Segment-Based Disparity Refinement With Occlusion Handling for Stereo Matching

Tingman Yan, Yangzhou Gan, *Member, IEEE*, Zeyang Xia, *Senior Member, IEEE*, and Qunfei Zhao

***Abstract*— In this paper, we propose a disparity refinement method that directly refines the winner-take-all (WTA) disparity map by exploring its statistical significance. According to the primary steps of the segment-based stereo matching, the reference image is over-segmented into superpixels and a disparity plane is fitted for each superpixel by an improved random sample consensus (RANSAC). We design a two-layer optimization to refine the disparity plane. In the global optimization, mean disparities of superpixels are estimated by Markov random field (MRF) inference, and then, a 3D neighborhood system is derived from the mean disparities for occlusion handling. In the local optimization, a probability model exploiting Bayesian inference and Bayesian prediction is adopted and achieves second-order smoothness implicitly among 3D neighbors. The two-layer optimization is a pure disparity refinement method because no correlation information between stereo image pairs is demanded during the refinement. Experimental results on the Middlebury and KITTI datasets demonstrate that the proposed method can perform accurate stereo matching with a faster speed and handle the occlusion effectively. It can be indicated that the "matching cost computation + disparity refinement" framework is a possible solution to produce accurate disparity map at low computational cost.**

***Index Terms*— Stereo vision, disparity refinement, Markov random fields, RANSAC, Bayesian inference.**

I. Introduction

STEREO matching is a key step in 3D reconstruction. It takes two rectified color images as input and matches object projections in the image domain to compute disparities. Depth can be directly reconstructed via disparity and camera parameters. The foreground-background occlusion which is almost inevitable makes the matching difficult since the occluded regions are only visible in one view. Matching is also ambiguous in scenes with low or repetitive textures. Other challenges include imperfect rectification and radiometric differences. The Middlebury 2014 benchmark [1] provides an evaluation that contains all these challenges, researchers can upload their results for fair comparison. Besides accuracy, fast computation is also required for real-time applications.

Stereo matching methods usually have (subsets of) four steps [2]: matching cost computation, cost aggregation, disparity computation/optimization, and disparity refinement. Matching cost measures the pixel-wise or patch-wise similarity between image locations. Common methods include absolute differences (AD), sampling insensitive measure (BT) [3], normalized cross-correlation (NCC), census and rank transforms [4], and the combination of these methods like AD-Census [5]. Recently, the powerful convolutional neural networks (CNN) has been applied to matching cost computation. Žbontar and LeCun [6] developed the MC-CNN method which learns the similarity measure between image patches. Fast network architecture which is able to produce an accurate result within one second was proposed by Luo *et al.* [7].

Cost aggregation and disparity computation/optimization are two key steps that determine the accuracy of stereo methods. Local stereo methods perform averaging or weighted averaging of matching costs [8] in a fixed size window and disparities are computed by the winner-take-all (WTA) operation to the cost volume. Yang [9] proposed a non-local cost aggregation on the minimum spanning tree (MST) structure. This idea was extended to 3D non-local cost aggregation on a 3D-multiple-MST structure [10]. Global stereo methods usually omit the cost aggregation step. Instead, a global energy function which penalizes depth discontinuities is optimized on the cost volume to compute disparity. Although the global method is much more accurate than the local one, it is far more computational complicated.

Disparity refinement is designed to further improve the results in hard regions, such as occluded regions and low texture regions. Most refinement methods follow the detection and filling scheme, followed by a filtering step. Left-right consistency check (LRC) [11] is commonly used to detect outliers. Jang and Ho [12] proposed an energy function to detect occlusion and classified the occlusions into leftmost occlusions and inner occlusions. Banno and Ikeuchi [13] labeled pixels that failed the LRC as low confidence and introduced a directed anisotropic diffusion to refine these pixels. Huang and Zhang [14] proposed a fast refinement including belief aggregation for outlier detection and belief propagation for filling. In the work of Mei *et al.* [5], outliers

Manuscript received March 17, 2018; revised October 1, 2018 and January 25, 2019; accepted February 22, 2019. Date of publication March 6, 2019; date of current version June 20, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61773365 and in part by the Major Project of the Guangdong Province Science and Technology Department under Grant 2014B090919002. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gene Cheung. *(Corresponding authors: Zeyang Xia; Qunfei Zhao.)*

T. Yan and Q. Zhao are with the Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: meander@sjtu.edu.cn; zhaoqf@sjtu.edu.cn).

Y. Gan and Z. Xia are with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China, and also with the CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Shenzhen 518055, China (e-mail: yz.gan@siat.ac.cn; zy.xia@siat.ac.cn).

Digital Object Identifier 10.1109/TIP.2019.2903318

were detected and classified into occlusions and mismatches and then an iterative region voting was applied to interpolate these outliers accordingly. Filtering like bilateral filtering and weighted median filtering [15] were also employed for disparity refinement. It is shown [16] that multi-step and iterative refinement strategies can result in competitive results. However, in large occluded regions, inner occlusions cannot be directly refined by these strategies and cumulative error may be introduced.

In the Middlebury 2014 benchmark [1], top-rank methods have achieved high accuracy in non-occluded regions. However, the evaluation error that contain occluded regions are almost doubled for most error metrics. Therefore, accurate estimation near occluded regions is still a challenging problem. In addition, all of the top-ten methods run more than 120s, which makes them hard to be applied in computationally intensive applications. To tackle these problems, we developed a stereo matching method that has a higher accuracy and lower computational cost with occlusion handling.

The proposed method directly refines the WTA disparity map computed from the raw matching cost with the guidance of the color reference image. The reference image is segmented into superpixels and the proposed method operates on superpixel-level, which is of high efficiency. First, a front-parallel disparity map is obtained by estimating mean disparities of superpixels. Then a slanted-surfaces disparity map is refined by assigning each superpixel a plane. The front-parallel to slanted-surfaces framework is achieved by a two-layer optimization. In the global optimization layer, the front-parallel disparity map is estimated by MRF optimization. In the local optimization layer, the slanted-surfaces disparity map is refined by the RANSAC plane fitting and the probability-based disparity plane refinement. The two layers are connected by two constraints: the slanted-surfaces disparity map cannot deviate far from the front-parallel disparity map; the two disparity map share the same depth discontinuities. The first constraint helps remove outliers and deal with degeneracy in the RANSAC plane fitting. The second constraint is embedded in a 3D neighborhood system and contributes for occlusion handling. The proposed method is evaluated on the Middlebury 2014 and the KITTI 2015 dataset and compared with the state-of-the-art disparity refinement methods. Experimental results demonstrate its accuracy, efficiency, and robustness.

In summary, the main contributions of this paper are: (1) A *pure* disparity refinement method that directly refines the WTA disparity map with the guidance of the color reference image and achieves the state-of-the-art performance with occlusion handling. (2) A 1D label MRF formulation with a novel data term that is based on disparity distributions. And a theoretical analysis that proves the 1D label MRF cannot model the highly slanted surfaces. (3) A front-parallel to slanted-surfaces framework with a Bayesian inference and Bayesian prediction based disparity plane refinement that makes the 1D label approach robust to slanted surfaces.

II. Related Work

This section mainly focuses on MRF stereo methods and segment-based stereo methods which are more related to our work. We refer readers to [2] and [17] for more comprehensive reviews.

A. MRF Stereo Methods

Markov Random Fields (MRF) stereo methods formalize stereo matching as a label problem and the goal is to optimize a global energy function which measures the quality of the labeling.

Conventional MRF stereo methods [18], [19] assign each pixel a 1D discrete disparity label. Optimizations such as graph cuts [20], [21], belief propagation [22], [23] and TRW [24] can be used to minimize the energy function. Graph cuts based expansion moves and swap moves [18] are shown to have good performance. These moves can update labels of all pixels simultaneously and therefore the optimization is hard to be trapped by the local minima. The drawback of 1D label stereo methods is modeling the highly slanted surfaces. 3D label stereo methods [25]–[27] are proposed to model the scene more accurately. These methods can not only model highly slanted surface but also achieve second order smoothness constraints [25], [28], [29]. Therefore, they usually have better accuracy than 1D label methods. However, the global optimization on pixel level complicates the computation.

Our method performs 1D label MRF inference on superpixel level. Since the number of superpixels is much less than that of pixels in an image, the inference is much faster. Unlike previous work, ours takes the discrete mean disparity of superpixel as the label. Even in highly slanted surfaces, the mean disparities can also be correctly estimated.

B. Segment-Based Stereo Methods

Segment-based stereo methods [30]–[32] assume the scene structure to be piece-wise planar and the estimation of disparity map transforms into assigning a 3D disparity plane to each segment. First, these methods segment the reference image into regions with homogeneous color. Then an initial disparity map is computed by a known stereo matching method and candidate disparity planes are generated by plane fitting to the disparity map. Finally, a global optimization, e.g., graph cuts and belief propagation, is utilized to assign each segment an optimal plane label. The final results rely on the quality of the segmentation. To relax the segment constraints, several improvements have been proposed. Over-segmentation is a common solution to ensure that depth discontinuities only occur in the boundary of segments. Bleyer *et al.* [33] proposed a pixel-wise MRF formulation that incorporated soft segment constraints. Joint segmentation and disparity computation [34], [35] can improve the segmentation quality during optimization. But these methods have a common drawback. The final plane label of a segment is assigned from the candidate label set. The finite set may not contain the correct label of the segment, in such case the estimation of the disparity plane is false and the error cannot be corrected. In the work of Wang and Zheng [36], the total energy function is optimized by cooperative optimization and is decomposed into the sum of sub-target energy functionals which are locally optimized. The optimization process is done iteratively and

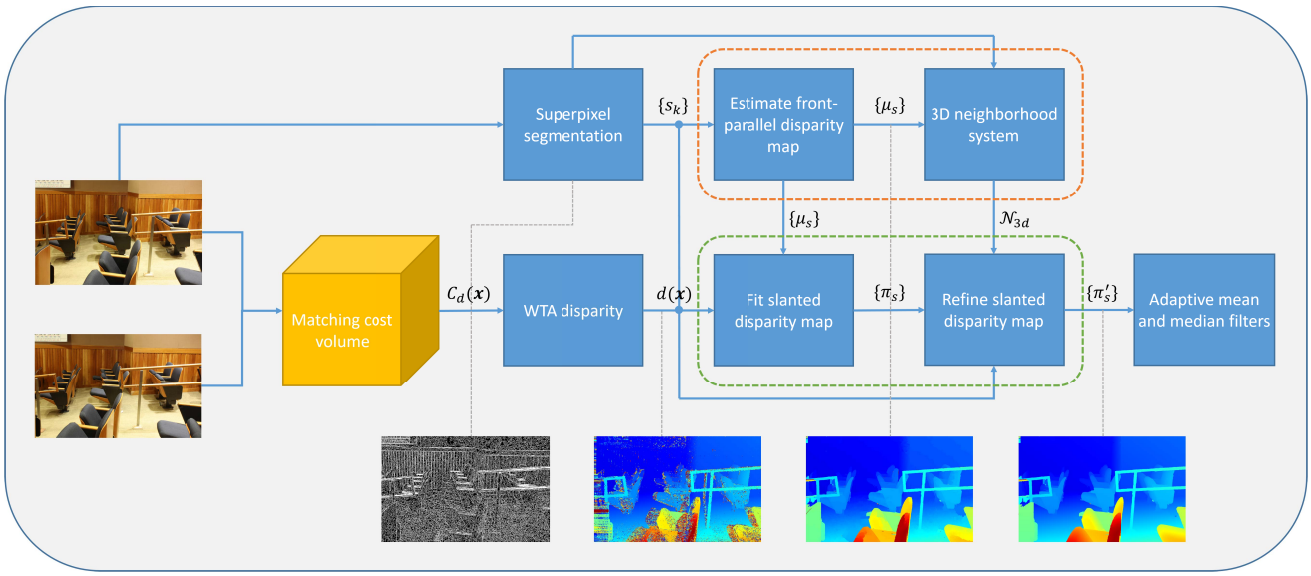


Fig. 1. Workflow of the proposed method. The WTA disparity map is refined by a two-layer optimization. The global and local optimization layers are in the orange and green rounded rectangles, respectively. Four intermediate results are shown in the bottom.

false plane label can be corrected by the local optimization. Therefore, their method is robust to the initial plane-fitting result.

In contrast to previous works which optimize a global energy function, our method refines the disparity plane by a local optimization and constraints smoothness implicitly. Moreover, the disparity refinement method demands no correlation information between the stereo image pairs and thus can be processed on a single view.

III. Overview of the Two Layer Optimization

The matching cost volume $C_d(\boldsymbol{x})$ is generated by MC-CNN[1] [6]. The disparity map $d(\boldsymbol{x})$ is computed by winner-take-all:

$$d(\boldsymbol{x}) = \arg\min_{d^*} C_{d^*}(\boldsymbol{x}) \tag{1}$$

The left reference image is segmented into superpixels $\{s_k\}$ by the graph-based segmentation [37]. The workflow of our method is shown in Fig. 1.

We propose a two-layer optimization to refine the WTA disparity map. In the global optimization layer (Section IV), a front-parallel disparity map is estimated by MRF optimization. The 3D neighborhood system $\mathcal{N}_{3d}$ is derived from superpixels mean disparities $\{\mu_s\}$. In the local optimization layer (Section V), slanted planes $\{\boldsymbol{\pi}_s\}$ are fitted for superpixels by RANSAC and mean disparities of superpixels $\{\mu_s\}$ are utilized to constraint the fitting. The initial slanted disparity map is refined by a probabilistic model that exploits Bayesian inference and Bayesian prediction in the 3D neighborhood system. Both optimization layers operate at superpixel level and have high efficiency.

[1]Downloaded from https://github.com/t-taniai/LocalExpStereo

IV. Front-Parallel Disparity Map

We use the global MRF optimization to estimate a front-parallel disparity map. Superpixels are formulated as graph nodes. MRF optimization aims to minimize the following energy:

$$E(\mu) = \sum_{s\in\Omega} \phi_s(\mu_s) + \lambda \sum_{(s,t)\in\mathcal{N}} \psi_{st}(\mu_s, \mu_t), \tag{2}$$

where μ_s is the label, in our case it is the mean disparity of superpixel s; Ω is the set of superpixels, $\Omega = \{s_k\}$, and $\mathcal{N}$ represent the set of neighboring superpixels; and $\phi_s(\mu_s)$ is called the *data term*, $\psi_{st}(\mu_s, \mu_t)$ is called the *smoothness term* and λ is a parameter to balance the influence of the smoothness term. In contrast to 3D label MRF, optimizing 1D label on superpixel level is efficient (Section IV-B).

We propose a novel data term which is based on disparity distribution (Section IV-A) instead of matching cost or similarity measure between left and right images. To handle the foreground-background occlusions, the 3D neighborhood system which represents depth discontinuities is derived by $\{\mu_s\}$ (Section IV-C). We also study a special case and prove that the 1D label MRF formulation cannot model the highly slanted surfaces (Section IV-D).

A. Disparity Distribution Interpretation

Segment-based stereo methods assume that disparities are approximately linear within a segmentation. With the piece-wise planar surfaces assumption, the disparity distribution of a planar surface with appropriate boundaries shall be evenly distributed. Considering the irregular boundary shape of superpixels, we model the disparity distribution within a superpixel s a normal distribution

$$\mathrm{Norm}_d(\mu_s, \sigma_s) = \frac{1}{\sqrt{2\pi}\sigma_s} \exp\left(-\frac{(d-\mu_s)^2}{2\sigma_s^2}\right), \tag{3}$$

where d represents the disparity, μ_s and σ_s are disparity mean and variance of superpixel s, respectively. Higher σ_s indicates a more slanted surface while for a front-parallel surface, σ_s is approximately equal to zero. The data term of (2) is based on disparity distribution histograms, as described in Section IV-B.

B. MRF Optimization

To estimate a front-parallel disparity map, we estimate mean disparities of superpixels. The front-parallel plane of superpixel s can be obtained by $\boldsymbol{\pi}_s^{fp} = (0, 0, \mu_s)$. The data term and smoothness term of (2) are defined as follows:

1) Data Term: To measure the confidence of disparity centers, the disparity distributions of superpixels are divided into histogram bins. We count the number that the WTA disparity $d_s(\boldsymbol{x})$ in superpixel s falls into a bin $\mathcal{B}(\mu_s)$ with bin-width L. The data term of s is defined as

$$\phi_s(\mu_s) = N_s - \sum_{i=1}^{N_s} \mathcal{I}(d_s(\boldsymbol{x}_i) \in \mathcal{B}(\mu_s)), \tag{4}$$

where N_s is the number of pixels in superpixel s, μ_s takes discrete values, $\mu_s = 0, L, 2L, \cdots$, and lower data term implies higher confidence due to the negative sign. $\mathcal{I}$ is a function of condition, defined as

$$\mathcal{I}(\cdot) = \begin{cases} 1, & \text{if } \cdot \text{ is true} \\ 0, & \text{if } \cdot \text{ is false} \end{cases}, \tag{5}$$

and in (4) $\mathcal{I}$ indicates whether the disparity $d_s(\boldsymbol{x}_i)$ falls into bin $\mathcal{B}(\mu_s)$, i.e. $d_s(\boldsymbol{x}_i) \in [\mu_s, \mu_s + L)$.

The design of data term is voting-based. More observations falling in the same bin results in a higher confidence. The WTA disparities in occluded regions are noise-corrupted and it is hard for them to reach a consensus. Therefore, the data term in occluded regions is relatively high and the label is dominated by the smoothness term.

2) Smoothness Term: The smoothness term enforces the similarity of disparity distribution centers among neighboring superpixels, which is defined as

$$\psi_{st}(\mu_s, \mu_t) = \max(\omega_{st}, \epsilon)\mathcal{L}(s, t)T(\mu_s, \mu_t), \tag{6}$$

where ω_{st} is a color-similarity weight which is defined as

$$\omega_{st} = e^{-\|I(s)-I(t)\|_2/\gamma}, \tag{7}$$

where γ is a parameter that controls the influence of color weight, and $I(s)$ denotes the average color of superpixel s; ϵ is a lower-bound truncated value [29]; $\mathcal{L}(s, t)$ [38] is the shared boundary length between neighboring superpixels s and t; and T could be a metric or a semi-metric which will be defined in Section IV-C.

C. 3D Neighborhood System

The superpixel mean disparities $\{\mu_s\}$ estimated by MRF optimization provide global scene information. The 3D neighborhood system which represents depth discontinuities is inferred from $\{\mu_s\}$ as

$$\mathcal{N}_{3d} = \{(s, t) \in \mathcal{N} \mid |\mu_s - \mu_t| \leq L\}. \tag{8}$$

(a) (b) (c)

Fig. 2. (a) Is the disparity map of a slanted plan is the disparity map of a slanted plan is the disparity map of a slanted plane $\boldsymbol{\pi}$, in which v_d is the rate of disparity changes along the u axis. (b) Is the disparity map after MRF optimization of the Jadeplant image pair and $v_d \approx 0.96$ in the red rectangle. (c) Playtable image pair with $v_d \approx 0.14$ in the red rectangle.

For a pair of superpixels $(s, t) \in \mathcal{N}$, if their mean disparities are not similar, i.e. $|\mu_s - \mu_t| > L$, then the pair $(s, t) \notin \mathcal{N}_{3d}$. Therefore the possible foreground-background connections are cut in $\mathcal{N}_{3d}$. This is an important property for occlusion handling. See Section VI-E for more discussion.

To estimate $\mathcal{N}_{3d}$ accurately, T in the smoothness term (6) shall be properly designed. By the definition of $\mathcal{N}_{3d}$, T should be large if $|\mu_s - \mu_t| > L$ and be small if $|\mu_s - \mu_t| \leq L$. Thus, T is set to

$$T_s(\mu_s, \mu_t) = \begin{cases} 0, & |\mu_s - \mu_t| = 0 \\ 1, & |\mu_s - \mu_t| = L \\ \tau, & |\mu_s - \mu_t| > L \end{cases}, \tag{9}$$

where $\tau > 2$. T_s is a semi-metric which is non-submodular and cannot be optimized by the popular expansion moves [18].

To optimize the semi-metric, we follow the first expansion then swap moves in [18]. First, we set T to

$$T_l(\mu_s, \mu_t) = |\mu_s - \mu_t| \tag{10}$$

which is a linear metric. The labels $\{\mu_s^l\}$ are obtained by the expansion moves. Then, we set T to T_s and perform the swap moves with $\{\mu_s^l\}$ as the initial labels to get the optimal labels $\{\mu_s\}$. This two-steps approximation could be simplified by using expansion moves with QPBO [39] or more advanced techniques [40]–[42] for dealing with non-submodular energies.

D. Estimate Front-Parallel Plane in Slanted Surfaces

We analyze the model's performance on slanted surfaces. Fig. 2a shows the disparity map of a slanted plane $\boldsymbol{\pi}$, in which axes u and v are aligned such that disparities are constant along the v axis. The rate of disparity changes along u axis is v_d. To simplify, here are two assumptions: the surface of $\boldsymbol{\pi}$ is segmented into $m \times n$ superpixels, and each superpixel is a square with side length a; the surface of $\boldsymbol{\pi}$ has homogeneous colors. We will prove that even if the WTA disparity map of plane $\boldsymbol{\pi}$ is the same as the ground truth, the MRF formulation fails to model the highly-slanted surface.

Two set of candidate labels, the slanted ground truth labels $\{\mu_{slanted}\}$ and the front-parallel labels $\{\mu_{front}\}$ which assign the same label to all superpixels are considered here. When $v_d a \geq L$ and $T = T_l$, from (4) and (6), the energy of $\{\mu_{slanted}\}$ is

$$E(\mu_{slanted}) = mn\left(a^2 - \frac{aL}{v_d}\right) + (m-1)n\lambda a^2 v_d, \tag{11}$$

and the energy of $\{\mu_{front}\}$ is

$$E(\mu_{front}) = mna^2 - n\frac{aL}{v_d}. \tag{12}$$

If $v_d \geq \sqrt{L/(\lambda a)}$, then $E(\mu_{front}) \leq E(\mu_{slanted})$.

Therefore, if the plane is slanted enough, i.e. $v_d > v_d^l = \max\left(L/a, \sqrt{L/(\lambda a)}\right)$, the energy of the ground truth label is not the minimum as the model desires. For the semi-metric T_s, the same conclusion holds if $v_d > v_d^s = \max\left(L/a, L/(\lambda\tau)\right)$.

Experimental results verify the conclusion. As described in Section VI-A, $L = 2$, $\lambda = 0.3$, $\tau = 16$ and $a \approx \sqrt{\bar{N}} = 8.5$, where $\bar{N}$ is the average number of pixels in superpixels. Hence $v_d^l = 0.89, v_d^s = 0.42$. In Fig. 2b and Fig. 2c, the superpixels in the red box have the front-parallel labels and the slanted labels, respectively. This agrees with our conclusion, since the v_d in Fig. 2b is larger than both v_d^l and v_d^s and the v_d in Fig. 2c is smaller than both v_d^l and v_d^s.

If a set of neighboring superpixels have the same label, it is possible that these superpixels are in the same slanted surface. Therefore, we merge them into a new superpixel to benefit the following local optimization layer.

V. Slanted-Surfaces Disparity Map

To handle the slanted surfaces, we propose a local optimization which is guided by the global optimization results, i.e. the front-parallel disparity map. Two constraints are introduced. First, the slanted surfaces disparity map cannot deviate far from the front-parallel disparity map. Second, the two disparity maps share the same depth discontinuities. The two constraints are employed in the plane fitting (Section V-A) and the disparity plane refinement (Section V-B) procedures, respectively.

A. RANSAC Plane Fitting

A slanted plane $\boldsymbol{\pi}_s = (a_s, b_s, c_s)$ is fitted for each superpixel s by the observations from the WTA disparity map $d(\boldsymbol{x})$. $\{\mu_s\}$ is utilized to constrain the fitting in two aspects: selecting reliable observations and dealing with degeneracy.

1) Reliable Observation Selection: As described before, we assume the disparity distribution of a superpixel as a normal distribution. Therefore, the effective distribution shall be unimodal and have continuous disparity domain. Here, we use density distribution instead of disparity distribution, for robustness.

The density of disparity d in a superpixel s is defined as

$$\rho_s(d) = \sum_{\boldsymbol{x}\in s} \mathcal{I}\left(|d - d(\boldsymbol{x})| \leq L\right). \tag{13}$$

To select effective densities, the density set $P_s = \{\rho_s(d), d = 1, \ldots, D_{max}\}$ are divided into two subsets: P_s^+ with densities larger than $\bar{\rho}_s$ and $P_s^- = P_s \setminus P_s^+$, $\bar{\rho}_s$ is the average density in the disparity range D_{max}. Hence, P_s^+ is the effective subset of P_s, which consists of several subsets with continuous disparity domain. We utilize the prior mean disparity μ_s to select the only subset of P_s^+ whose disparity domain contains μ_s. The observations with its disparities in the subset domain are reserved. If no such subset exists, i.e. $\rho_s(\mu_s) \leq \bar{\rho}_s$, there are no effective observations and the fitting is regarded as failed.

The test threshold of RANSAC is determined as

$$\theta_s = \begin{cases} 1, & D_s \leq D_{slanted} \\ L, & D_s > D_{slanted} \end{cases}, \tag{14}$$

where D_s is the disparity range of the selected subset of P^+ and $D_{slanted}$ is a cutoff range. The slanted surface has large D_s and thus the test threshold is increased to L to improve the robustness to test error.

2) Degeneracy: In RANSAC plane fitting, the sampling strategy of GroupSAC [43] is adopted. The groups are obtained by clustering the observations in the superpixel according to their disparities. The density-based clustering [44] is used because of its efficiency and ability to determine the number of clusters automatically.

The RANSAC [45] sets no constraints to the fitting result, which may lead to the problem called degeneracy [46]. To handle the problem, the fitting result is constrained by μ_s as

$$\left|\hat{\mu}_s - \tilde{\mu}_{s,k}\right| \leq L \quad and \quad \left|\hat{\sigma}_s^2 - \tilde{\sigma}_{s,k}^2\right| \leq L^2, \tag{15}$$

where $\tilde{\mu}_{s,k}$ and $\tilde{\sigma}_{s,k}^2$ are the disparity mean and variance generated by the fitting result $\boldsymbol{\pi}_s^k$, respectively; k is the number of samples drawn so far; and $\hat{\mu}_s$ and $\hat{\sigma}_s^2$ are estimated via maximum a posteriori (MAP) inference with normal inverse gamma prior which is conjugated to the normal distribution [47]. The hyperparameters of the prior distribution are set as

$$[\alpha, \beta, \gamma, \delta] = [1, 1, N_s^-, \mu_s], \tag{16}$$

where N_s^- is the number of reliable observations in s. During the sampling, if (15) is not satisfied, $\boldsymbol{\pi}_s^k$ will be rejected.

The termination condition of RANSAC [45] is that the probability of no model having higher consistency is smaller than a predefined threshold η

$$(1 - \varepsilon^m)^k \leq \eta, \tag{17}$$

where ε is the inlier ratio tested on all reliable observations and m is the minimum number of data points needed to generate a plane model, $m = 3$. If the constraint in (17) is not satisfied after all samples are drawn, or the inlier ratio is less than a threshold ε_P, the fitting result will be regarded as failed.

B. Disparity Plane Refinement

The initial planes $\{\boldsymbol{\pi}_s\}$ are fitted independently for each superpixel, of which the results in textureless and occluded regions are often of poor quality. We refine $\boldsymbol{\pi}_s$ by the prior information of the local neighbors of superpixel s in a probabilistic manner. The refinement is performed on the 3D neighborhood system which contributed to occlusion handling.

The likelihood of the observations $\boldsymbol{p}_{1,\ldots,N_s} = \left\{\boldsymbol{p}_i \mid \boldsymbol{p}_i = (u_i, v_i, d_i), i = 1, \ldots, N_s\right\}$ in superpixel s given the plane label $\boldsymbol{\pi}_t = (a_t, b_t, c_t)$ is defined as

$$Pr(\boldsymbol{p}_{1,\ldots,N_s}|\boldsymbol{\pi}_t) = \frac{1}{N_s}\sum_{i=1}^{N_s} \mathcal{I}(|a_t u_i + b_t v_i + c_t - d_i| \leq 1) \tag{18}$$

which is the inlier ratio of plane $\boldsymbol{\pi}_t$ tested on the observations of superpixel s.

The prior of a neighbor superpixel t to s is defined as

$$Pr(\boldsymbol{\pi}_t) \propto \max(\omega_{st}, \epsilon) Pr(\boldsymbol{p}_{1,\ldots,N_t}|\boldsymbol{\pi}_t), \tag{19}$$

where ω_{st} denotes the color similarity which is same as (7), and $Pr(\boldsymbol{p}_{1,\ldots,N_t}|\boldsymbol{\pi}_t)$ is the inlier ratio of $\boldsymbol{\pi}_t$ tested on the observations in superpixel t. Therefore, the neighbor with similar color and good fitting result serves as a strong prior to s.

Our local optimization only considers the prior knowledge from 3D neighbors of superpixel s. As a result, for a superpixel $t \in \mathcal{N}_{3d}(s)$, the posterior probability of $\boldsymbol{\pi}_t$ is

$$Pr(\boldsymbol{\pi}_t|\boldsymbol{p}_{1,\ldots,N_s}) = \frac{Pr(\boldsymbol{p}_{1,\ldots,N_s}|\boldsymbol{\pi}_t) Pr(\boldsymbol{\pi}_t)}{\sum\limits_{t' \in \mathcal{N}_{3d}(s)} Pr(\boldsymbol{p}_{1,\ldots,N_s}|\boldsymbol{\pi}_{t'}) Pr(\boldsymbol{\pi}_{t'})}, \tag{20}$$

The posterior predictive probability of an arbitrary data point $\boldsymbol{p}_*$ is estimated by Bayesian prediction

$$Pr(\boldsymbol{p}_*|\boldsymbol{p}_{1,\ldots,N_s}) = \sum_{t \in \mathcal{N}_{3d}(s)} Pr(\boldsymbol{p}_*|\boldsymbol{\pi}_t) Pr(\boldsymbol{\pi}_t|\boldsymbol{p}_{1,\ldots,N_s}), \tag{21}$$

where $Pr(\boldsymbol{p}_*|\boldsymbol{\pi}_t)$ is the binary inlier ratio of $\boldsymbol{\pi}_t$ tested on $\boldsymbol{p}_*$, and the sum is a weighted average of $Pr(\boldsymbol{p}_*|\boldsymbol{\pi}_t)$ with the posterior $Pr(\boldsymbol{\pi}_t|\boldsymbol{p}_{1,\ldots,N_s})$ as the weight.

To generate the refined labels, we take a sampling and fitting strategy and draw samples from s's neighbors to enforce the smoothness among neighbors. Concretely, we draw sample points $\boldsymbol{p}_i^b = (u_i^b, v_i^b, d_i^b)$ in boundaries of superpixel $t \in \mathcal{N}_{3d}(s)$ by $\boldsymbol{\pi}_t = (a_t, b_t, c_t)$ and

$$d_i^b = a_t u_i^b + b_t v_i^b + c_t, \ i = 1, \ldots, N_t^b, \tag{22}$$

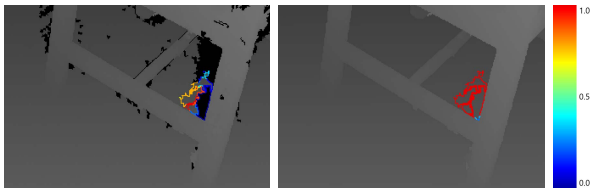
where N_t^b is the number of boundary pixels and (u_i^b, v_i^b) locates in the boundary of superpixel t. To generate the final solution, a plane is fitted by the sample points $\left\{\boldsymbol{p}^b_{1,\ldots,N_t^b}, t \in \mathcal{N}_{3d}(s)\right\}$ using weighted least squares (WLS). The weight of $\boldsymbol{p}_i^b$ is the posterior predictive probability $Pr(\boldsymbol{p}_i^b|\boldsymbol{p}_{1,\ldots,N_s})$.

When labels of all s's neighbors are the same, i.e. $\forall t \in \mathcal{N}_{3d}(s)$, $f_t = f_s$, we have $Pr(\boldsymbol{p}_i^b|\boldsymbol{\pi}_t) = 1$ because all the samples are generated by the same label. From (20) and (21) we know

$$Pr(\boldsymbol{p}_i^b|\boldsymbol{p}_{1,\ldots,N_s}) = \sum_{t \in \mathcal{N}_{3d}(s)} Pr(\boldsymbol{\pi}_t|\boldsymbol{p}_{1,\ldots,N_s}) = 1. \tag{23}$$

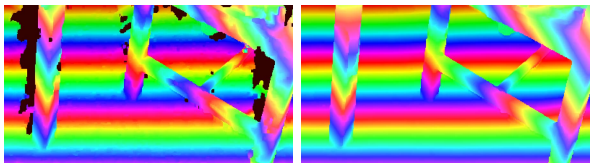
In such case, all the sample points reach their maximum weight.

According to (21), the samples in the shared plane of neighboring superpixels have higher posterior predictive probability and thus have higher weights in the WLS fitting. Therefore, our method encourages the consistency of neighbors labels and achieves implicit second order smoothness constraints among neighboring superpixels. The visualization of weights of sample points before refinement is shown in Fig. 3a. The posterior predictive probabilities of sample points after the refinement are closed to 1 as shown in Fig. 3b, which demonstrates the effectiveness of the method. The disparity plane refinement is iterated twice since the WLS weights of the second iteration is much larger than that of the first iteration. Iterations more than twice have no further improvements since the weights of the new sampled points are already close to the maximum.



(a) (b)

Fig. 3. The visualization of sample points and their corresponding weights (posterior predictive probabilities) for WLS fitting of an occluded region. (a) Shows the weights before disparity plane refinement. (b) The weights after refinement are close to 1, this demonstrates the effectiveness of the refinement.



(a) (b)

Fig. 4. The comparison of results (a) before and (b) after refinement. The method achieves appropriate interpolation of occluded regions and second order smoothness among 3D neighboring superpixels.

For a failed RANSAC fitting result $\boldsymbol{\pi}_s$ of superpixel s, the likelihood $Pr(\boldsymbol{p}_{1,\ldots,N_s}|\boldsymbol{\pi}_s)$ is equal to 0. From (19) we know the prior $Pr(\boldsymbol{\pi}_s)$ also equal to 0 and hence the failed result $\boldsymbol{\pi}_s$ has no contribution to the WLS weights. The final result $\boldsymbol{\pi}'_s$ of superpixel s is determined by its neighbors that pass the RANSAC fitting. Therefore, we achieve the interpolation of occluded regions via the Bayesian inference. Moreover, to make sure failed superpixels have neighbors that pass the fitting, we merge the 3D neighboring superpixels that failed the RANSAC fitting into a new superpixel. The merged superpixel has a larger boundary and therefore provides more information for interpolation. The comparison of disparity results before/after refinement with rainbow mapping is shown in Fig. 4.

As a pure disparity refinement method, our method can be processed on a single view. We did not utilize the commonly used left-right-consistency (LRC) check because it would double the computational cost. At last, the slanted-surfaces disparity map is further refined by the adaptive mean and median filters.

VI. Experimental Results

A. Experiment Setup

1) Testing Database: Our method is mainly tested on the Middlebury dataset and compared with other state-of-the-art disparity refinement methods. The Middlebury dataset consists of 15 test image pairs and 15 training image pairs with ground truth and includes various static indoor scenes. The resolution of the full resolution dataset is about 3000×2000 and most submitted methods use the half resolution dataset for evaluation. We also follow this setting. The high resolution,

TABLE I

ERROR RATES ON FIVE ERROR METRICS UNDER TWO MASKS OF THE MIDDLEBURY 2014 BENCHMARK. STATE-OF-THE-ART DISPARITY REFINEMENT METHODS SNP-RSM [48], MC-CNN+RBS [49], AND MC-CNN-ACRT [6] ARE LISTED FOR COMPARISON

Metrics	bad 1.0		bad 2.0		bad 4.0		avgerr		rms	
Masks	noc	all	noc	all	noc	all	noc	all	noc	all
Middlebury 2014 Test Set										
SDR	18.8	**25.1**	**7.69**	**13.8**	**4.90**	**10**	2.94	**6.16**	15.4	**24.2**
SNP-RSM	18.0	26.3	8.08	16.6	4.95	11.4	2.73	7.63	15.6	30.4
MC-CNN+RBS	18.1	27.5	8.42	18.1	5.08	13.9	**2.67**	8.19	**15.0**	29.9
MC-CNN-acrt	**17.1**	27.3	8.75	19.1	4.91	15.8	3.82	17.9	21.3	55.0
Middlebury 2014 Training set										
SDR	21.1	**26.9**	10.1	**15.7**	**6.13**	**10.9**	**2.30**	**4.32**	**9.90**	**15.9**
SNP-RSM	19.9	27.6	10.8	18.0	6.15	12.0	2.44	5.19	11.6	18.7
MC-CNN+RBS	19.5	28.0	10.8	19.3	6.34	14.5	2.60	6.66	10.2	20.9
MC-CNN-acrt	**18.4**	27.7	**10.1**	19.7	6.87	15.7	3.81	11.8	18.0	36.6
r200high	42.0	49.4	34.3	42.5	30.4	38.9	23.1	31.9	53.8	64.5
r200high+SDR	57.4	61.1	24.4	30.8	15.8	22.5	13.0	19.6	35.4	48.3

complex scene structure and different lighting conditions of the dataset make it challenging yet appropriate to evaluate the robustness and accuracy.

Our method is also evaluated in the KITTI 2015 dataset. The KITTI 2015 dataset consists of 200 test image pairs of outdoor driving scenes and additional 200 image pairs with ground truth for training.

2) Parameter Settings: For the energy function (2), we set λ to 0.3 and γ to 20. The bin width L in data term (4) is set to be 2 and the parameter τ in smoothness term (9) is set to 16. The ϵ in (6) is set to 0.01.

For the RANSAC plane fitting, η is set to 0.001 and the minimum inlier ratio ε_P required to pass the fitting is set to be 0.50. The cutoff range $D_{slanted}$ is set to be 6.

The above-mentioned parameters are only related to the input WTA disparity map except γ and are insensitive to the appearance variances of the color reference image. In fact, the parameters we tuned for the Middlebury dataset also work well for the KITTI dataset. The only different settings for the two datasets are the parameters σ and k for superpixel segmentation because the color images of the two datasets have different appearance and resolutions. The notation of σ and k is the same as [37], in which σ is the parameter of a Gaussian filter used to pre-smooth the color image and k affects the preference of the size of components (superpixels). σ is set to 0.1 for both datasets and k is set to 30 for Middlebury and 40 for KITTI. Note that the disparity plane refinement is parameter-free which reflects its robustness.

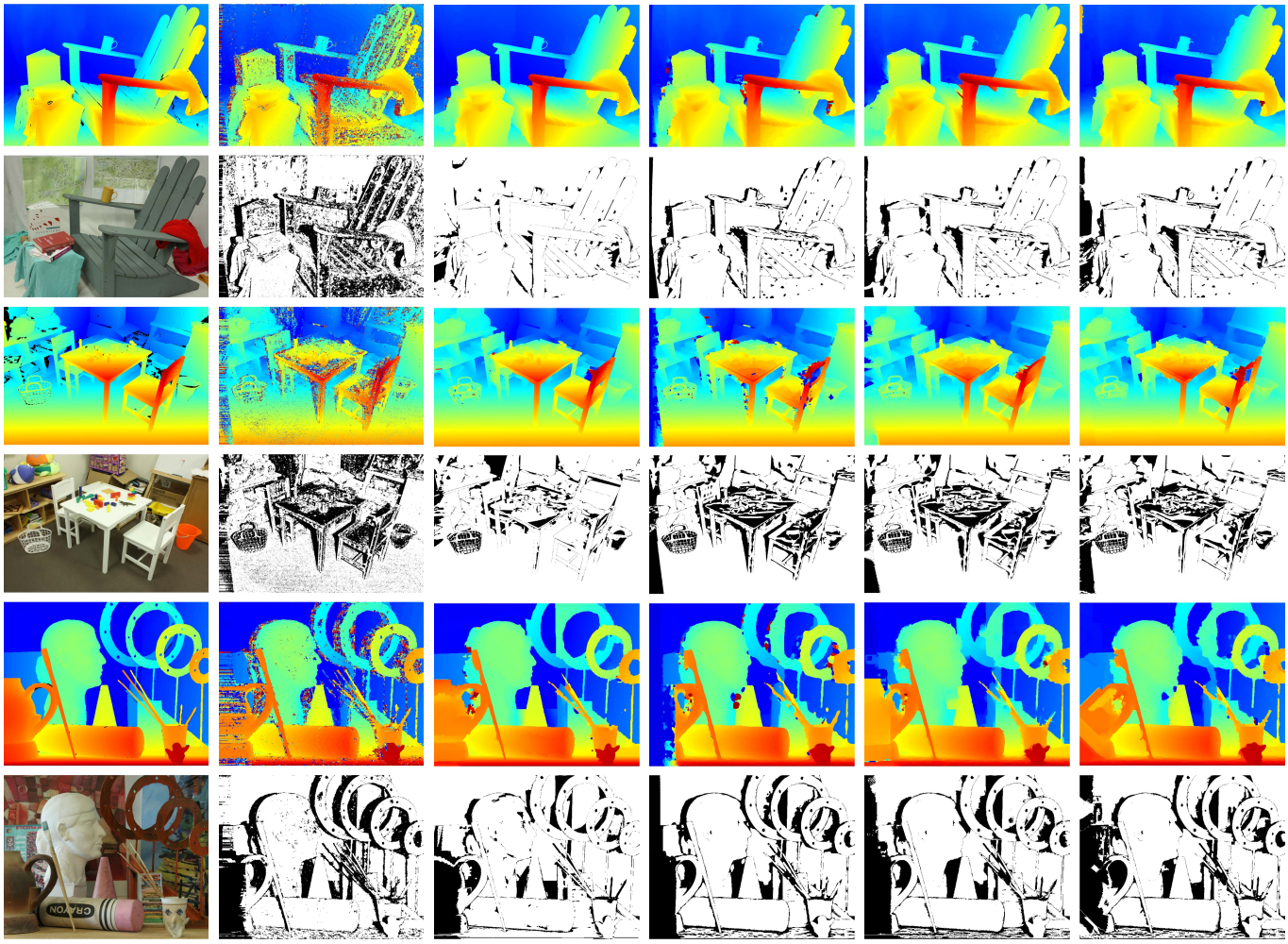
B. Comparison With the State-of-the-Arts

Our segment-based disparity refinement (SDR) is compared with the state-of-the-art disparity refinement methods MC-CNN+RBS [49], SNP-RSM [48], and Mei *et al.* [5]. Since MC-CNN-acrt [6] adopts the post-processing and the refinement of [5] we compare with MC-CNN-acrt instead. In MC-CNN-acrt [6], the matching cost volume is aggregated by SGM [50] and cross based cost aggregation (CBCA) [51], and then a multi-step refinement included LRC, occlusion and mismatch filling, sub-pixel enhancement, a median filter, and a bilateral filter are applied. In contrast, SDR directly refines the WTA disparity map computed from the raw cost volume without cost aggregation. MC-CNN+RBS [49] uses the bilateral solver to refine the output disparity map of MC-CNN-acrt. SNP-RSM [48] utilizes the surface normal predicted by a CNN to refine the output disparity map of MC-CNN-acrt. Although the input disparity map of SDR has a much higher error rate than that of the other three methods, SDR still achieves the best performance among the evaluated methods in most error metrics, as listed in Table I. 'bad 2.0' is the default error metric of the Middlebury dataset which is evaluated on the full resolution ground truth and corresponds to one-pixel error ('bad 1.0') of half resolution images.

The 'noc' mask represents non-occluded pixels and the 'all' mask represents all pixels. As can be seen in Table I, SDR outperforms the other methods with a large margin on 'bad 2.0, all' in both training and test sets. Under 'all' mask, where occluded pixels are considered, SDR performs best in all metrics. This demonstrates the effectiveness of SDR for occlusion handling. A qualitative comparison is shown in Fig. 5, in which the results of Adirondack, PlaytableP and ArtL image pairs are listed. SDR solves the occlusions near the table and chairs accurately. Also, SDR has a more accurate estimation of the slanted surfaces in Adirondack and ArtL than the other three methods.

C. Overall Performance

The overall ranks on the Middlebury 2014 test set evaluated on 'bad 2.0, all' is shown in Table II. SDR ranks 6th by the time of submission and ranks 1st on the Plant image pairs. SDR generates results comparable to the state-of-the-art stereo matching methods even without the cost aggregation which is the main component of these methods. The WTA disparity maps and the disparity maps refined by SDR is shown in Fig. 6a and Fig. 6b, respectively. Although the WTA disparity maps are noise-corrupted especially in occluded and textureless regions, SDR refines the disparities efficiently.



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

Fig. 5. Comparison on Adirondack, PlaytableP, and ArtL image pairs. (a) Color reference images and ground truth disparity maps. (b) WTA disparity maps and corresponding error maps. (c)–(f) Disparity and error maps of SDR, MC-CNN-acrt [6], MC-CNN+RBS [49], and SNP-RSM [48], respectively.

TABLE II

EVALUATION RESULTS ON TEST SET OF MIDDLEBURY 2014 BENCHMARK ON 'BAD 2.0, ALL', TOP-TEN METHODS ARE LISTED HERE. (SNAPSHOT ON MARCH 4, 2018)

Name	Res	Avg	Austr	AustrP	Bicyc2	Class	ClassE	Compu	Crusa	CrusaP	Djemb	DjembL	Hoops	Livgrm	Nkuba	Plants	Stairs
LocalExp	H	**11.7** 1	5.35 2	4.78 2	**6.83** 1	**7.76** 1	**11.6** 1	**14.1** 1	9.71 2	15.2 6	**4.28** 1	12.6 2	23.2 3	14.7 4	21.3 5	12.5 4	18.3 4
MC-CNN+TDSR	F	12.1 2	8.68 9	7.41 12	9.71 8	10.0 3	17.4 6	18.0 5	10.7 4	12.7 3	5.17 3	13.0 3	**21.1** 1	12.6 2	**17.4** 1	12.0 3	**11.0** 1
3DMST	H	12.5 3	5.61 3	**4.77** 1	9.49 5	10.6 5	14.6 3	16.5 2	**9.40** 1	14.7 5	4.92 2	15.1 5	26.3 4	13.0 3	20.3 3	11.8 2	19.3 6
PMSC	H	13.6 4	**5.29** 1	4.97 3	10.7 11	10.4 4	14.2 2	17.7 3	10.1 3	15.6 7	5.30 4	15.4 6	27.2 6	17.2 5	23.2 11	13.0 5	21.1 9
APAP-Stereo	H	13.7 5	7.87 6	7.09 11	8.81 4	13.2 10	28.4 24	21.1 11	12.4 5	**11.4** 1	6.94 9	17.6 12	21.6 2	**10.9** 1	20.3 2	14.6 7	12.9 2
SDR	H	13.8 6	7.90 7	6.66 8	7.97 2	8.95 2	17.1 5	19.1 6	12.6 6	12.2 2	7.35 11	13.9 4	28.4 8	19.7 9	23.5 14	**10.8** 1	20.3 7
NTDE	H	15.3 7	9.68 14	8.37 18	10.9 12	10.9 6	18.0 9	19.8 9	13.2 7	17.8 8	6.20 7	16.9 9	29.4 10	18.0 6	22.0 7	15.7 10	23.5 12
MeshStereoExt	H	15.7 8	8.74 10	8.14 15	10.5 10	12.6 8	19.2 11	17.9 4	19.6 17	19.9 16	7.14 10	16.2 8	26.5 5	18.0 7	22.0 8	16.1 11	18.6 5
OVOD	H	15.8 9	7.04 4	5.61 5	9.51 6	14.0 11	21.4 15	19.3 7	15.7 9	19.2 13	5.90 5	19.1 13	30.0 13	20.7 12	22.3 9	14.7 8	23.7 13
FEN-D2DRR	H	16.0 10	8.66 8	7.95 13	9.59 7	13.0 9	17.9 8	19.4 8	19.3 16	19.6 15	7.87 17	16.0 7	29.0 9	20.1 10	21.5 6	13.9 6	25.1 16

In addition, SDR can handle the inputs that have tremendous noises. The disparity maps of r200high are shown in Fig. 6d, which is computed by the Intel RealSense R200 stereo model [52]. The disparity maps refined by SDR is shown in Fig. 6e. As can be seen, the surfaces of the motorcycle, chairs, and the recycle bin are efficiently refined. SDR reduces the error of r200high significantly as listed in Table I.

SDR is also evaluated on the KITTI 2015 dataset. The KITTI dataset consists of images with similar scene structures and has a loose error threshold 'D1' which evaluates three

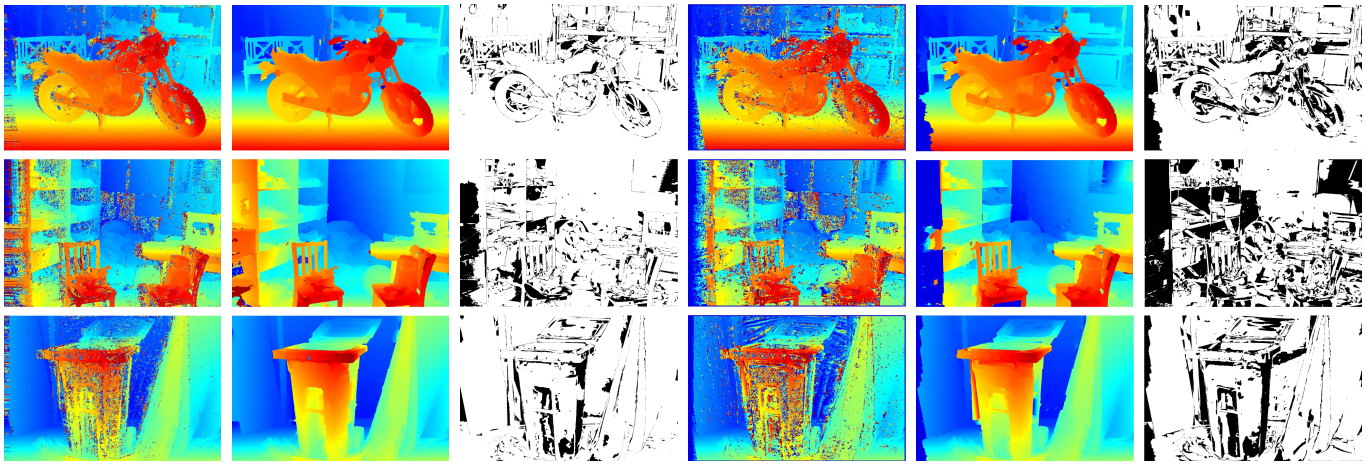


Fig. 6. (a) WTA disparity maps computed by matching cost of MC-CNN [6]. (b) and (c) Are the refined disparity maps of (a) and error maps, respectively. (d) Disparity maps of r200high [52]. (e) and (f) Are the refined disparity maps of (d) and error maps, respectively.

TABLE III

EVALUATION RESULTS ON KITTI 2015 AND MIDDLEBURY 2014 DATASET. IRESNET [54], PSMNET [53], MC-CNN-ACRT [6], AND SDR ARE LISTED HERE. 'BAD 1.0' IS THE ONE PIXEL ERROR AND 'D1' IS THE THREE PIXEL AND 5% ERROR. THE DEFAULT METRICS FOR KITTI AND MIDDLEBURY ARE 'D1' AND 'BAD 2.0', RESPECTIVELY

Dataset	KITTI 2015 Train		KITTI 2015 Test		Middlebury 2014 Train		Middlebury 2014 Test	
Method	bad 1.0	D1	noc	all	noc	all	noc	all
iResNet	-	**1.35**	2.19	2.44	20.3	25.8	22.9	29.5
PSMNet	-	1.83	**2.14**	**2.32**	58.5	61.8	42.1	47.2
MC-CNN-acrt	20.17	3.06	3.33	3.89	**10.1**	19.7	8.08	19.1
SDR	**18.20**	4.88	5.09	5.86	10.1	**15.7**	**7.69**	**13.8**

pixels and 5% error. In contrast, the default error threshold of the Middlebury dataset is one pixel. The color images in KITTI have more noises than that in Middlebury. Due to the above-mentioned differences, there are few methods that achieve top performance on both datasets simultaneously. The evaluation of methods that have both submissions to KITTI and Middlebury benchmarks are listed in Table III. PSMNet [53] and iResNet [54] achieve top performance on KITTI dataset, while much worse performance on Middlebury dataset. MC-CNN-acrt outperforms SDR on the default metric of KITTI dataset. We assume the reason for the degeneration of performance is that the WTA disparity maps have more noises and the color reference images are not as clean as that in the Middlebury dataset. Aggregate the matching cost before the WTA operation can reduce the noises and improve performance. SDR outperforms MC-CNN-acrt when evaluated on 'bad 1.0' metric. This demonstrates that the disparity map estimated by SDR is more precise. Fig. 7 shows the visual comparison of SDR and MC-CNN-acrt. SDR estimates the road signs (lines 1, 4) and the distant cars (line 2) more accurately and is able to handle the textureless overexposure regions (line 2, 3). Even there are tremendous noises, the surfaces of the road and cars can be recovered (line 5).

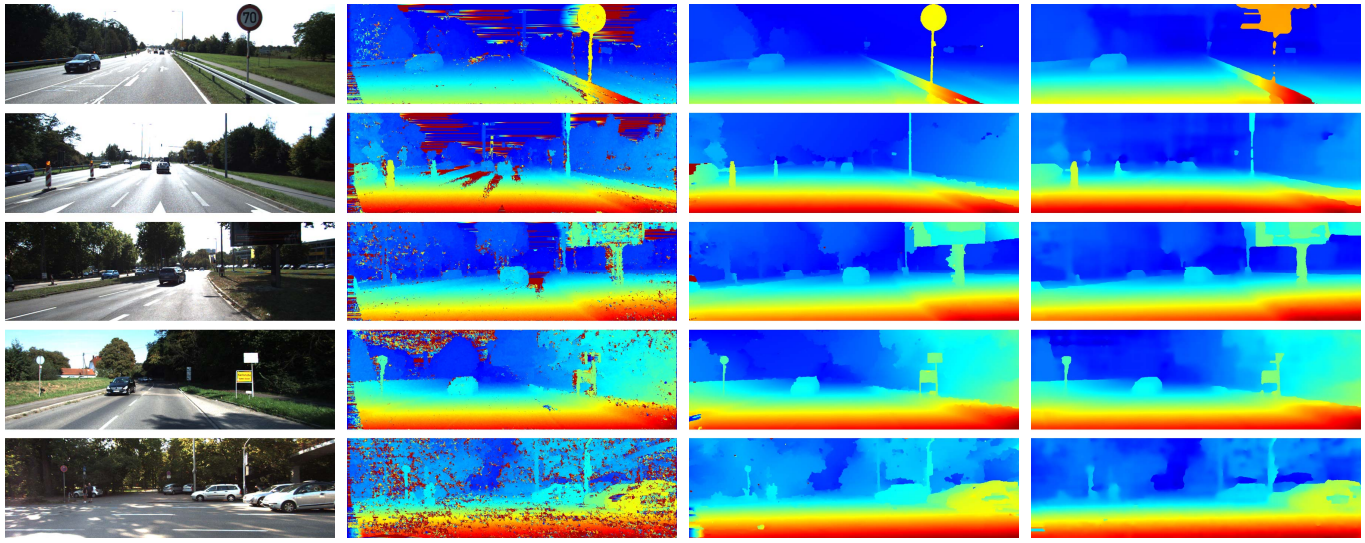
D. Effect of Parameters

The key parameter in the MRF inference is the regularization parameter λ in (2). This parameter controls the influence of smoothness constraints. When λ is set to 0, the smoothness constraints lose their effect and hence the estimation degenerates to a local one. As λ increases to infinity, the smoothness constraints dominate the estimation, which will assign the same label to all vertices. Therefore, the proper setting of parameter λ is necessary for good performance. The evaluation results of λ are shown in Fig. 8a. We plotted the error rates of three image pairs and the average errors of Middlebury and KITTI training set. The plotted five curves have a low error rate when λ varies from 0.1 to 0.7. The best setting of λ is 0.3 for both Middlebury and KITTI, which demonstrates the insensitivity of our method to parameter λ.

Parameter k affects the size of superpixels and a larger k causes a preference for larger superpixels [37]. As a result, small k deals with details of images while large k can handle the large textureless regions. Our method is insensitive to k in the given range as shown in Fig. 8b.

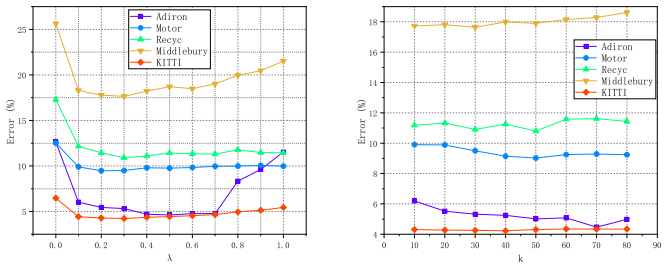
E. Effect of Disparity Plane Refinement

To demonstrate the contribution of disparity plane refinement, we propose three comparative experiments. The local optimization layer of SDR consists of disparity plane fitting (PF) and the Bayesian inference based disparity plane refinement (BR), hence the original method is denoted as $PF + BR$. We keep the plane fitting module unchanged and compare the disparity plane refinement method. The first comparison is the MAP based disparity plane refinement



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

Fig. 7. Results on KITTI 2015 dataset. (a) Color reference images. (b) WTA disparity maps. (c) Disparity maps refined by SDR. (d) Disparity maps of MC-CNN-acrt [6].



(a) (b)

Fig. 8. Effect of parameters (a) λ and (b) k. Errors of 3 image pairs Adirondack, Motorcycle, and Recycle and average errors of Middlebury ('bad 2.0, all') and KITTI ('D1, noc') training sets are listed.

TABLE IV

EVALUATION ERROR ON THE MIDDLEBURY TRAINING SET FOR DIFFERENT DISPARITY PLANE REFINEMENT METHODS. BOTTOM ROW IS OUR METHOD WITH 2D NEIGHBORHOOD SYSTEM

Method	bad 1.0	bad 2.0	avgerr	rms
PF+BR	**26.9**	**15.7**	**4.32**	**15.9**
PF+MAP	28.4	19.1	5.22	17.8
PF+WMF	34.6	21.9	5.64	17.4
PF	28.5	20.7	15.7	42.5
PF+BR (2DN)	28.6	17.7	5.00	16.9

which is denoted as $PF + MAP$. In $PF + MAP$, the label with the highest posterior probability of neighbors is assigned to superpixel s. In BR, we use weighted least squares (WLS) to generate models with the posterior predictive probability of sample points as weight. In contrast with WLS, we propose a weighted mean filtering (WMF) to generate new models, denoted as $PF+WMF$. The disparity plane label $\boldsymbol{\pi}_s$ of superpixel s is refined by the weighted average of its neighbors' labels with their posterior probability as the corresponding weight. At last, we consider only disparity plane fitting PF into comparison, to verify the importance of disparity plane refinement.

The results of this comparison are listed in Table IV. As can be seen, the methods with disparity plane refinement have better results than PF only for the four evaluation metrics. $PF+WMF$ has low error rates in metrics 'avgerr' and 'rms'. But it has high error rates in 'bad 2.0' metric because the weighted average of neighbors labels takes no consideration of the location information. $PF + MAP$ has better performance than $PF + WMF$ and PF in 'bad 2.0" and 'avgerr' metrics while its high 'rms' error uncovers its poor robustness. $PF + BR$ achieves best results for all metrics, which demonstrates its superiority for disparity plane refinement. The 3D neighborhood system is a core component of the disparity plane refinement. To demonstrate its importance, a comparison that performs the Bayesian inference based refinement in 2D neighborhood system, denoted by $PF + BR$ $(2DN)$, is listed in Table IV. The 2D neighborhood system is the common image domain neighborhood system. The 3D neighborhood system is the subset of the 2D neighborhood system because 2D neighbors with similar mean disparities are defined as 3D neighbors. The mean disparities are estimated by the MRF inference and therefore provide global scene information. In 2D neighborhood system, the occluded regions with failed plane fitting results are interpolated by its occluding and non-occluded neighbors. Whereas the same regions are interpolated only by its non-occluded neighbors in 3D neighborhood system. As a result, methods with 3D neighborhood system performs better on occlusion handling. We provide the results of our method with the 2D neighborhood and 3D neighborhood system for visual comparison, as shown in Fig. 9. The occluded regions are wrongly interpolated for $PF + BR$ $(2DN)$ due to the interference of foreground occluding neighbors.

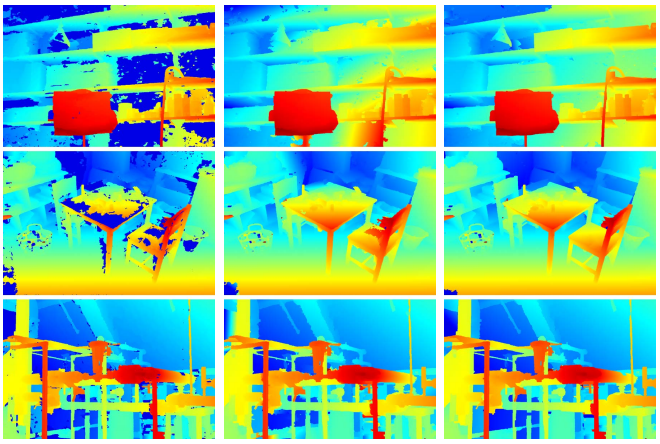


Fig. 9. Visual comparison of different methods on Shelves, Playtable, and Pipes image pairs. (a) PF only. (b) PF+BR with 2D neighborhood system. (c) PF+BR with 3D neighborhood system.

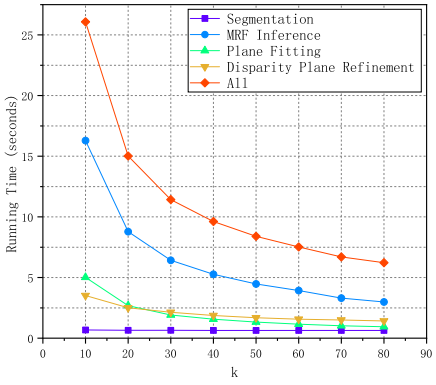


Fig. 10. Running time of each component of our method and the overall running time.

F. Efficiency Analysis

For MRF stereo methods, the computational complexity is expressed as $O(MD)$, where M is the number of nodes in the graph model and D is the number of labels. Due to the superpixel representation, M in our approach is equal to the number of superpixels which is only about one percent of the number of pixels in grid graphs and $D = D_{max}/L$, which results in a much lower computational cost. Furthermore, the proposed disparity refinement method required no matching cost and thus saves the memory for storing the cost volume.

We implement SDR in C++ and OpenCV on a PC with an 2.6GHz CPU (using a single core). We changed the parameter k to generate different numbers of superpixels and evaluate the corresponding running time of each component in Middlebury dataset as shown in Fig. 10. The number of superpixels decreases with increasing k, thereby leading to shorter running time. The most time-consuming part is the MRF inference which is considered as a bottleneck for further improvement in speed. The average running time in KITTI 2015 dataset is 4.2s. If the solver of the MRF optimization is changed to expansion moves [18] with $T = T_l$, and k is set to 80, the average running time reduced to 1.7s and the accuracy is still comparable to the original (4.40 on 'D1, noc'). In addition, both the plane fitting and the disparity plane refinement are the local method and can be accelerated by parallelization.

VII. Conclusions

We have developed a faster disparity refinement that directly refines the WTA disparity map by exploring its statistical significance. There are two key components in SDR: one is the MRF inference which is able to estimate mean disparities of superpixels as well as the mean disparities of occluded regions, and the other is the Bayesian inference based disparity plane refinement which generates a smooth disparity map. Experiments demonstrate the good performance in occlusion handling with low computational cost.

Further improvements include introducing explicit occlusion term in the MRF inference to estimate mean disparities more accurately and adding B-splines models to fitting to approximate surfaces of objects more precisely.

References

[1] D. Scharstein *et al.*, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proc. German Conf. Pattern Recognit.*, vol. 8753, 2014, pp. 31–42.
[2] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, nos. 1–3, pp. 7–42, Apr. 2002.
[3] S. Birchfield and C. Tomasi, "Depth discontinuities by pixel-to-pixel stereo," in *Proc. Int. Conf. Comput. Vis.*, 2002, pp. 1073–1080.
[4] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Computer Vision—ECCV*. Berlin, Germany: Springer, 1994, pp. 151–158.
[5] X. Mei, X. Sun, M. Zhou, H. Wang, X. Zhang, and S. Jiao, "On building an accurate stereo matching system on graphics hardware," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Nov. 2012, pp. 467–474.
[6] J. Žbontar and Y. Lecun, "Computing the stereo matching cost with a convolutional neural network," in *Proc. Comput. Vis. Pattern Recognit.*, 2015, pp. 1592–1599.
[7] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5695–5703.
[8] K.-J. Yoon and I. S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 650–656, Apr. 2006.
[9] Q. Yang, "A non-local cost aggregation method for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1402–1409.
[10] L. Li, X. Yu, S. Zhang, X. Zhao, and L. Zhang, "3D cost aggregation with multiple minimum spanning trees for stereo matching," *Appl. Opt.*, vol. 56, no. 12, pp. 3411–3420, 2017.
[11] G. Egnal, M. Mintz, and R. P. Wildes, "A stereo confidence metric using single view imagery with comparison to five alternative approaches," *Image. Vis. Comput.*, vol. 22, no. 12, pp. 943–957, 2004.
[12] W.-S. Jang and Y.-S. Ho, "Discontinuity preserving disparity estimation with occlusion handling," *J. Vis. Commun. Image Represent.*, vol. 25, no. 7, pp. 1595–1603, 2014.
[13] A. Banno and K. Ikeuchi, "Disparity map refinement and 3D surface smoothing via directed anisotropic diffusion," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Sep./Oct. 2011, pp. 1870–1877.
[14] X. Huang and Y. J. Zhang, "An O(1) disparity refinement method for stereo matching," *Pattern Recognit.*, vol. 55, pp. 198–206, Jul. 2016.
[15] 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.*, Dec. 2014, pp. 49–56.
[16] Y. Zhan, Y. Gu, K. Huang, C. Zhang, and K. Hu, "Accurate image-guided stereo matching with efficient matching cost and disparity refinement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 9, pp. 1632–1645, Sep. 2016.

[17] B. Tippetts, D. J. Lee, K. Lillywhite, and J. Archibald, "Review of stereo vision algorithms and their suitability for resource-limited systems," *J. Real-Time Image Process.*, vol. 11, no. 1, pp. 5–25, 2016.

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

[19] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Jul. 2001, pp. 508–515.

[20] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004.

[21] V. Kolmogorov and R. Zabin, "What energy functions can be minimized via graph cuts?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 147–159, Feb. 2004.

[22] P. F. Felzenszwalb and D. R. Huttenlocher, "Efficient belief propagation for early vision," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2004, pp. I-261–I-268.

[23] Q. Yang, L. Wang, R. Yang, H. Stewénius, and D. Nistér, "Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 2347–2354.

[24] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1568–1583, Oct. 2006.

[25] C. Olsson, J. Ulèn, and Y. Boykov, "In defense of 3D-label stereo," in *Proc. Comput. Vis. Pattern Recognit.*, 2013, pp. 1730–1737.

[26] F. Besse, C. Rother, A. W. Fitzgibbon, and J. Kautz, "PMBP: Patchmatch belief propagation for correspondence field estimation," *Int. J. Comput. Vis.*, vol. 110, no. 1, pp. 2–13, 2014.

[27] M. Bleyer, C. Rhemann, and C. Rother, "PatchMatch stereo—Stereo matching with slanted support windows," in *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 14-1–14-11.

[28] L. Li, S. Zhang, X. Yu, and L. Zhang, "PMSC: PatchMatch-based superpixel cut for accurate stereo matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 3, pp. 679–692, Mar. 2018.

[29] T. Taniai, Y. Matsushita, Y. Sato, and T. Naemura, "Continuous 3D label stereo matching using local expansion moves," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2725–2739, Nov. 2018.

[30] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in *Proc. Int. Conf. Pattern Recognit.*, vol. 2006, pp. 15–18.

[31] L. Hong and G. Chen, "Segment-based stereo matching using graph cuts," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun./Jul. 2004, p. 1.

[32] H. Tao, H. S. Sawhney, and R. Kumar, "A global matching framework for stereo computation," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Jul. 2001, pp. 532–539.

[33] M. Bleyer, C. Rother, and P. Kohli, "Surface stereo with soft segmentation," in *Proc. Comput. Vis. Pattern Recognit.*, 2010, pp. 1570–1577.

[34] K. Yamaguchi, T. Hazan, D. Mcallester, and R. Urtasun, "Continuous Markov random fields for robust stereo estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 45–58.

[35] S. Xu, F. Zhang, X. He, X. Shen, and X. Zhang, "PM-PM: Patchmatch with potts model for object segmentation and stereo matching," *IEEE Trans. Image Process.*, vol. 24, no. 7, pp. 2182–2196, Jul. 2015.

[36] Z.-F. Wang and Z.-G. Zheng, "A region based stereo matching algorithm using cooperative optimization," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

[37] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, Sep. 2004.

[38] Y. Deng, Q. Yang, X. Lin, and X. Tang, "Stereo correspondence with occlusion handling in a symmetric patch-based graph-cuts model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1068–1079, Jun. 2007.

[39] V. Kolmogorov and C. Rother, "Minimizing nonsubmodular functions with graph cuts—A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 7, pp. 1274–1279, Jul. 2007.

[40] M. Tang, I. Ben Ayed, and Y. Boykov, "Pseudo-bound optimization for binary energies," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2014, pp. 691–707.

[41] T. Taniai, Y. Matsushita, and T. Naemura, "Superdifferential cuts for binary energies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2030–2038.

[42] L. Gorelick, Y. Boykov, O. Veksler, I. Ben Ayed, and A. Delong, "Local submodularization for binary pairwise energies," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 10, pp. 1985–1999, Oct. 2017.

[43] K. Ni, H. Jin, and F. Dellaert, "GroupSAC: Efficient consensus in the presence of groupings," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 2193–2200.

[44] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.

[45] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[46] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J.-M. Frahm, "USAC: A universal framework for random sample consensus," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 2022–2038, Aug. 2013.

[47] S. J. D. Prince, *Computer Vision: Models, Learning, and Inference*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[48] S. Zhang, W. Xie, G. Zhang, H. Bao, and M. Kaess, "Robust stereo matching with surface normal prediction," in *Proc. IEEE Int. Conf. Robot. Autom.*, May/Jun. 2017, pp. 2540–2547.

[49] J. T. Barron and B. Poole, "The fast bilateral solver," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 617–632.

[50] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

[51] K. Zhang, J. Lu, and G. Lafruit, "Cross-based local stereo matching using orthogonal integral images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 7, pp. 1073–1079, Jul. 2009.

[52] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel realsense stereoscopic depth cameras," in *Proc. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 1267–1276.

[53] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5410–5418.

[54] Z. Liang *et al.*, "Learning for disparity estimation through feature constancy," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2811–2820.

Tingman Yan received the B.S. degree in guidance, navigation, and control from Beihang University, Beijing, China, in 2016. He is currently pursuing the Ph.D. degree in automation with Shanghai Jiao Tong University, Shanghai, China. His current research interests include stereo matching and multi-view 3D reconstruction.

Yangzhou Gan received the B.S. degree in automation from the University of Electronic Science and Technology of China in 2010, and the Ph.D. degree in control science and engineering from Shanghai Jiao Tong University, China, in 2015. He is currently an Assistant Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. His research interests include image processing and biomechanics.

Zeyang Xia received the B.S. degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2002, and the Ph.D. degree in mechanical engineering from Tsinghua University, Beijing, China, in 2008. He is currently a Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, and also the Director of the Medical Robotics and Biomechanics Laboratory. He has published over 80 peer-reviewed papers and has applied for over 40 patents. His research interests include biped humanoid robotics, medical robotics, and dental biomechanics. He is the Chairman of the Guangzhou Branch of the Youth Innovation Promotion Association, Chinese Academy of Sciences, and the Co-Chair of the Guangdong Chapter of the IEEE Robotics and Automation Society. He served as the Program Co-Chair of the IEEE RCAR 2016 and the ICVS 2017, and will be the General Chair of the IEEE RCAR 2019.

Qunfei Zhao received the B.S.E.E. degree from Xi'an Jiao Tong University, Xi'an, China, in 1982, and the Sc.D. degree in system science from the Tokyo Institute of Technology, Tokyo, Japan, in 1988. He is currently a Professor with the School of Electronic Information and Electric Engineering, Shanghai Jiao Tong University, China. His research interests include robotics, machine vision, and optimal control of complex mechatronic systems.