches have been proposed recently, as reported for instance in [30] and [31]. Enlightened by psychophysical evidence of visual processing in human Vision System (HVS) and Natural Scene Statistics (NSS) models of image and range, [30] proposed a Bayesian framework to recover the range information from monocular image by adopting the statistical relationships between luminance and depth in natural scenes. In [31], Xiaoyan Wang et al. proposed a depth estimation conditional random field (CRF) model with the field of experts (FoE) as the prior.

The DepthTransfer algorithm [2] [3] is closely allied to the nonparametric depth sampling method proposed by J. Konrad et al. in [24, 25]. It leverages an instance-based learning framework to extract a depth map from monocular images. Then it automatically converts them to stereoscopic images. Both make similar assumptions, e.g., appearance is correlated with depth. The DepthTransfer algorithm adheres to the “big data” philosophy of machine learning, and is inspired by the recent trend to employ large image databases for various computer vision tasks, such as object recognition [32] and image saliency detection [33]. In contrast to a model-based method like Make3D [19, 20] or Semantic-Label [21], an instance-based learning method like DepthTransfer avoids explicit definition of a parametric model. It also requires fewer assumptions and no training time. It scales better with respect to the training data size, and is still capable of generating a compelling 3D image.

We build on this work by transferring depth to superpixels rather than pixels, and with the assistance of semantic labels, have enhanced depth warping precision. Additionally, modified Cross Bilateral Filter controlled by the query image has been leveraged into refining our final result.

3. System Description

This system shares the DepthTransfer [2, 3] principle that two images that are photometrically similar also have a similar 3D structure and depth distribution. The depth extraction approach, as outlined in **Fig. 2**, has four stages:

(1) **Candidate Set Construction**: Given a database with known depth and semantic information, retrieve candidate images whose depth may approximate that of the query image (input 2D image).

(2) **Pre-process**: Segment both the query 2D image and candidate images into superpixels.

(3) **Depth map warping and fusion with semantic labeling at the superpixel level**: Based on the selected candidate set, estimate the semantic label of superpixels in the input image. Next, generate a superpixel-level image mapping between the query and candidate images. Finally, create an initial depth map via weighted fusion of multiple warped candidate depth maps.

(4) **Depth map correction**: Guided by the query image, harness the modified Cross Bilateral Filter to correct the initial depth map.

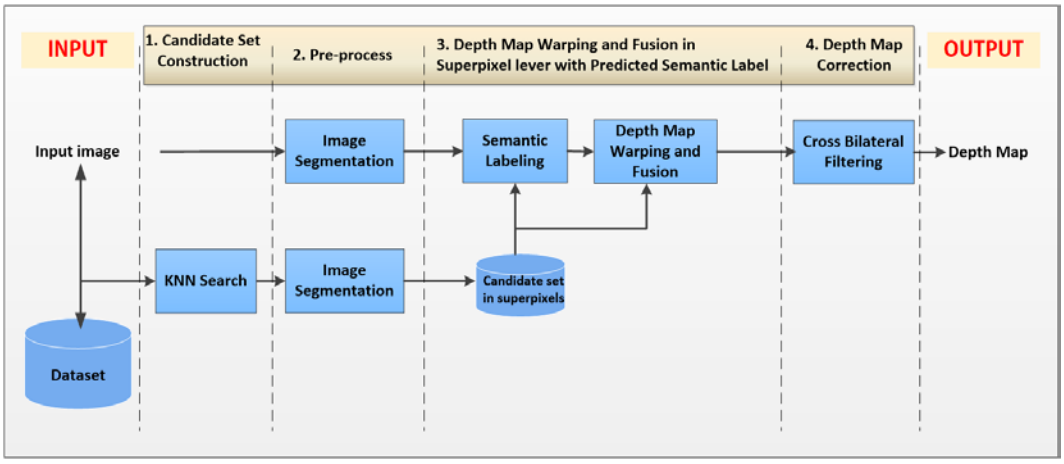


Fig. 2. Block diagram of the proposed overall algorithm.

3.1 Candidate Set Construction

Similar to various other data-driven methods, the first step is to find a relatively small candidate set from training images that can serve as the source of candidate matches at the superpixels level. This is accomplished not only for computational efficiency and tractability, but also to provide a scene-level context for the subsequent superpixels matching step.

Given an input image, high-level features are computed of each image both in the dataset and the input image. To determine the matching score between two images, the GIST descriptor [34] is leveraged, which provides a holistic representation of the scene by measuring its global properties. Then the top K (a relatively very small portion of the dataset) matching images (i.e., KNN algorithm) are selected as a subset. This is necessary because images that are not photometrically similar to the input image are ineffective for estimating the depth. Although it incurs a loss of several depth relevant images, a significant reduction in volume of involved images must be accomplished. Remaining matching images are called candidate images, and their corresponding depths called candidate depths.

3.2 Pre-Process

At the pre-process stage, the input and candidate set images are segmented into superpixels. From that point on superpixels are regarded as the basic processing unit for subsequent processing. Superpixels [35] are local, coherent regions, which not only reduce the complexity of the problem, but also give better spatial support for aggregating features that could belong to a single object than what pixels do. Superpixels are generated by exploiting the fast graph-based segmentation algorithm developed by Felzenszwalb and Huttenlocher [36].

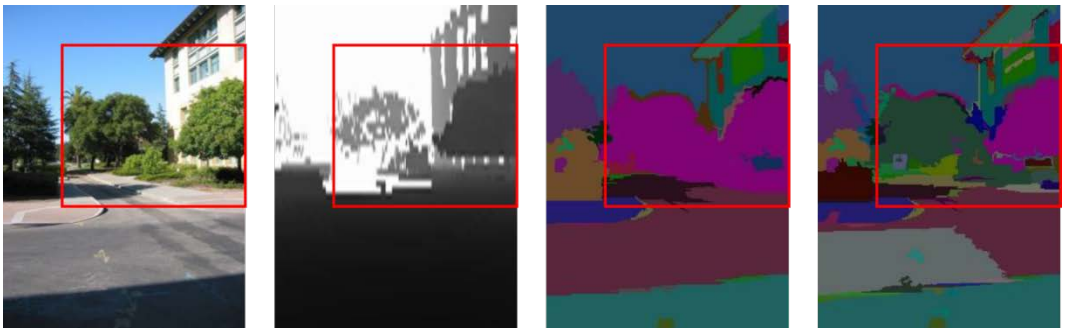


Fig. 3. The comparison of conventional fast graph-based segmentation method and our improved one incorporating corresponding depth information. The left two images are the RGB image to be segmented and its matching depth map from the dataset. The third image is the result of conventional fast graph-based segmentation algorithm, and the last one is the segmented result of ours.

For candidate images with known corresponding depth maps, an improvement on the fast graph-based segmentation algorithm [36] is achieved by incorporating RGB color and depth features into a similarity measure between adjacent nodes, as shown in **Algorithm 1**. This results in better segmentation, as illustrated in **Fig. 3**, possessing consistency in appearance as well as in depth. Admittedly, with the assistance of depth information, the improved segmentation method can naturally seize and recognize more object details due to the known depth of the scene.

Algorithm 1: Improved Fast Graph-Based Segmentation.

Input: RGB image and its corresponding Depth map (RGBD).
Output: The segmented RGB image based on the Depth map.
Namely, $Seg = (C_1, C_2, \dots, C_r)$.

1. Relative to the RGBD image, construct a graph $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, with $\boldsymbol{n}$ vertices and $\boldsymbol{m}$ edges. The vertices Set is $\boldsymbol{V} = (v_1, v_2, \dots, v_n)$ and the edge set is $\boldsymbol{E} = (e_1, e_2, \dots, e_m)$. The edge weight is calculated by **Eq.1**.
2. Sort $\boldsymbol{E}$ by non-decreasing edge weight.
3. Start with an initial segmentation Seg^0, where each vertice is in its own component, i.e., $Seg^0 = (\{v_1\}, \{v_2\}, \dots, \{v_n\})$.
4. **for** t = 1 to m **do**
5. **if** $|\boldsymbol{e_t}| < \boldsymbol{min}\left\{\boldsymbol{l}_i + \frac{\boldsymbol{k}}{|\boldsymbol{C}_i^{t-1}|}, \boldsymbol{l}_j + \frac{\boldsymbol{k}}{|\boldsymbol{C}_j^{t-1}|}\right\}$, $\boldsymbol{k}$ is a constant.
6. Merge $\boldsymbol{C}_i^{t-1}$ and $\boldsymbol{C}_j^{t-1}$.
7. **else** $Seg^t = Seg^{t-1}$.
8. **End for**
5. Return $Seg = Seg^m$

where C_i^{t-1} indicates the component of Seg^{t-1} containing v_i and C_j^{t-1} containing v_j in the $t-1th$ iteration. Let e_t be t-th edge connecting v_i and v_j in the ordering, i.e., $e_t = (v_i, v_j)$. l_i and l_j denote the maximum edge in the Minimum Spanning Tree (MST) of C_i^{t-1} and C_j^{t-1} separately.

$$w(v_i, v_j) = \sqrt{\left\|I_{RGB}(x_i) - I_{RGB}(x_j)\right\|_2^2 + \lambda\left\|D(x_i) - D(x_j)\right\|_2^2} \tag{1}$$

After segmentation, the depth of a superpixel is defined as the mean depth of pixels within its region, and the semantic label of a superpixel is the category with the highest frequency within its region.

$$d(s) = \frac{1}{N_s}\sum_{x \in s} D(x) \tag{2}$$

$$\omega(s) = arg\max_{\omega_j}\{count(s, \omega_j)\} \tag{3}$$

where $x \in s$ represents a single pixel that belongs to its superpixel s and N_s denotes the number pixels within it. d(s) and $\omega(s)$ are the defining depth and semantic label of the superpixel s. $count(s, \omega_j)$ is the number of pixels with identical label ω_j in superpixel s.

To describe superpixels more succinctly, 16 different features somewhat in alignment with those of Malisiewcz and Efros' work [37] are adopted. These have influenced a number of modifications and additions. A complete list of the feature descriptors is shown in **Table 1**.

Table 1. Features for Superpixels.

Type	Name	Dimension
Shape	Mask of superpixels shape over its bounding box (8×8).	64
	Bounding box width/height relative to image width/height.	2
	Superpixels area relative to the area of the image.	1
Location	Mask of superpixels shape over the image.	64
	Top height of bounding box relative to image height.	1
Color	RGB color mean and std dev.	3×2
	Color histogram (RGB, 11 bins per channel, dilated by 10 pixels color histogram.	33×2
DSIFT	Quantized SIFT histogram, dilated by 10 pixels quantized SIFT histogram	100×2
	Left / right / top / bottom boundary quantized SIFT histogram.	100×4
Textures	Statistics texture energy and texture kurtosis of the filter bank containing of 17 filters	34

3.3 Depth map warping and fusion

3.3.1 Semantic Labeling

Once the input image is segmented and the features $f_1, f_2, \ldots, f_M$ $(M = 16)$ as described in **Table 1** of all superpixels are extracted, a log likelihood ratio score $logLR(s|\omega_j)$ for each target superpixel s, as well as for each semantic class ω_j present in the candidate set, is attained. As reported in [38], the log likelihood ratio is defined as (4) and (5), where $\overline{\omega_j}$ is the set of all classes excluding ω_j. Each likelihood ratio is computed with the help of nonparametric density estimates of features from the required classes in the neighborhood of f_m.

$$\widehat{\omega}(s) = arg\max_{\omega_j} logLR(s|\omega_j) \tag{4}$$

$$logLR(s|\omega_j) = log\frac{P(s|\omega_j)}{P(s|\overline{\omega_j})} = log\prod_{m=1}^{M}\frac{P(f_m|\omega_j)}{P(f_m|\overline{\omega_j})} \tag{5}$$

Specifically, let D denote the set of superpixels in the training set, and N_m be the set of all superpixels in the candidate set whose feature distance from f_m is below a fixed threshold t_m. Thus :

$$\frac{P(f_m|\omega_j)}{P(f_m|\overline{\omega_j})} = \frac{n(\omega_j, N_m)/n(\omega_j, D)}{n(\overline{\omega_j}, N_m)/n(\overline{\omega_j}, D)} \tag{6}$$

where $n(\omega_j, S)$ is the number of superpixels in set S with class label ω_j. The superpixel neighbors N_m are found by nearest search measured by Euclidean distance.

3.3.2 Depth Map Warping

The texture feature has high utility to describe the target superpixels with a known semantic label. Similarity can be defined as $sim(s_i, s_j)$ between the two superpixels s_i and s_j as in (7). When belonging to an identical semantic class, similarity is non-negative, and when they have different semantic labels, the similarity is equal to zero.

$$sim(s_i, s_j) = \begin{cases} \frac{1}{1+exp(\left(\left\|f_{s_i} - f_{s_j}\right\|_2\right)-\mu)/\sigma} & ,if\ \ \omega(s_i) = \omega(s_j) \\ 0 & ,otherwise \end{cases} \tag{7}$$

$$S_i^k = arg\max_{S_j \in I^k}\left(sim(s_i, s_j)\right) \tag{8}$$

where μ and σ are constants. f_{s_i} denotes the texture feature of superpixel s_i. Through the similarity comparison, every superpixel s_i in the input image will match the most similar superpixel S_i^k in each candidate image I^k. We then define this corresponding superpixel S_i^k in (8) as the one which possesses the maximum similarity among all superpixels $S_j \in I^k$.

With similarity and mapping principle thus defined, the input image is warped with each respective candidate image at the superpixels level, as shown in **Fig. 4**. The weight of the warping is precisely defined by the similarity between two matched superpixels.

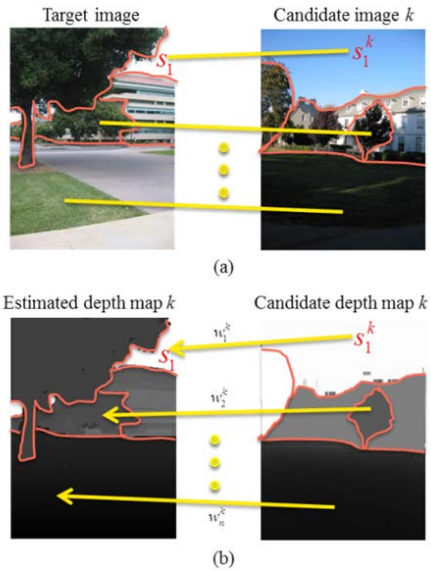


Fig. 4. Depth Map Warping in Superpixels Level.

3.3.3 Depth Map Fusion

Each warped candidate depth figures to be a rough approximation of the target's depth map. Thus the initial depth map can be computed by applying the mean operator across the warped candidate depth maps at each superpixel as (9) describes, where s_i is superpixel i in target image and $\hat{d}$ is its fusing depth. s_i^k is the matched superpixel in candidate image k, of which d^kdenotes its depth and w_i^k is its responding weight defined by (10). This process is depicted in **Fig. 5** below.

$$\hat{d}(s_i) = \frac{\sum_{k=1}^{K} w_i^k d^k\left(s_i^k\right)}{\sum_{k=1}^{K} w_i^k} \tag{9}$$

$$w_i^k = sim(s_i, s_i^k) \tag{10}$$

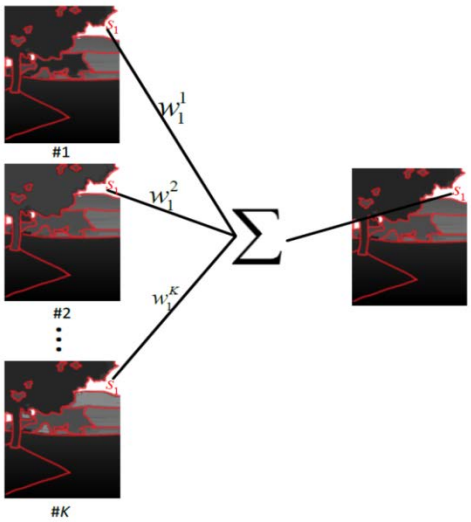


Fig. 5. Depth Map Fusion in Superpixels Level.

3.4 Depth Map Correction

At this point, the fused depth may suffer inaccuracies and lack certain spatial smoothness. To alleviate this, modified cross bilateral filtering (CBF), as demonstrated by [39, 40], is applied. CBF is an edge-preserving image smoothing method that basically applies anisotropic diffusion controlled by local image content [39]. In keeping with Konrad et al. [22], the original is modified and the external input RGB image guided the diffusion.

Modified CBF is defined in (11) (12) (13), where x represents the pixel, and y denotes the eight connected neighbor pixels around x. D and D' are the depth map before filtering and after filtering. $w(x, y)$ is the weight of pixel y to x defined in (12), which is determined not only by space discontinuities h_{σ_s} of the internal fused depth field but also by luminance discontinuities h_{σ_r} of the external RGB query image. In this fashion, the final result preserves the global depth properties, but smooths the depth field among superpixels while still keeping the depth edges sharp and aligned with the query RGB image structure. h_σ is a Gaussian weighting function. We display the qualitative effect of depth map correction in **Fig. 6** as follows.

$$D'(x) = \frac{\sum_y D(y)w(x,y)}{\sum_y w(x,y)} \tag{11}$$

$$w(x,y) = h_{\sigma_s}(y-x)h_{\sigma_r}(I(y)-I(x)) \tag{12}$$

$$h_\sigma(\cdot) = \frac{1}{2\pi\sigma^2} exp(-\frac{\|\cdot\|^2}{2\sigma^2}) \tag{13}$$

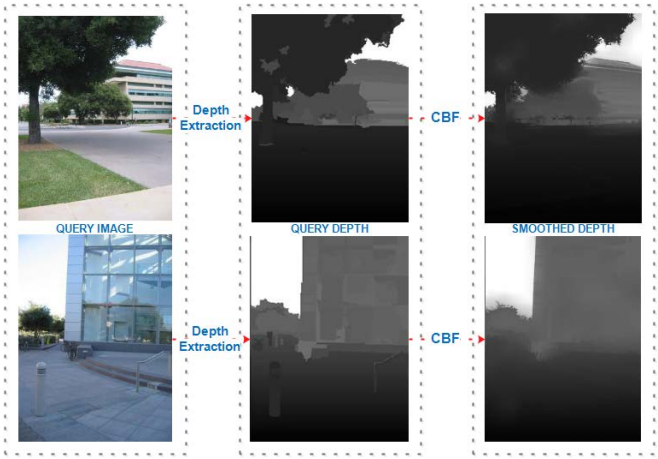


Fig. 6. An illustration of depth map correction. The left column is the RGB query image, and the middle column is its corresponding estimated depth field. After modified CBF processing, the smoothed result is demonstrated in the right column. It is apparent that the final depth map is entirely smoothed, while depth edges, if any, are aligned with features in the query image.

4. Experiments

The approach is tested on the publicly available dataset from Saxena et al.: The Make3D dataset #1[19] [41]. It is composed of 534 outdoor images with corresponding depth maps generated by a laser range finder. The corresponding semantic class labels are hand-annotated by Koller et al. as one of: sky, tree, road, grass, water, building, mountain, and foreground

object [21]. The Make3D images are of 1704×2272 resolution, but notably the corresponding fields are only of 55×305 spatial resolution. The sensor used to collect ground truth measurements has a range of 81m. The dataset is divided into 400 training samples and 134 testing samples. All images were resized to 240×320 before performing the experiment.

The proposed algorithm is compared with four state-of-the-art learning-based depth extraction methods: Depth MRF [20], Make3D [19], SemanticLabel [21] and DepthTransfer [2, 3]. Depth MRF takes the advantage of discriminatively-trained Markov Random Field (MRF) considering multiscale local- and global-image features to model depths at individual points and relation between depths at diverse points. Make3D estimates a 3D scene structure from a single still image of an unstructured environment by supervised learning of 3D position and orientation of small homogeneous patches in the image. The SemanticLabel algorithm first performs a semantic segmentation of the scene and uses the semantic labels to guide the 3D reconstruction. DepthTransfer consists of finding nearest neighbors using high-level features, followed by SIFT-flow to warp the depth fields to the current image. It also performs optimization to combine the warped depths while imposing a smoothness constraint and a global depth prior.

For quantitative performance evaluation of separate algorithms, the selected metrics are average relative error (rel), correlation coefficient (ρ), mean log10 error ($log_{10}\ error$) and root mean square ($RMS\ error$). These are defined in (14), (15), (16) and (17) respectively, where $D(x)$ is the ground truth depth for pixel x, $\widehat{D}(x)$ is the estimated depth, N indicates the number of pixels in the input image, and μ_D, $\mu_{\widehat{D}}$ are the empirical means of D and $\widehat{D}$, while σ_D and $\sigma_{\widehat{D}}$ are the corresponding empirical standard deviations. Quantitative measures are averaged over all images in the test set.

$$rel = \frac{1}{N}\sum_x\left(\left|\widehat{D}(x) - D(x)\right|/D(x)\right) \tag{14}$$

$$\rho = \frac{1}{N\sigma_{\widehat{D}}\sigma_D}\sum_x(\widehat{D}(x) - \mu_{\widehat{D}})(D(x) - \mu_D) \tag{15}$$

$$log_{10}\ error = \frac{1}{N}\sum_x\left|log_{10}\left(\widehat{D}(x)\right) - log_{10}(D(x))\right| \tag{16}$$

$$RMS\ error = \sqrt{\frac{1}{N}\sum_x(\widehat{D}(x) - D(x))^2} \tag{17}$$

An important parameter to be determined in the proposed algorithm is the size K of the candidate set. To obtain a proper K, both metrics are evaluated over the entire dataset by assigning different values to K among $\{5,\ 10, 30, 50, 100, 200\}$. Both quantitative metrics rise rapidly when a comparatively small k value is adopted, a maximum is attained upon an increase to $K = 30$, and then a tapering off takes effect. Therefore, ongoing experiments use $K = 30$.

Table 2 shows quantitative results obtained from 134 test images of the Make3D dataset using various algorithms. The currently proposed method achieves state-of-the-art performance for the correlation coefficient metric (ρ), mean log10 error and root mean squared error (RMS-error). And it achieves comparable performance for the average relative error metric (rel), which measures the average depth error on every single pixel intuitively indicating the estimating precision. Whereas our proposed method is slightly inferior to the DepthTransfer algorithm on rel despite succeeding the identical instance-based learning framework. We suppose DepthTransfer emphasizes estimating and optimizing the depth map

via every single pixel while sacrificing the computational complexity. Comparatively, our method considers the consistency and depth relevance of a group of photometrically similar pixels (superpixels). As regards the metric based on the precision of every pixel, to some extent, we may mistakenly predict some depth information, yet we achieve significantly better overall depth map (see **Fig. 7**). Furthermore, it's noted that ground truth of Make3D dataset has a relatively low resolution due to the limitation of laser range finder, and it may also exert uncertain impact on the final results. The correlation coefficient metric (ρ) measures the correctness of the depth order. Relative depth, is universally recognized as the more appropriate metric for scene understanding. Thus, this proposed method extracts a more suitable depth map for applications of 2D-to-3D conversion.

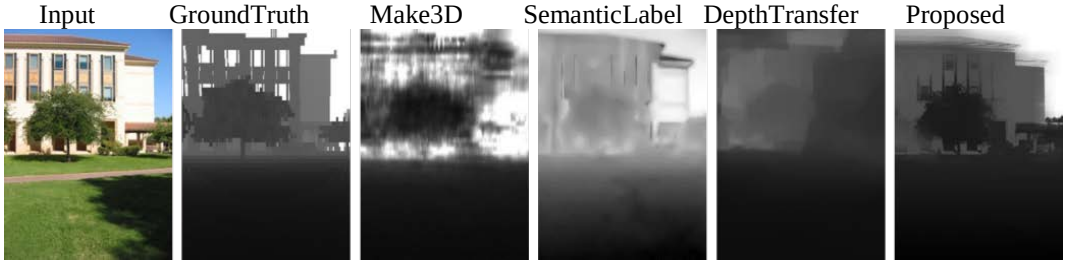
Note that the relative error metric (rel), mean log10 error and root mean squared error (RMS-error) all follow the rule of "lower is better" (i.e., the lower these metrics are, the better the performance becomes), and conversely, the correlation coefficient metric (ρ) follows the rule of "higher is better" as shown in **Table 2.**

Table 2. Comparison of Quantitative Evaluation Results.

Methods	Lower is better			Higher is better
	rel	RMS	log_{10}	ρ
Depth MRF [20]	0.530	16.7	0.198	-
Make3D [19]	0.370	-	0.187	0.65
SemanticLabel [21]	0.375	-	0.148	0.68
DepthTransfer [2, 3]	**0.361**	15.2	0.148	0.71
Proposed	0.367	**15.0**	**0.144**	**0.78**

Several examples of depth maps estimated by diverse approaches are illustrated in **Fig. 7**. Compared to other methods, the proposed method extracted a depth map with superior correctness, richness, and obvious depth level. Final results confirmed higher consistency of depth fields and a better depth edge alignment with ground truth.

Due to the limited precision of the laser range finder that collected the Make3D dataset, the resolution of the ground truth depth map is relatively low. In some cases, the depth map extracted from the proposed algorithm has a richer and more obvious depth level than ground truth one. **Fig. 7(c)** demonstrates, for instance, that there are only three depth levels (ground, nearby tree and sky) in the ground truth depth map. However, in the proposed method's depth map, there are evidently five levels (ground, nearby tree, the distant tree, building and sky). Therefore, the depth map extracted by this method possesses richer layers. And it comprises more abundant scene contents with greater fidelity to human perception of the 3D world.



(a)

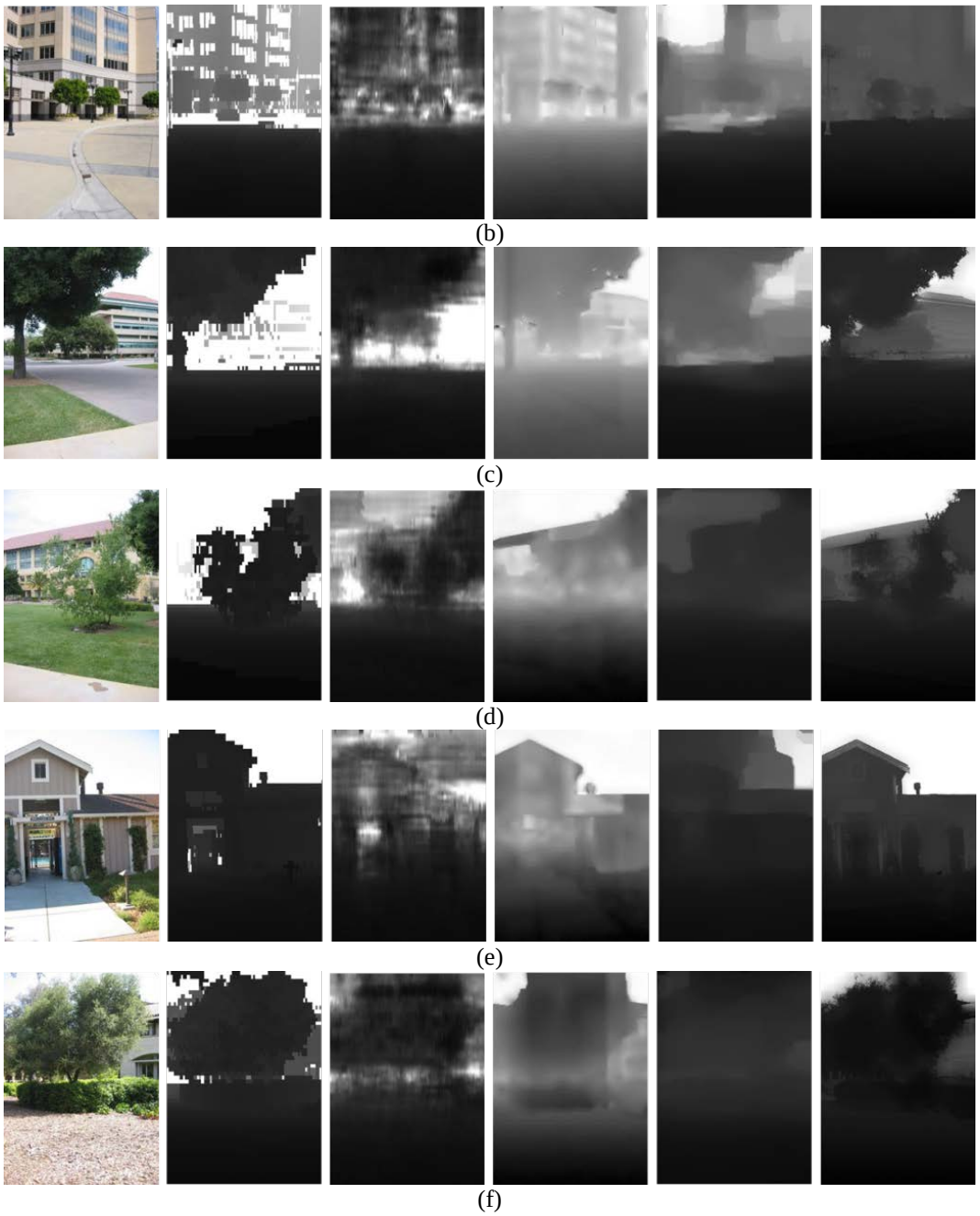


Fig. 7. Single image results obtained from test images in the Make3D dataset. Each result contains the following six images (from left to right): input image, ground truth depth map, depth map produced by Make3D algorithm, depth map produced by SemanticLabel algorithm, depth map produced by DepthTransfer algorithm, depth map produced by the proposed algorithm. The depth maps are show in linear scale. Darker pixels indicate nearby objects (black appears at approximately 1m away) and lighter pixels indicate objects farther away (white appears approximately at 81m away).

For further validation of every improvement included in this paper, comparative experiments were conducted on the Make3D dataset. In leveraging the control variate scheme, one improvement was changed at a time, while leaving the others unaltered. Error reporting was accomplished with the above mentioned three commonly-used metrics, defined in (14), (16), and (17). **Table 3** summarizes the quantitative comparison results as follows: In method 1, the improved fast graph-based segmentation method was changed into the conventional one [30] with no consideration of any corresponding depth information. In method 2, no semantic label was included when performing depth map warping. The cross bilateral filter of method 3 was not selected as the final correction step. As for the effectiveness of regarding the superpixels as the processing unit rather than pixels, it was verified by revision of the algorithm of Depth Fusion [23] in method 4. This is also based on the framework of an instance-based learning method, though the Depth Fusion algorithm is totally based on pixels. Additionally, results of our full model and state-of-the-art DepthTransfer method serving as a contrast are listed.

Table 3. Quantitative Make3D dataset comparison of final results with those obtained 1) without depth information in fast graph-based segmentation, 2) without consideration of semantic label in depth warping, 3) without cross bilateral filter as correction, 4) using a revised Depth Fusion method [23] based on pixels, 5) using our entire model, and 6) using the DepthTransfer method, where the control variate scheme was adopted to quantitatively test every proposed improvement. For clarification, best results are marked in red.

	Methods (Control Variate Scheme)	**Lower is better**		
		rel	log_{10}	RMS
1.	KNN search+FGBS without depth information+Depth warping (superpixels)+CBF	0.420	0.960	16.700
2.	KNN search+IFGBS+Depth warping without semantic label (superpixels)+CBF	0.382	0.161	15.900
3.	KNN search+IFGBS+Depth warping(superpixels) without CBF	0.375	0.155	15.400
4.	Revised Depth Fusion [23]: KNN search+Depth warping(pixels)+CBF	0.377	0.154	15.500
5.	**Full Model**	0.367	**0.144**	**15.000**
6.	DepthTransfer [2,3]	**0.361**	0.148	15.200

- FGBS: Fast Graph-Based Segmentation.
- IFGBS: Improved Fast Graph-Based Segmentation.
- CBF: Cross Bilteral Filter.

Advantages of using the four improvements are shown as three metrics in **Table 3.** They are: 1) improved fast graph-based segmentation, 2) taking semantic information used as a reference, 3) depth correction with CBF, and 4) depth estimation at the superpixel level. Verification of effectiveness was accomplished by a stepwise comparison of each method, 1 through 4, against method 5. Among these four improvements, the improved fast graph-based segmentation is the major contributor, while the other three also have enhanced the algorithm to varying degrees.

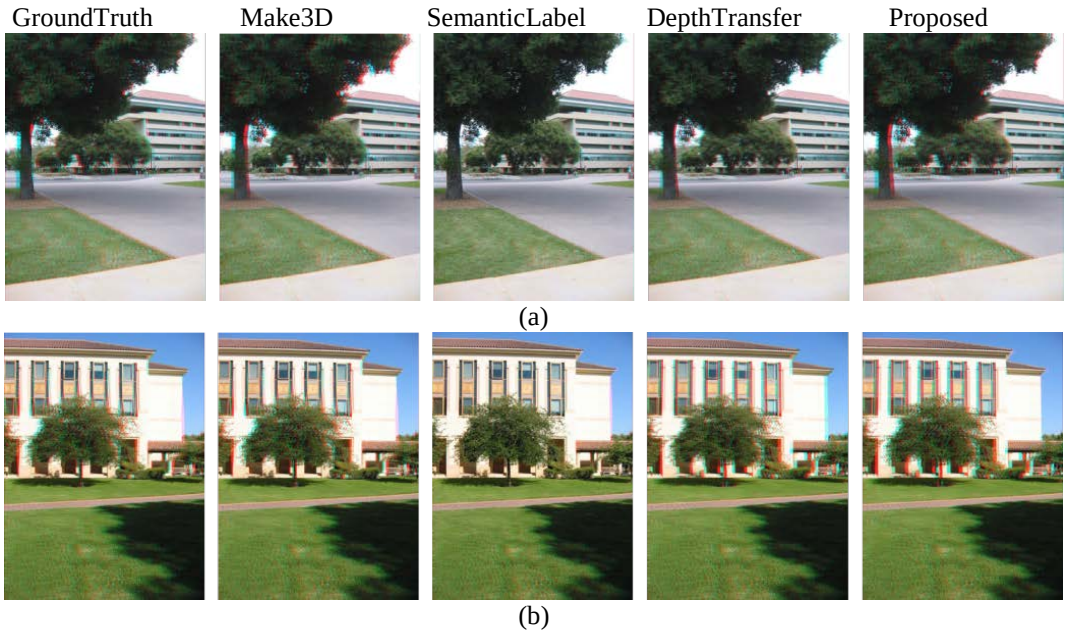


Fig. 8. Anaglyph images generated by employing depth maps of different algorithms. From left to right, respectively: The ground truth, the Make3D algorithm, SemanticLabel algorithm, the DepthTransfer algorithm, and the proposed algorithm. Note that the anaglyph images here should be viewed in color through red-blue anaglyph glasses.

Further implementation of the DIBR (Depth Image Based Rendering) algorithm is done in accordance with [42], and converted the input image into stereo pairs, accomplishing the entire 2D-to-3D conversion. Anaglyph images based on different depth maps extracted by various methods are shown in **Fig. 8**, shows. Among the following conversions, none are flawless. But anaglyph images generated by the proposed algorithm are undeniably more visually pleasing and appealing.

Computational complexity comparison of various methods is illustrated in **Table 4.** The running-time experiment was conducted in an identical hardware and software configuration environment (CPU: Intel Core i7-4790 3.60GHz, RAM: 8 GB; MATLAB version: R2014a). Although Make3D displays optimal efficiency, it is a model-based learning algorithm that requires a prodigious amount of pre-training time, and whenever the dataset is updated or expanded, re-training is always mandated. The proposed method comprises improvements based on the DepthTransfer algorithm. Although image segmentation and semantic labeling are required, an average running time of 1 minute per frame has been attained, a remarkable velocity far superior to that of DepthTransfer. Compared to pixel-based mapping, superpixel-based image mapping exponentially diminishes mapping complexity and computational complexity.

Table 4. Running Time Comparison Results (Minute/Frame).

Method	Running time
Make3D [19]	**0.6**
SemanticLabel [21]	2
DepthTransfer [2, 3]	2.3
Proposed	1.0

5. Conclusion

The key but elusive challenge of 2D-to-3D conversion lies in estimating depth from a single image. Significant improvement of the DepthTransfer algorithm is attained by incorporating superpixels and semantic labels. Tests of the proposed algorithm against state-of-the-art methods demonstrate the superiority in the metrics of the correlation coefficient metric, mean log10 error and root mean squared error and the comparable performance for the average relative error metric. The proposed algorithm performs favorably both in terms of estimated depth quality as well as in computational complexity. Depth maps extracted from the proposed algorithm have more accurate and richer depth levels, and in sum, produce highly compelling 3D images.

Acknowledgments

his work is supported by the Shenzhen Key Lab of Information Theory & Future Network Arch under Grant ZDSYS201603311739428, the Shenzhen Municipal Development and Reform Commission (Disciplinary Development Program for Data Science and Intelligent Computing), and the Shenzhen Engineering Laboratory of Broadband Wireless Network Security.

References

[1] Holliman, N.S., Dodgson, N.A., Favalora, G.E., Pockett, L., "Three-dimensional displays: a review and applications analysis," *Broadcasting, IEEE Transactions on*, vol. 57, pp. 362-371, 2011. Article (CrossRef Link)

[2] Karsch, K., Liu, C., Kang, S.B., "Depth extraction from video using non-parametric sampling," in *Proc. of Computer Vision-ECCV 2012*, pp. 775-788, 2012. Article (CrossRef Link)

[3] Karsch, K., Liu, C., Kang, S.B., "Depth transfer: Depth extraction from video using non-parametric sampling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, pp. 2144-2158, 2014. Article (CrossRef Link)

[4] Chen, J.C., Huang, M., "2D-to-3D conversion system using depth map enhancement," *KSII Transactions on Internet & Information Systems*, vol. 10, 2016. Article (CrossRef Link)

[5] Yang, H., Zhang, H., "Efficient 3d room shape recovery from a single panorama," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5422-5430, 2016. Article (CrossRef Link)

[6] Song, Y., Tang, J., Liu, F., Yan, S., "Body surface context: A new robust feature foraction recognition from depth videos," *IEEE Transactions on Circuits & Systems for Video Technology*, vol. 24, 2014. Article (CrossRef Link)

[7] Xiang, Y., Kim, W., Chen, W., Ji, J., Choy, C., Su, H., Mottaghi, R., Guibas, L., Savarese, S., "ObjectNet3D: A Large Scale Database for 3D Object Recognition," *Springer International Publishing*, 2016. Article (CrossRef Link)

[8] Li, Z., Liu, J., Tang, J., Lu, H., "Robust structured subspace learning for data representation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 37, 2015. Article (CrossRef Link)

[9] Urtasun, R., Lenz, P., Geiger, A., "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 3354-3361, 2012. Article (CrossRef Link)

[10] Liao, M., Gao, J., Yang, R., Gong, M., "Video stereolization: Combining motion analysis with user interaction," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 18, pp. 1079-1088, 2012. Article (CrossRef Link)

[11] Guttmann, M., Wolf, L., Cohen-Or, D., "Semi-automatic stereo extraction from video footage," in *Proc. of Computer Vision, 2009 IEEE 12th International Conference on*, pp. 136-142, 2009. Article (CrossRef Link)
[12] Zhang, R., Tsai, P.S., Cryer, J.E., Shah, M., "Shape-from-shading: a survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, pp. 690-706, 1999. Article (CrossRef Link)
[13] Forsyth, David A., and J. Ponce, "Computer Vision: A Modern Approach, 2/E," *Prentice Hall Professional Technical Reference*, 2002. Article (CrossRef Link)
[14] Subbarao, M., Surya, G., "Depth from defocus: a spatial domain approach," *International Journal of Computer Vision*, vol. 13, pp. 271-294, 1994. Article (CrossRef Link)
[15] Huang, C., Liu, Q., Yu, S., "Regions of interest extraction from color image based on visual saliency," *The Journal of Supercomputing*, vol. 58, pp. 20-33, 2011. Article (CrossRef Link)
[16] Huang, X., Wang, L., Huang, J., Li, D., Zhang, M., "A depth extraction method based on motion and geometry for 2d to 3d conversion," in *Proc. of 2009 Third International Symposium on Intelligent Information Technology Application*, pp. 294-298, 2009. Article (CrossRef Link)
[17] Hoiem, D., Efros, A.A., Hebert, M., "Geometric context from a single image," in *Proc. of Computer Vision, Tenth IEEE International Conference on*, vol. 1, pp. 654-661, 2005. Article (CrossRef Link)
[18] Delage, E., Lee, H., Ng, A.Y., "A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image," in *Proc. of Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 2, pp. 2418-2428, 2006. Article (CrossRef Link)
[19] Saxena, A., Sun, M., Ng, A.Y., "Make3d: Learning 3d scene structure from a single still image," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, pp. 824-840, 2009. Article (CrossRef Link)
[20] Saxena, A., Chung, S.H., Ng, A.Y., "Learning depth from single monocular images," in *Proc. of Advances in Neural Information Processing Systems*, pp. 1161-1168, 2005. Article (CrossRef Link)
[21] Liu, B., Gould, S., Koller, D., "Single image depth estimation from predicted semantic labels," in *Proc. of Computer Vision and Pattern Recognition, IEEE Conference on*, pp. 1253-1260, 2010. Article (CrossRef Link)
[22] Konrad, J., Wang, M., Ishwar, P., Wu, C., Mukherjee, D., "Learning-based, automatic 2d-to-3d image and video conversion," *Image Processing, IEEE Transactions on*, vol. 22, pp. 3485-3496, 2013. Article (CrossRef Link)
[23] Herrera, J.L., Konrad, J., del Bianco, C.R., Garcia, N., "Learning-based depth estimation from 2d images using gist and saliency," in *Proc. of Image Processing, IEEE International Conference on*, pp. 4753-4757, 2015. Article (CrossRef Link)
[24] Konrad, J., Brown, G., Wang, M., Ishwar, P., Wu, C., Mukherjee, D., "Automatic 2d-to-3d image conversion using 3d examples from the internet," in *Proc. of IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics*, pp. 82880F-82880F, 2012. Article (CrossRef Link)
[25] Konrad, J., Wang, M., Ishwar, P., "2d-to-3d image conversion by learning depth from examples," in *Proc. of Computer Vision and Pattern Recognition Workshops, IEEE Computer Society Conference on*, pp. 16-22, 2012. Article (CrossRef Link)
[26] Liu, M., Salzmann, M., He, X., "Discrete-continuous depth estimation from a single image," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 716-723, 2014. Article (CrossRef Link)
[27] Eigen, D., Puhrsch, C., Fergus, R., "Depth map prediction from a single image using a multi-scale deep network," in *Proc. of Advances in neural information processing systems*, pp. 2366-2374, 2014.

[28] Liu, F., Shen, C., Lin, G., "Deep convolutional neural fields for depth estimation from a single image," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5162-5170, 2015. Article (CrossRef Link)
[29] Baig, Mohammad Haris, and L. Torresani, "Coupled depth learning," *Applications of Computer Vision*, pp. 1-10, 2016. Article (CrossRef Link)
[30] Su, C.C., Cormack, L.K., Bovik, A.C., "Depth estimation from monocular color images using natural scene statistics models," in *Proc. of IVMSP Workshop*, pp. 1-4, 2013. Article (CrossRef Link)
[31] Wang, X., Hou, C., Pu, L., Hou, Y., "A depth estimating method from a single image using foe crf," *Multimedia Tools and Applications*, vol. 74, pp. 9491-9506, 2015. Article (CrossRef Link)
[32] Liu, C., Yuen, J., Torralba, A., "Nonparametric scene parsing via label transfer," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, pp. 2368-2382, 2011. Article (CrossRef Link)
[33] Wang, M., Konrad, J., Ishwar, P., Jing, K., Rowley, H., "Image saliency: From intrinsic to extrinsic context," in *Proc. of Computer Vision and Pattern Recognition, IEEE Conference on*, pp. 417-424, 2011. Article (CrossRef Link)
[34] Oliva, A., Torralba, A., "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, pp. 145-175, 2001. Article (CrossRef Link)
[35] Ren, X., Malik, J., "Learning a classification model for segmentation," in *Proc. of Computer Vision, Proceedings, Ninth IEEE International Conference on*, pp. 10-17, 2003. Article (CrossRef Link)
[36] Felzenszwalb, P.F., Huttenlocher, D.P., "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, pp. 167-181, 2004. Article (CrossRef Link)
[37] Malisiewicz, T., Efros, A.A., "Recognition by association via learning per-exemplar distances," in *Proc. of Computer Vision and Pattern Recognition, IEEE Conference on*, pp. 1-8, 2008. Article (CrossRef Link)
[38] Tighe, J., Lazebnik, S., "Superparsing: scalable nonparametric image parsing with superpixels," in *Proc. of Computer Vision-ECCV, Springer*, pp. 352-365, 2010. Article (CrossRef Link)
[39] Durand, F., Dorsey, J., "Fast bilateral filtering for the display of high-dynamic-range images," in *Proc. of ACM transactions on graphics*, vol. 21, pp. 257-266, 2002. Article (CrossRef Link)
[40] Angot, L.J., Huang, W.J., Liu, K.C., "A 2d to 3d video and image conversion technique based on a bilateral filter," in *Proc. of IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics*, pp. 75260D-75260, 2010. Article (CrossRef Link)
[41] Saxena, A., Sun, M., Ng, A.Y., "Learning 3-d scene structure from a single still image," in *Proc. of Computer Vision, IEEE 11th International Conference on*, pp. 1-8, 2007. Article (CrossRef Link)
[42] Zhang, L., Tam, W.J., "Stereoscopic image generation based on depth images for 3d tv," *Broadcasting, IEEE Transactions on*, vol. 51, pp. 191-19, 2005. Article (CrossRef Link)

Yuesheng Zhu received his B.Eng. degree in radio engineering, M. Eng. degree in circuits and systems and Ph.D. degree in electronics engineering in 1982, 1989 and 1996, respectively. He is currently working as a professor at the Lab of Communication and Information Security, Shenzhen Graduate School, Peking University. He is a senior member of IEEE, fellow of China Institute of Electronics, and senior member of China Institute of Communications. His interests include digital signal processing, multimedia technology, communication and information security.

Yifeng Jiang received his B.E. degree from Beijing Institute of Technology (major in information security and countermeasure technique) in 2014, and currently he is a graduate student in Shenzhen Graduate School, Peking University. He has been an exchange student and research assistant in Nanyang Technological University in 2014. His research interests are focused on image segmentation and depth extraction technique.

Zhuandi Huang received her M.S. degree in Computer Application Technology from Shenzhen Graduate School, Peking University in 2014. Her research interests cover depth extaction and 2D-to-3D conversion.

Guibo Luo received his MSc degree in 2013 from Peking University. He is currently working as an engineer at the Lab of Communication and Information Security, Shenzhen Graduate School, Peking University. His research interests are computer vision, machine learning, and multimedia technology.