Front. Comput. Sci.
https://doi.org/10.1007/s11704-018-6146-6

RESEARCH ARTICLE

Joint view synthesis and disparity refinement for stereo matching

Gaochang Wu[1,2], **Yipeng Li**[2], **Yuanhao Huang**[3], **Yebin Liu** (✉)[2]

1 State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, China

2 Broadband Network & Digital Media Lab, Department of Automation, Tsinghua University, Beijing 100084, China

3 Orbbec Company, Shenzhen 518061, China

Abstract Typical stereo algorithms treat disparity estimation and view synthesis as two sequential procedures. In this paper, we consider stereo matching and view synthesis as two complementary components, and present a novel iterative refinement model for joint view synthesis and disparity refinement. To achieve the mutual promotion between view synthesis and disparity refinement, we apply two key strategies, disparity maps fusion and disparity-assisted plane sweep-based rendering (DAPSR). On the one hand, the disparity maps fusion strategy is applied to generate disparity map from synthesized view and input views. This strategy is able to detect and counteract disparity errors caused by potential artifacts from synthesized view. On the other hand, the DAPSR is used for view synthesis and updating, and is able to weaken the interpolation errors caused by outliers in the disparity maps. Experiments on Middlebury benchmarks demonstrate that by introducing the synthesized view, disparity errors due to large occluded region and large baseline are eliminated effectively and the synthesis quality is greatly improved.

Keywords stereo matching, view synthesis, disparity refinement.

1 Introduction

Stereo matching is one of the most extensively issues in computer vision. It plays an important role in a large variety of computer vision applications including object recognition, human tracking, image segmentation [1], image-based rendering [2] as well as 3D photography. The goal of stereo matching is to determine a disparity map that indicating the corresponding pixels in two views of a scene. The disparity map can be easily converted to a pixel-level depth map representing three-dimensional information of the scene. Once the depth of a scene is recovered, a virtual image of any viewpoint between the two views can be obtained by using depth-image-based rendering (DIBR) techniques.

In the field of stereo vision, stereo matching and view synthesis are always considered as two sequential procedures. On the one hand, high-accurate depth maps or disparity maps produce high quality synthesized views. On the other hand, stereo accuracy benefits from multiview techniques [3] or 4D light field [4], i.e., additional views improve disparity quality [5]. For the existing binocular stereo matching approaches, they tend to fail in occluded and disparity discontinuous region (Fig. 1(b)). If a middle view is employed, the stereo quality can be greatly improved due to the less occluded regions (Fig. 1(c)). For the practical binocular stereo, only two real views can be captured by the device. Although the DIBR techniques are able to provide additional views [6], we encounter a chicken-and-egg problem that the interpolation quality depends on stereo accuracy.

In this paper, we present a joint view synthesis and disparity refinement model that takes both view synthesis quality and disparity accuracy into account. To the best of our knowledge, no one has ever attempted to solve this problem

Received March 17, 2018; accepted July 30, 2018

E-mail: liuyebin@mail.tsinghua.edu.cn

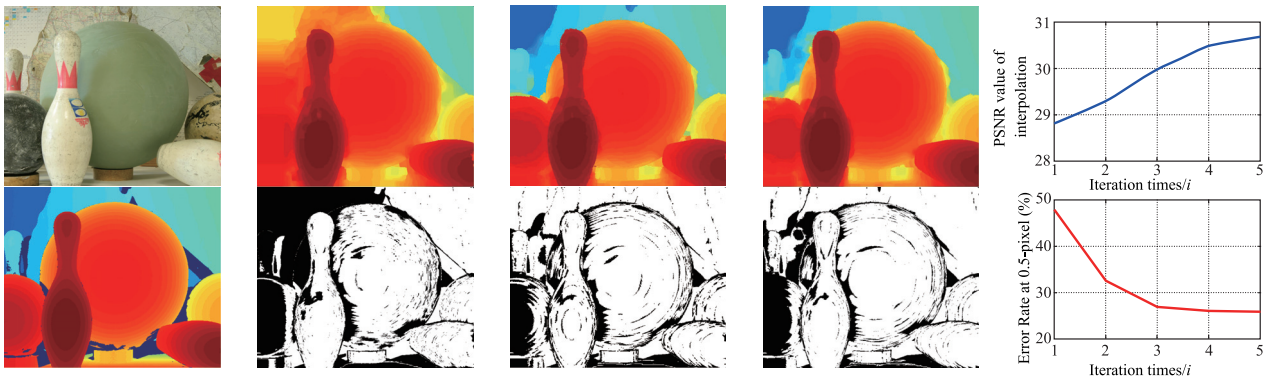


Fig. 1 Disparity maps using different views. (a) Left view and its ground truth disparity map; (b) disparity map produced by the original input image pair and its error map at 0.5-pixel threshold (31.83%); (c) disparity map produced by using the original image pair and an additional ground truth middle view and its error map (20.58%); (d) disparity map produced by the proposed joint view synthesis and disparity refinement model and its error map (25.80%). Note that relative to the error map of (c), most errors are centered on the left borders which are invisible in the input right view; (e) PSNR value of the synthesized view and disparity error rate during the iteration

in the existing binocular stereo approaches. In order to tackle the chicken-and-egg problem, we therefore resort to an elaborately designed coarse-to-fine framework that uses both the original image pair and a synthesized virtual view to gradually refine the disparity maps (Fig. 1(d)). The framework is mainly composed by two strategies (Fig. 2) to achieve the mutual promotion between the disparity maps and the synthesized view: 1) a disparity maps fusion scheme to detect and neutralize the stereo errors after the view synthesis step; and 2) a disparity-assisted plane sweep-based rendering (DAPSR) method to weaken the influence on synthesized view from bad pixels of the disparity maps after the disparity estimation step. These two strategies also ensure that, along with the iterations the model outputs not only refined disparity maps but also an elegant virtual view with higher visual coherency, see Figs. 1(d) and 1(e).

More specifically, the main obstacles we are facing are two-folds: 1) The stereo quality will certainly be affected by view synthesis artifacts; 2) Bad pixels in the disparity maps will cause view synthesis errors. On the one hand, to handle the first obstacle, the proposed disparity maps fusion scheme is employed to produce the final disparity maps that are robust to the synthesis artifacts (Fig. 2(a)). Two intermediate disparity maps will be produced using the synthesized view as reference image and one of the inputs (left or right) views as target image. We indicate that the minor errors are always coupled in these two intermediate disparity maps, i.e., appear on the same position of the two disparity maps and share the same absolute value but opposite in sign. These errors will be counteracted during the fusion. On the other hand, to handle the second obstacle, the proposed DAPSR method is employed to interpolate and update the virtual view (Fig. 2(b)). The key idea of the DAPSR is that the virtual view can be produced by a soft blending of the swept input images with the assistance of disparity maps. This interpolation strategy is robust to small errors in the input disparity, which is crucial for the initial interpolation.

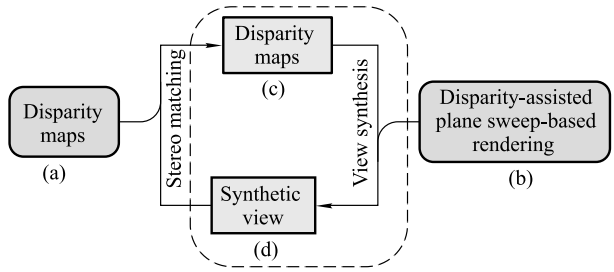


Fig. 2 Two strategies (a) disparity maps fusion and (b) disparity-assisted plane sweep-based rendering (DAPSR) are applied to the proposed model. The dash block indicates the conventional relationship between stereo matching and view synthesis, i.e., they are considered as two sequential procedures

The contributions of the paper are summarized as follows:

1. We proposed a novel joint virtual view synthesis and disparity refinement model that outputs not only refined disparity maps but also a synthesized middle view with high visual coherency;
2. We develop a disparity maps fusion scheme to eliminate the error caused by potential interpolation artifacts, providing new disparity maps for the synthesized view updating;
3. We introduce a disparity-assisted plane sweep-based rendering (DAPSR) method to weaken interpolation errors caused by bad pixels in the disparity maps.

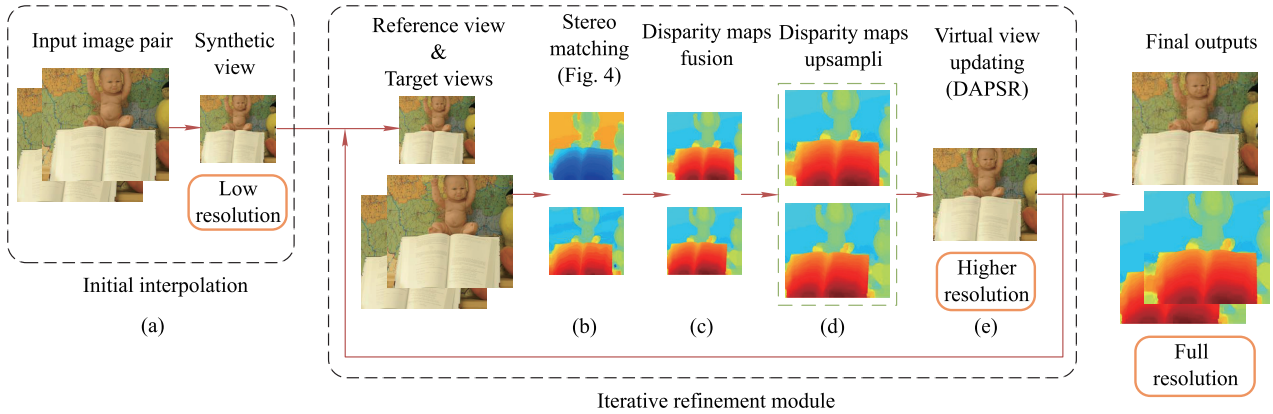


Fig. 3 Framework of the proposed joint view synthesis and disparity refinement model

We provide a new thought of the relationship between stereo matching and view synthesis. We also demonstrate that the proposed model can improve the performance of many local and global stereo matching algorithms [7–9] (Fig. 2(c)) after integrating the proposed blocks in Figs. 2(a) and 2(b).

The rest of the paper is organized as follows: Section 2 reviews some existing approaches in view synthesis and stereo matching. Section 3 provides an overview of the proposed model. Section 4 presents the detailed description of the patch-based stereo matching and the DAPSR method. Section 5 gives the detailed description of the proposed iterative refinement model for joint view synthesis and disparity refinement. Except the proposed patch-based stereo matching method, other methods will also be employed to the model to demonstrate the universal applicability. The experimental result in terms of interpolation quality and stereo accuracy and the robustness of the proposed stereo matching method will be presented in Section 6. Section 7 gives a conclusion and some discussions of the model.

2 Related works

In this paper, we mainly tackle the problem of the combination of view synthesis and stereo matching. We present the first model for joint view synthesis and disparity refinement that use a synthesized image as reference view to perform stereo matching. Therefore, the problem is split into two parts: view synthesis and stereo matching.

2.1 View synthesis

View synthesis is a common technique extrapolating or interpolating a view using other available views, which is widely applied to 3D display, such as 3D Video (3DV) and Free-viewpoint TV (FTV) [10]. To synthesize a novel view, most of state-of-the-art view synthesis approaches are based on DIBR techniques that use the combination of a textured image with a corresponding depth map or disparity map [6, 11]. Point based rendering approaches [12] warp each pixel in the texture image to the novel viewpoint based on the depth map. The direct warping of pixels can produce undesired holes or conflicting information from different views, particularly in occluded regions. Mori et al. [13] proposed a 3D warping approach to solve the problem. The depth maps from different views are first warped to the desired view, then a median filter is applied to fill the holes and a bilateral filter is applied to smooth the warped depth maps. Finally, the images in the real viewpoint are warped to the novel view according to the warped depth maps, and the synthesized view is the blending of them. The 3D warping approach produce a good result in hole-filling, while the pixel conflict problem may corrupt the synthesized view. Fickel et al. [2] proposed a triangle mesh based rendering approach that uses multiple triangle meshes as rendering proxies. Wanner and Goldluecke [14] introduced a variational light field angular super-resolution framework by utilizing the estimated depth map to warp the input images to the novel views. Zhang et al. [15] further presented a generative variational model to enable per-pixel viewpoint assignment for stereoscopic 3D images generation. Zhang et al. [16] followed the idea and produce high quality disparity maps as well as rendering results. However, the interpolation result using triangle based rendering approach depends much on the depth or disparity quality. In the proposed model, the initial disparity maps are in low resolution, and quality of stereo is assumed to be coarse, therefore, the mesh based rendering approach is not competent in our model.

In recent years, some studies for maximizing the quality of synthesized views have been presented that are based on CNNs. Kalantari et al. [17] used two sequential convolutional

neural networks to model depth and color estimation simultaneously by minimizing the error between synthesized views and ground truth images. Wu et al. [5] presented a depth-free framework to reconstruct a high angular resolution light field using extracted epipolar plane images.

2.2 Stereo matching

Stereo matching approaches are usually composed of four major components [18]: 1) computation of matching cost for each pixel; 2) cost aggregation in the support regions; 3) disparity computation; and 4) refinement of the disparity map. The most common matching cost features are absolute difference (AD), zero-mean normalized cross correlation (ZNCC), gradient and census-based measures. Single feature measurement is sensitive to the defect of the measure, while the combination of features can achieve better result, such as the combination of AD, census and scale-invariant feature transform (SIFT) [19] and AD, census and gradient [20]. In addition to these common features, convolutional neural networks (CNN) are also employed to extract features of small patches of the stereo image pair [21, 22].

The existing stereo matching approaches can be divided into two categories, local and global, depending on how the cost aggregation and disparity computation procedures are performed. Local approaches are mainly based on a simple assumption that for the pixels having similar color or intensity within a support window share similar disparities. Support weights [23–25] for every pixels within the window is utilized to aggregate the costs together and assign the best disparity value based on the minimum aggregated cost. Rhemann et al. described the local aggregation as a cost filtering problem in [7]. Local approaches are considered faster yet less accurate than global approaches.

Global approaches takes account the overall structure of the image by building an explicit data term and a pairwise smoothness term and are assigning all disparities simultaneously by employing energy minimization techniques such as graph cuts (GC) [26, 27], belief propagation (BP) [28], dynamic programming and markov random field (MRF). Zhang et al. [16] partitioned the input images into 2D triangles, and lifted the 2D triangles to 3D mesh using a two-layer MRF. Lee et al. [8] and Guney et al. [29] decomposed images into a set of superpixels, and obtained the disparity by applying MRF. Psota et al. [30] converted the minimizing global energy problem into maximizing a posteriori (MAP) using Hidden Markov Trees (HMT). Mozerov and Van [31] combined the cost filtering and energy minimization methods by applying a two step energy minimization approach: a fully connected model and a conventional locally connected model, and achieved a remarkable result.

Whether for local approaches or global techniques, the main challenges are disparity smoothness and occlusion handling. These two kinds of approaches work in different ways to tackle this paradox and have their own merits. For local algorithms such as dynamic window [32,33] and weighted window [25], the potential disparity plane of the target pixel is first estimated based on the RGB information prior, then the disparity is selected by winner-takes-all (WTA) algorithm. For global technique, the overall scene structure and disparity smoothness are taken into account to determine the disparity map. Main methods to solve this optimization problem are MRF [31], belief propagation (BP) [28, 34, 35] and graph cut (GC) [26, 27, 36].

3 Algorithm overview

The overall framework of the proposed joint view synthesis and disparity refinement model is shown in Fig. 3. First, an initial middle view $I_S^{(0)}$ in low resolution (180 pixels in image width in our implementation) is first synthesized using an initial disparity map, and the left view I_L and the right view I_R as input (Fig. 3(a)). The initial disparity is generate by a proposed patch-based stereo matching approach as shown in Fig. 4, and the employed view synthesis method is same as the one for virtual view updating (Fig. 3(e)).

The initial low resolution middle view $I_S^{(0)}$ is input to the iterative refinement module, and will be updated by its output. In the next (ith) iteration, we use the virtual view $I_S^{(i)}$ as reference view, and the left view I_L and the right view I_R as target view, respectively, to obtain two intermediate disparity maps $D_{SL}^{(i)}$ and $D_{SR}^{(i)}$ by employing the proposed stereo matching method (Fig. 3(b)). The desired disparity maps $D_{LR}^{(i)}$ and $D_{RL}^{(i)}$ are the fusion of the two disparity maps $D_{SL}^{(i)}$ and $D_{SR}^{(i)}$ (Fig. 3(c)). This disparity maps fusion block is able to detect and neutralize the initial interpolation error caused by the relatively low quality stereo, where the detail will be described in Section 4.2. The disparity maps $D_{LR}^{(i)}$ and $D_{RL}^{(i)}$ are upsampled (Fig. 3(d)) using method in [37]. This block should be skipped when the disparity maps reach the full resolution in the last few iterations. Finally, the proposed DAPSR method is used to update the virtual view $I_S^{(i+1)}$ into a higher resolution (Fig. 3(e)).

The proposed model is a coarse-to-fine framework that eventually outputs high quality synthesized view and dispar-

ity maps. The initial low quality and low resolution synthesized view will be updated to full resolution iteratively until obtaining a desired high quality view. It should be highlighted that other stereo matching methods are also appropriate for the model by replacing the stereo matching step in (Fig. 3(b)).

4 Joint view synthesis and disparity refinement model

4.1 Stereo matching

Figure 4 depicts the framework of the proposed patch-based stereo matching method. First, the left image I_L and the right image I_R are converted into gray-scale and downsampled to the same resolution as $I_S^{(i)}$, denoted as $I_L^{(i)}$ and $I_R^{(i)}$, where i means the ith iteration. Then the proposed patch-based stereo matching method is applied to obtain disparity maps $D_{SL}^{(i)}$ and $D_{SR}^{(i)}$ using $I_S^{(i)}$ as reference view, $I_L(i)$ and $I_R(i)$ as target view, respectively. For initial disparity estimation, only the left image $I_L^{(0)}$ and the right image $I_R^{(0)}$ are used as input. In the following, we introduce the proposed stereo matching method using I_L and I_R as the input.

4.1.1 Patch-based local matching

In this procedure, we build a matching cost volume by performing patch-based matching. The most common cost features are absolute differences (AD), Birchfield and Tomasi's pixel dissimilarity, normalized cross correlation (NCC), gradient-based measures and census-based measures. Single feature measurement is sensitive to the defect of the feature, while combination of features achieves a more robust measurement. Klaus et al. [34] proposed a linear combination of sum of absolute differences (SAD) and gradient. Mei et al. [32] proposed a normalized combination of AD and census. Their combinations of features achieve applausive results, but seems straightforward and relying on parameters. Therefore, we proposed an improved feature combination strategy that automatically determine the weight between the features. Given a 5×5 patch P_L centered on pixel p_L in the left image I_R and a patch $P_{R,d}$ in the right image I_R with disparity d, the matching cost in our work contains two parts, gradient feature $C_{gradient}(P,d)$ and census feature $C_{census}(P,d)$. Gradient information is incorporated into patch matching to improve searching accuracy for similar patches [38]. We also use first- and second-order derivatives as features to calculate matching cost. The features is extracted by applying four 1-D gradient filters:

$$\begin{aligned} G_1 &= [-1,0,1], \quad G_2 = G_1^T, \\ G_3 &= [1,0,-2,0,1], \quad G_4 = G_3^T. \end{aligned} \tag{1}$$

The extracted feature $\boldsymbol{F}_g$ is represented as concatenation of the vectorized filter outputs. The cost value $C_{gradient}(P,d)$ is defined as the $\mathcal{L}_2$ distance between the features:

$$C_{gradient}(P,d) = \frac{1}{dim(\boldsymbol{F}_g)} \|\boldsymbol{F}_g(P_L) - \boldsymbol{F}_g(P_{R,d})\|_2, \tag{2}$$

where $\boldsymbol{F}_g(P_L)$ is the feature of the patch P_L, $\boldsymbol{F}_g(P_{R,d})$ is the feature of the patch $P_{R,d}$, and $dim(\boldsymbol{F}_g)$ is dimension of the feature.

For census feature $C_{census}(P,d)$, it is defined as the weighted sum of Hamming distance of each pixel in the patches. For each pixel q_L in the patch P_L, a 5×5 window is used to encode its local structure in a 32-bit vector, denoting as $\boldsymbol{F}_c$ and $C_{census}(P,d)$ is computed as:

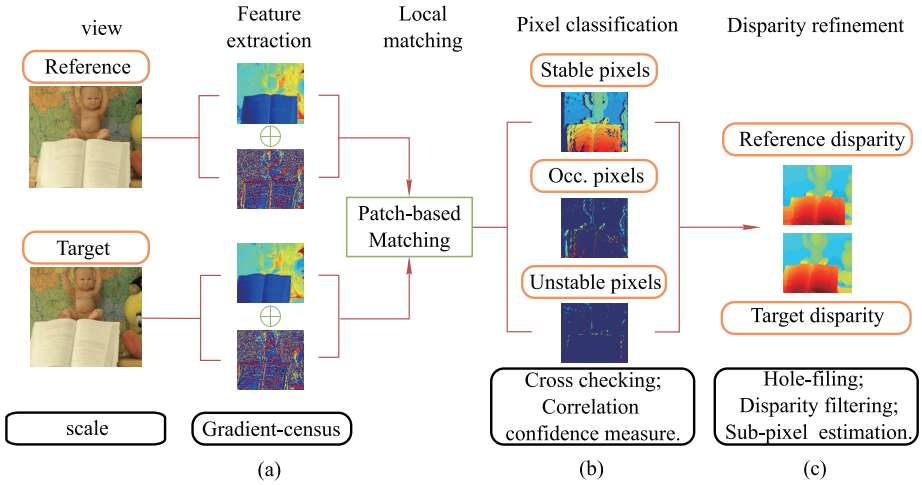


Fig. 4 Framework of the proposed patch-based stereo matching method

$$C_{census}(P,d) = \frac{\sum_{q_L \in P_L} \omega(p_L, q_L)\text{Ham}(\boldsymbol{F}_c(q_L), \boldsymbol{F}_c(q_{R,d}))}{\sum_{q_L \in P_L} \omega(p_L, q_L)},$$
$$\omega(p_L, q_L) = \exp\left(-\frac{\|I_c(p_L) - I_c(q_L)\|_1}{\lambda_c}\right), \quad (3)$$

where $q_{R,d}$ is the correspondence pixel of q_L, $\omega(p_L, q_L)$ is a weight function to overcome the edge-fattening problem, and I_c denotes the image in RGB space, p_L is the center pixel of the patch P_L.

Typical stereo matching methods combine the two features with a constant weight and adopt a classical winner-takes-all (WTA) strategy that selects disparity with the lowest cost. However, the result of the strategy depends on the robustness of the measures, i.e., the selected disparity is error-prone when several cost values are close to the minimum, and the true disparity may not locates at the position with the minimum cost (see Fig. 5(c)). The inaccurate measurement of one feature can corrupt the other if we simply add the two features with a fixed weight. Alternatively, we proposed an novel combination method using an optimal weight. The matching cost $C(P,d)$ between the two patches P_L and $P_{R,d}$ is combined as follows:

$$C(P,d) = \alpha\rho(C_{gradient}(P,d), c_{gradient}) + (1-\alpha)\rho(C_{census}(P,d), c_{census}),$$
$$\rho(x,c) = 1 - exp(-\frac{x}{c}),$$
$$\alpha = exp(-\frac{Var_{census}/Var_{gradient}}{c_{Var}}), \quad (4)$$

where $c_{gradient} = 30$ and $c_{census} = 5$ are two constants, α is the optimal weight between $C_{gradient}$ and C_{census}, $Var_{gradient}$ is variance of the four minimum values of $C_{gradient}(P,d)$, Var_{census} is variance of the four minimum values of $C_{census}(P,d)$, and $c_{Var} = 1.4427$. By measuring the variance of the costs of these two features, the optimal weight α endows the higher efficient feature with a more proportion. Figure 5 shows the result with and without the proposed optimal weight α. We select disparity with the minimum cost as the initial disparity result, denoted as D_L for the left view and D_R for the right view.

4.1.2 Pixel classification

By implementing the cross checking and the correlation confidence measure [28], the pixels in the initial disparity map are classified into three components: stable pixels, occlusion pixels and unstable pixels (as shown in Fig. 4(b)). For cross checking, the right disparity map D_R is re-projected to the left image, denoted as D_{R2L}. Finally, a pixel is declared unocclusion if the following relation holds:

$$|D_L(p) - D_{R2L}(p)| < 1. \quad (5)$$

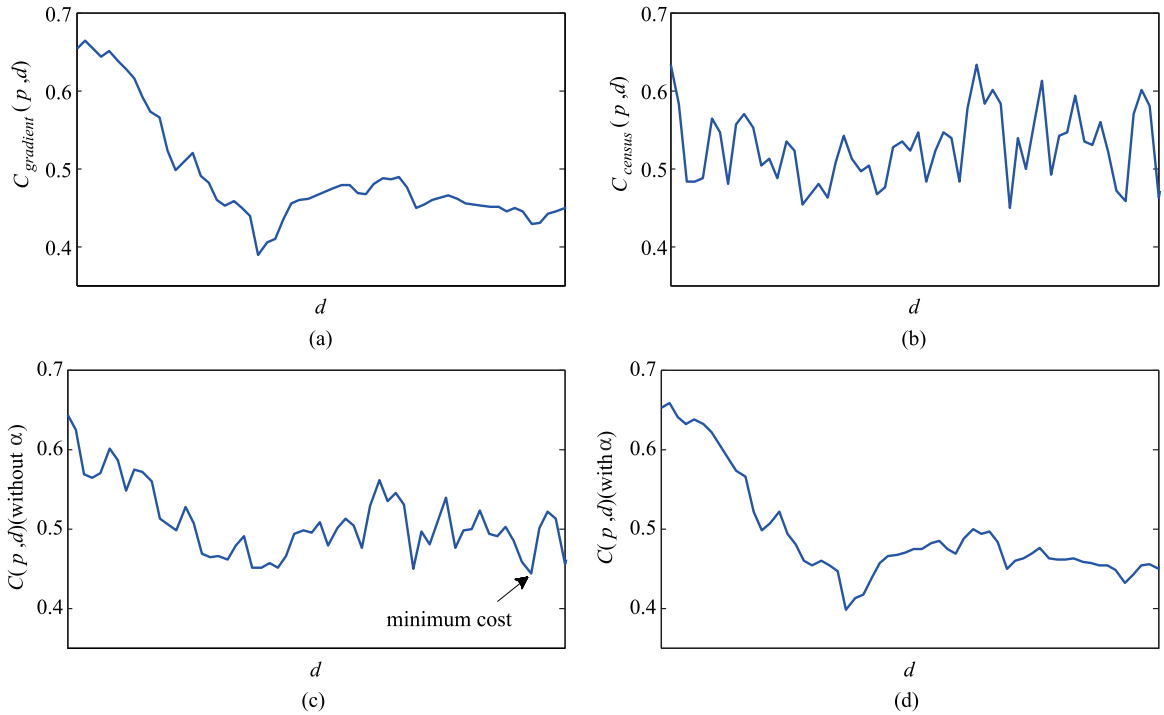


Fig. 5 WTA strategy without and with the optimal weight α. (a) $C_{gradient}$; (b) C_{census}; (c) final matching cost $C(P,d)$ without the optimal weight α; (d) final matching cost $C(P,d)$ with the optimal weight α. The gradient feature provides reliable matching costs, while census feature fails to. If we combine the costs evenly, the WTA strategy will fails to select the true disparity. However, the combination with α can provide a clear minimum value

If the relation does not hold, the pixel is declared as occlusion. For correlation confidence measure, the minimum and the second minimum cost values of a pixel p are selected, being denoted as $C_1(p)$ and $C_2(p)$, respectively. Then the correlation confidence V_{Conf} is defined as:

$$V_{Conf} = |\frac{C_1(p) - C_2(p)}{C_2(p)}|. \tag{6}$$

If the confidence V_{Conf} is above a threshold $b = 0.04$, the pixel is declared as stable, otherwise unstable. For a stable pixel, we take the disparity value with the minimum cost between the left disparity map and the re-projected right disparity map as the final disparity in this procedure.

4.1.3 Disparity refinement

After pixel classification, the disparity map contains holes caused by occlusion pixels and the unstable pixels. Besides, the quantized disparity by performing WTA strategy contains large discontinuities and errors. This block focuses on disparity refinement including hole-filling, disparity filtering and sub-pixel estimation, as shown in Fig. **??**(c).

Hole-filling A bilateral filter is applied to the cost volume in hole pixel based on the following assumptions: 1) The pixels with similar colors around a region are more likely to have similar disparity; 2) For occlusion areas, the pixels are always located at background with small disparity.

The filter is designed as follows:

$$\begin{aligned} Occlusion: F_O(p,q) &= f_c(p,q) f_s(p,q) f_d(q), \\ Unstable: F_U(p,q) &= f_c(p,q) f_s(p,q), \\ f_c(p,q) &= \exp\left(-\frac{\|I_c(p) - I_c(q)\|_1}{\lambda_c}\right), \\ f_s(p,q) &= \exp\left(-\frac{\|p-q\|_F}{\lambda_s}\right), \\ f_d(q) &= \exp\left(-\frac{|D(p') - D_{\min}|}{\lambda_d}\right), \end{aligned} \tag{7}$$

where p is the current hole pixel centered in a window with radius r_B, $q \in N(p)$ is the valid neighbor pixels of p, and u, I_c denotes the RGB image, D is the disparity map, and $D_{\min}$ is the minimum valid disparity in the current patch. λ_c and λ_s are two constants used as thresholds of the color and distance difference. $\lambda_d = 0.5 D_{\min}$. Traditional bilateral filter is consisted of f_c and f_s as color term and distance term, respectively. For occlusion pixels, an additional term f_d is added to the filter based on the second prior assumption, i.e., the costs corresponding to a pixel that holds small disparity in the filter window are endowed with large weight. After the cost volume filtering for unstable and occluded pixels, the WTA strategy is used to select the final disparity values.

Weighted median filter As an extension of bilateral filter, weighted median filter is widely applied to stereo matching [31,39]. Weighted median filter replaces the current pixel x with the weighted median of the neighbor pixels $q \in N(p)$ within a window (usually a box) by accumulating a weighted histogram in the pixel:

$$\begin{aligned} h(p,i) &= \sum_{q \in N(p)} \omega(p,q) \delta(D(q), i), \\ \omega(p,q) &= exp(-\frac{\|I_c(p) - I_c(q)\|_F^2}{2\sigma_w^2}), \end{aligned} \tag{8}$$

where $\omega(p,q)$ is the weight depending on the RGB image I_c, D is the disparity map obtained in the previous step, i is the discrete bin index, $\delta(\cdot)$ is the Kronecker delta function: $\delta(\cdot) = 1$ when the argument is 0 and $\delta(\cdot) = 0$ otherwise, and σ_w is an intrinsic parameters of the filter. To accelerate the execution, a fast weighted median filter is applied to our work [40]. The weighted median filter uses a joint-histogram representation, median tracking, and a new data structure that enables fast data access, reducing computation complexity from $O(r_w^2)$ to $O(r_w)$ (r_w is the the window radius).

Sub-pixel estimation In this procedure, a sub-pixel estimation algorithm based on quadratic polynomial interpolation is performed to reduce the discontinuities caused by quantization disparity selection [28]. First, for pixel p, we select disparity candidates d_p, d_p^- and d_p^+, where d_p is the initial disparity in this step, $d_p^- = d_p - 1$ and $d_p^+ = d_p + 1$. Then the interpolated disparity $\hat{d}_p$ is estimated as follows:

$$\hat{d}_p = d_p - \frac{C(p, d_p^+) - C(p, d_p^-)}{2(C(p, d_p^+) + C(p, d_p^-) - 2C(p, d_p))}. \tag{9}$$

Finally, a box-car filter (F_B) is applied to the interpolated result:

$$F_B(p,q) = \begin{pmatrix} 1, & |\hat{d}(q) - \hat{d}(p)| < 1, \\ 0, & \text{else}, \end{pmatrix} \tag{10}$$

where $q \in N(p)$ is the neighbor pixels of p within a box window (window radius $r_B = 4$ by default).

4.2 Disparity maps fusion

In this procedure (Fig. 3(c)), disparity maps $D_{SL}^{(i)}$ and $D_{SR}^{(i)}$ obtained by the stereo matching method described above are applied to produce disparity maps $D_{LR}^{(i)}$ and $D_{RL}^{(i)}$, which are located at positions of the input views. We generate them by

adopting the following equation:

$$
\begin{aligned}
D^{(i)}_{LR}(x_l, y) &= \begin{cases} -D^{(i)}_{SL}(x,y) + D^{(i)}_{SR}(x,y), & \text{else,} \\ 0, |D^{(i)}_{SL}(x,y) + D^{(i)}_{SR}(x,y)| > D_m(x,y), \end{cases} \\
D^{(i)}_{RL}(x_r, y) &= \begin{cases} D^{(i)}_{SL}(x,y) - D^{(i)}_{SR}(x,y), & \text{else,} \\ 0, |D^{(i)}_{SL}(x,y) + D^{(i)}_{SR}(x,y)| > D_m(x,y), \end{cases} \\
x_l &= x - D^{(i)}_{SL}(x,y) \\
x_r &= x - D^{(i)}_{SR}(x,y) \\
D_m(x,y) &= \tau^{(i)} \cdot \max(-D^{(i)}_{SL}(x,y), D^{(i)}_{SR}(x,y)),
\end{aligned} \tag{11}
$$

where $-D^{(i)}_{SL} + D^{(i)}_{SR}$ stands for the fused disparity map at the position μ. It should be noted that $D^{(i)}_{SL}$ is considered to have an opposite disparity values with $D^{(i)}_{SR}$ due to a reversed stereo matching. To obtain the disparity maps $D^{(i)}_{LR}$ and $D^{(i)}_{RL}$ at positions of the input views, we simply warp $-D^{(i)}_{SL} + D^{(i)}_{SR}$ to the left view and right view, respectively. Consider a pixel in $-D^{(i)}_{SL}+D^{(i)}_{SR}$ located at (x, y), to regain the corresponding pixel in $D^{(i)}_{LR}$ located at (x_l, y), we re-project the pixel back to $I^{(i)}_L$ by a horizontal displacement $-D^{(i)}_{SL}(x, y)$. The pixel's value $-D^{(i)}_{SL}(x, y) + D^{(i)}_{SR}(x, y)$ stands for the disparity from $I^{(i)}_L$ to $I^{(i)}_R$ at (x_l, y). The disparity map $D^{(i)}_{RL}$ is regained analogously. In the initial iterations, our virtual view is coarse, i.e., contains pixel position error comparing with the real image at position μ. However, this error can be counteracted effectively after the disparity maps fusion. Assume a pixel at (x, y) in $I^{(i)}_S$, the estimated disparity to $I^{(i)}_L$ is d_{SL} and the true disparity is $d^{(gt)}_{SL}$, $d_{SL} = d^{(gt)}_{SL} + Error_1$, and the estimated disparity from $I^{(i)}_S$ to $I^{(i)}_R$ is d_{SR} and the true disparity is $d^{(gt)}_{SR}$, $d_{SR} = d^{(gt)}_{SR} + Error_2$.

$$
\begin{aligned}
d_{LR} &= -d_{SL} + d_{SR} && (12) \\
&= -(d^{(gt)}_{SL} + Error_1) + (d^{(gt)}_{SR} + Error_2). && (13)
\end{aligned}
$$

When the errors are small and the correspondences between $I^{(i)}_S(x, y)$ and $I^{(i)}_L(x_r, y)$, $I^{(i)}_S(x, y)$ and $I^{(i)}_R(x_t, y)$ are found precisely, $Error_1 - Error_2 = 0$, and the final disparity $d_{LR} = -d^{(gt)}_{SL} + d^{(gt)}_{SR}$. On the other hand, when the errors are too big, i.e., the pixel holds the relation $||D^{(i)}_{SL}(x, y)| - |D^{(i)}_{SR}(x, y)|| > D_m(x, y)$, the disparity at this location will be set invalid. The threshold $\tau^{(i)}$ controls this error tolerance degree. Figure 6 illustrates the mechanism of the strategy. For the initial iteration, the threshold is set as a large value to tolerate the interpolation error. With the increasing of iteration times, we gradually tighten the threshold to detect more disparity outliers. For invalid pixel (whose error is beyond the threshold), we adopt a similar strategy in Eq. (7).

4.3 Disparity maps upsampling

The resolution of the fused disparity maps are restricted by the interpolated view $I^{(i)}_S$. To render a higher resolution virtual view $I^{(i+1)}_S$, we upsample the disparity maps $D^{(i)}_{LR}$ and $D^{(i)}_{RL}$ (Fig. 3(d)) by employing the spatial-depth super resolution approach proposed by Yang et al. [37]. The upsampling factor in every iteration of our iterative novel view refinement model is 2×. The disparity maps after performing super resolution are denoted as $D'^{(i)}$. It should be noticed that this procedure is ignored if the disparity maps already hold the same resolution with the original RGB image.

4.4 Disparity-assisted plane sweep-based rendering (DAPSR)

The proposed virtual view rendering method is used for both the initial interpolation (Fig. 3(a)) and the virtual view

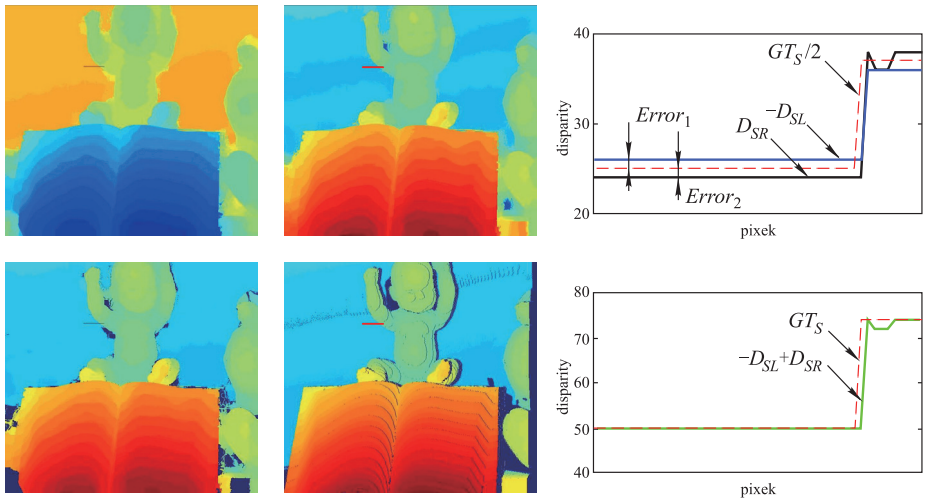


Fig. 6 Mechanism of the proposed disparity maps fusion strategy. Top left shows the disparity map D_{SL}; Top middle shows the disparity map D_{SR}; Bottom left shows the disparity map $-D_{SL} + D_{SR}$; Bottom middle shows the warped ground truth disparity map GT_S. The extracted disparity data being located at the red line in the disparity maps explain the error counteraction mechanism. Small errors, like $Error_1$ and $Error_2$, will be neutralized during the fusion. While the pixels with big error are set invalid

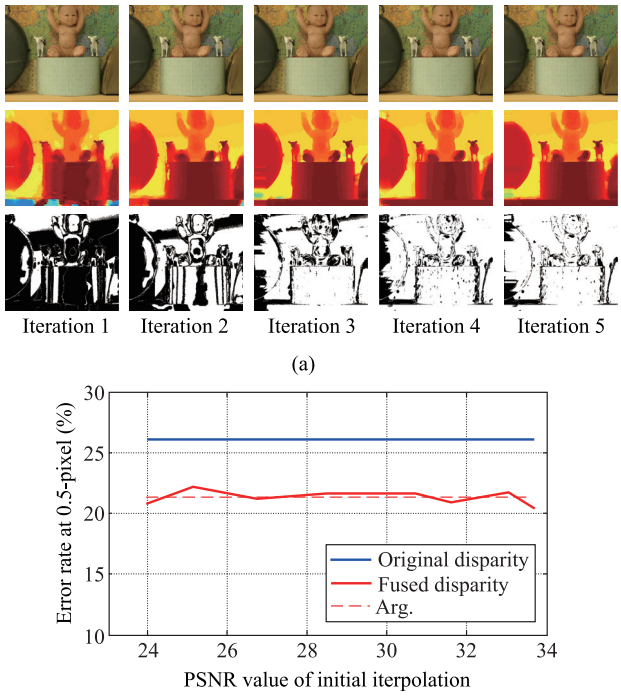


Fig. 7 Illustration of robustness of the proposed disparity maps fusion strategy. (a) Interpolation result and the fused disparity map versus iteration times; (b) initial interpolation quality versus stereo quality of output disparity map

updating (Fig. 3(e)). The key idea for the proposed DAPSR is that for each candidate disparity, the input views are swept to the virtual viewpoint to interpolate a set of candidate images, then the virtual view is the soft blending of the candidate images with the assistance of disparity maps. For disparity errors within one-pixel range, an plausible interpolation result can still be produced by using linear interpolation [41]. Therefore, the interpolation of candidate images in the DAPSR is robust to small disparity errors. Due to the initial interpolation is always implemented at a small resolution, the robustness to small disparity error is crucial to the entire model.

Specifically, for the initial interpolation, the two RGB image pair $I_L^{(0)}$ and $I_R^{(0)}$ (downsampled to the desired resolution) and the corresponding disparity map pair $D_L^{(0)}$ and $D_R^{(0)}$ are employed. For each d, $I_L^{(0)}$ and $I_R^{(0)}$ are swept to the virtual viewpoint μ by shift $-\mu d$ and $(1-\mu)d$, respectively, where $\mu \in [0,1]$ is considered as the normalized distance from the virtual view $I_S^{(0)}$ to $I_L^{(0)}$ (with $\mu = 0$ the position of left view $I_L^{(0)}$, and $\mu = 1$ the position of right view $I_R^{(0)}$). The resulting images are denoted as $I_{d,L}^{(0)}$ and $I_{d,R}^{(0)}$, and applied to interpolate the candidate images using linear interpolation $I_{d,S}^{(0)} = (1-\mu)I_{d,L}^{(0)} + \mu I_{d,R}^{(0)}$. At the same time, the input disparity maps $D_L^{(0)}$ and $D_R^{(0)}$ and images $I_L^{(0)}$ and $I_R^{(0)}$ are warped to the virtual view, and the resulting maps (images) are denoted as $D_{\mu,L}^{(0)}$, $D_{\mu,R}^{(0)}$, $I_{\mu,L}^{(0)}$ and $I_{\mu,R}^{(0)}$, respectively. The warp operation produces invalid pixels in the occluded regions. We then use the valid pixels in the disparity maps to produce a blended disparity $D_b^{(0)} = (1-\mu)D_{\mu,L}^{(0)} + \mu D_{\mu,R}^{(0)}$. The initial virtual view is computed as follows:

$$
\begin{aligned}
I_S'^{(0)}(p) &= \frac{\sum\limits_{d \in D_b^{(0)}} \omega_d(p) I_{d,S}^{(i)}(p)}{\sum\limits_{d \in D_b^{(0)}} \omega_d(p)}, \\
\omega_d(p) &= \begin{cases} \dfrac{2 - |D_b^{(0)}(p) - d|}{2}, & |D_b^{(0)}(p) - d| \leqslant 1, \\ 0, & else, \end{cases}
\end{aligned}
\tag{14}
$$

where p is a pixel. The equation above describes the blending of pixels that are visible in both input views. For occluded regions, the warped images $I_{\mu,L}^{(0)}$ and $I_{\mu,R}^{(0)}$ are used for rendering the final initial virtual view $I_S^{(0)}$:

$$
I_S^{(0)}(p) = I_S'^{(0)}(p)\delta_S(p) + I_{\mu,L}^{(0)}(p)\delta_L(p) + I_{\mu,R}^{(0)}(p)\delta_R(p), \tag{15}
$$

where $\delta_S(p) = 1$ if $I_S'^{(0)}(p)$ is valid in the non-occluded regions, $\delta_L(p) = 1$ if $I_{\mu,L}^{(0)}(p)$ is valid in the occluded regions, and is 0 otherwise. $\delta_R(p)$ is defined analogously. Finally, a 5×5 median filter is employed to fill the holes.

To update the virtual view $I_S^{(i+1)}$ in the following iterations, in addition to the RGB image pair $I_L^{(}i)$ and $I_R^{(}i)$ (downsampled to the desired resolution) and their corresponding disparity maps $D_{LR}'^{(i)}$ and $D_{RL}'^{(i)}$, the virtual view in the previous iteration $I_S^{(i)}$ (upsampled to the desired resolution) is also employed. The rendering procedure is the same as that for the initial interpolation. The only difference is that we use $I_S^{(i)}$ to fill the holes in stead of the median filter.

By employing the proposed strategies of disparity maps fusion and DAPSR, the quality of the output disparity maps as well as interpolated view can be greatly improved with the iteration, as shown in Fig. 7(a). To further demonstrate the robustness of the strategies, we deliberately interfere the initial interpolation quality by adding noise to the initial disparity map. Figure 7(b) shows initial interpolation quality versus stereo quality of an output disparity map. The PSNR value of the initial interpolation varies from 24 to 34, yet the final output disparity map get little influence.

5 Experimental results

In this section the Middleburry datasets is applied to evaluate the proposed model. The input views are *view*1 and *view*5 of half size version in Middleburry 2.0 [42] (*Tsukuba*, *Venus*,

Teddy and *Cones* cases are the default resolution in the evaluation v.2), and *im*0 and *im*1 of quarter size version in Middleburry 3.0 [43]. Besides the ground truth disparity maps, the Middleburry 2.0 [42] also has the ground truth middle views, providing a reference to evaluate the synthesized view. The evaluation is performed on two aspects: interpolation result and stereo result. The datasets with varying illumination and exposure conditions are also employed to demonstrate the robustness of the proposed framework. For the occlusion pixels we set $\lambda_s = 2r_B$, $r_B = 20$, and for the unstable pixels we set $\lambda_s = r_B$, $r_B = 10$. The iteration times in the model is 5, and $\tau^{(i)} = [0.5, 0.4, 0.3, 0.2, 0.2]$. The interpolation position $\mu = 0.5$ as default, corresponding to *view*3. The rest parameters applied to the model are given in Table 1, where r_W and σ_W are two parameters being applied to the weighted median filter. The employed methods include Classic+NLP (global) [9], CostFilter (local) [7] and Adaptive Random Walk (ARW, global) [8]. The computational complexity of the proposed method is $O(kNL)$, where k is the iteration times, N is the number of pixels in the image and L is the number of disparity labels. The average running time of the proposed method on a 0.35Mpix image is about 60 seconds in Matlab without GPU acceleration, whereas the stereo matching is the most time consuming step.

Table 1 Parameters settings

$c_{gradient}$	c_{census}	c_{Var}	λ_c
40	5	1.4427	10
λ_s	r_W	σ_W	μ
15	5	15.5	0.5

5.1 Interpolation result

In this subsection, both the interpolated initial virtual view and refined virtual view are evaluated using PSNR metric. In the stereo matching block of our model, we also employ other methods to show the universal applicability.

We use *view*3 in the datasets as the ground truth to evaluate the interpolation result. Figure 8 provides parts of the interpolation result with and the model. Some failed regions are zoomed in to present a clearer comparison. The proposed model produces a better interpolation result, and the CostFilter also achieves a better performance after being refined by the model, for example the flank and the arm of the baby in *Baby2*; the head of the bowling pin and the edge of the ball in *Bowling2*; the head of the doll and the ear of the bear in *Dolls*; and the surrounding of the reindeer's neck in *Reindeer*. Table

Fig. 8 Comparisons of interpolation result on *Baby2, Bowling2, Dolls and Reindeer datasets (a) Ground truth; (b) CostFilter; (c) CostFilter (refined); (d) ours. PSNR values are given at left side of the image. Red boxes highlight the failed regions, one of which is zoomed in*

Table 2 PSNR evaluation of the initial interpolations and the refined results

	Initial	Classic+NLP	CostFilter	ARW	Classic+NLP (R)	CostFilter (R)	ARW (R)	Proposed
Baby2	33.11	29.60	32.55	31.35	31.90	33.66	32.35	**33.94**
Baby3	33.00	29.42	30.41	30.36	31.47	33.55	31.20	**33.71**
Bowling2	28.13	25.09	30.34	29.22	30.66	33.24	31.11	**34.25**
Dolls	29.01	24.01	28.33	28.10	28.69	30.49	30.22	**30.56**
Reindeer	28.07	20.71	27.12	25.27	26.31	**32.03**	27.37	31.78
Teddy	31.59	30.25	29.96	30.16	32.81	30.63	31.66	**32.87**
Cones	28.44	23.25	28.14	27.31	28.89	29.12	27.86	**30.44**

2 presents a detailed interpolation result in terms of PSNR and SSIM, where *(R)* in the table indicate the refined result by using our proposed model. The interpolation result using the proposed model surpass the original result in all cases due to the better stereo quality. And the proposed stereo matching method outperforms other approach except the *Reindeer* case, in which the refined CostFilter [7] provides the best interpolation results. Table 2 also lists initial interpolation results (ground truth views are downsampled for comparison) to provide reference comparison between initial interpolations and their refined versions. The results show that the virtual views are improved when using the proposed model.

5.2 Stereo result

We evaluate the proposed model using the patch-based stereo matching and other approaches as the stereo kernel on both Middlebury 2.0 [42] and 3.0 benchmarks [43]. The evaluation is performed at 0.5-pixel threshold. Table 3 provides the evaluation results on the Middlebury 2.0 benchmarks [42], *Tsukuba*, *Venus*, *Teddy* and *Cones*, and the resulting disparity maps are shown in Fig. 9. The average rank of the proposed stereo matching approach is 35.2, ranking at 2nd place among the local methods. Besides, the average rank of the refined CostFilter is promoted to 35.8 from the original 48.0.

In addition, to further evaluate two strategies in the proposed framework, we perform a number of ablation studies by replacing the disparity maps fusion with a naïve adding of the disparity maps, denoted as *Ours (w/o DF)* for short, and by replacing the DAPSR with a simple DIBR method, denoted as *Ours (w/o DAPSR)*, respectively. The results in Table 3 show that the proposed strategies, disparity maps fusion and DAPSR, are important to the entire framework.

We demonstrate some evaluation result using Middlebury datasets 2005, 2006 [42] and 2014 [43] in Fig. 10. The figure shows the result produced by the original stereo matching approach (Classic+NLP [9] for the *Baby3*, ARW [8] for the *Motorcycle* case and CostFilter [7] for the rest cases), the refined version using our model and the proposed stereo matching method. The stereo quality in discontinuous and occlusion region is greatly enhanced after introducing the synthesized virtual view, such as the background between the left foot and the cow in *Baby3*, and the bench in *Adiron*. The proposed model using patch-based stereo matching also produces disparity maps that are more respectful to the ground truth scene structure. Bad pixel rates at 0.5-pixel error thresholds in all regions within the image of each method are presented in Table 4. The quantized result shows that the stereo matching approaches are improved after being applied to the model.

Table 3 Bad pixel rates evaluated at 0.5-pixel error threshold on Middlebury evaluation 2.0 [42]

	Avg.	Tsukuba			Venus			Teddy			Cones		
	Rank	nonocc.	all	disc.	nonocc.	all	disc.	nonocc.	all	disc.	nonocc.	all	disc.
Our approach	35.2	11.8	12.1	21.9	1.62	2.07	8.32	9.44	16.5	23.8	5.07	11.5	12.6
CostFilter (R)	35.8	9.42	11.4	13.4	2.33	2.82	10.4	10.9	17.0	27.8	6.69	12.3	17.2
TSGO [31]	37.8	8.78	9.45	14.9	0.72	1.12	5.24	10.1	16.4	21.3	8.49	14.7	16.5
CostFilter [7]	48.0	11.2	11.7	15.6	5.99	6.43	10.8	11.3	18.1	25.3	7.71	13.7	15.1
AdaptingBP [34]	52.0	19.1	19.3	17.4	4.84	5.08	7.84	12.8	16.7	26.3	7.02	13.2	14.0
Ours (w/o DAPSR)	64.0	12.6	13.1	23.5	2.69	3.53	9.39	14.7	17.8	24.6	10.2	16.5	17.3
Classic+NLP (R)	77.3	9.04	9.48	24.6	2.70	3.22	12.5	13.6	21.7	33.8	10.0	16.7	25.1
Ours (w/o DF)	77.5	14.1	15.4	26.6	3.07	3.92	10.7	16.0	18.9	25.7	12.1	18.3	19.6
Classic+NLP [9]	102.2	14.4	14.9	27.1	2.01	2.83	16.3	14.0	21.9	34.4	24.0	29.6	29.0
ARW (R)	104.4	14.0	15.2	31.5	5.24	6.49	28.4	15.5	23.3	35.5	10.9	18.0	25.0
ARW [8]	134.9	18.2	20.0	34.5	10.5	11.9	37.3	18.0	25.9	38.8	12.8	21.2	29.6

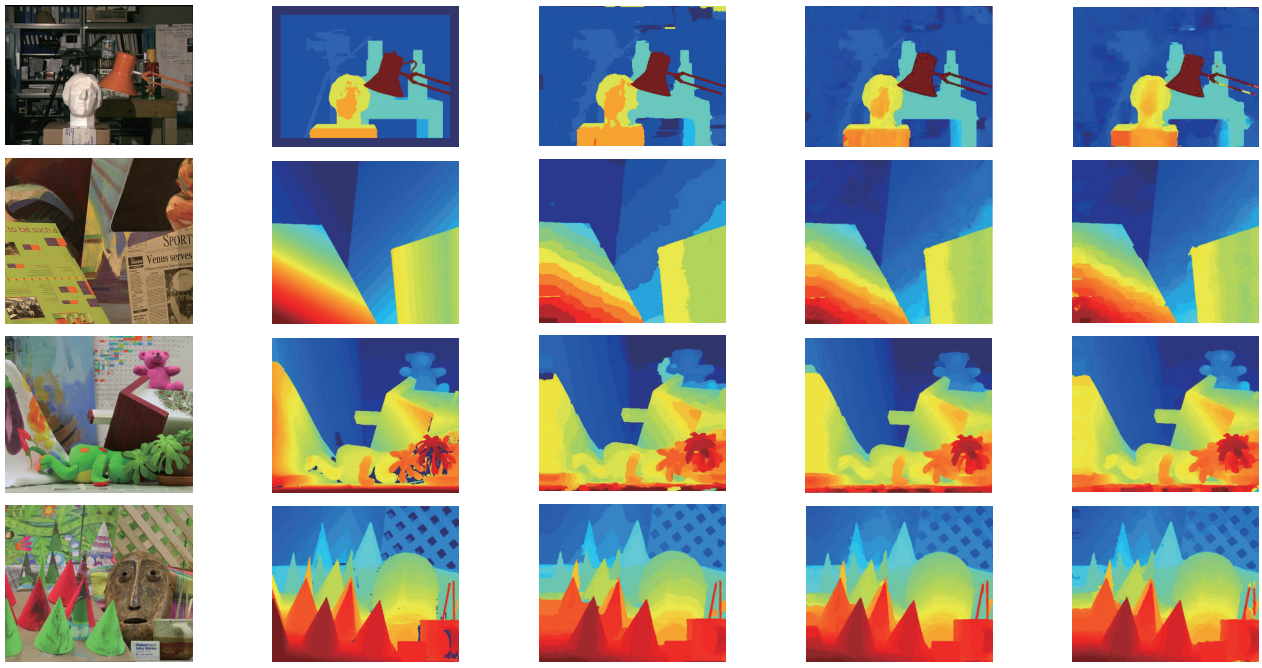


Fig. 9 Stereo result on Middlebury 2.0 benchmarks *Middle2.0, Tsukuba, Venus, Teddy and Cones (a) Left image; (b) ground truth; (c) CostFilter; (d) CostFilter (refined); (e) ours*

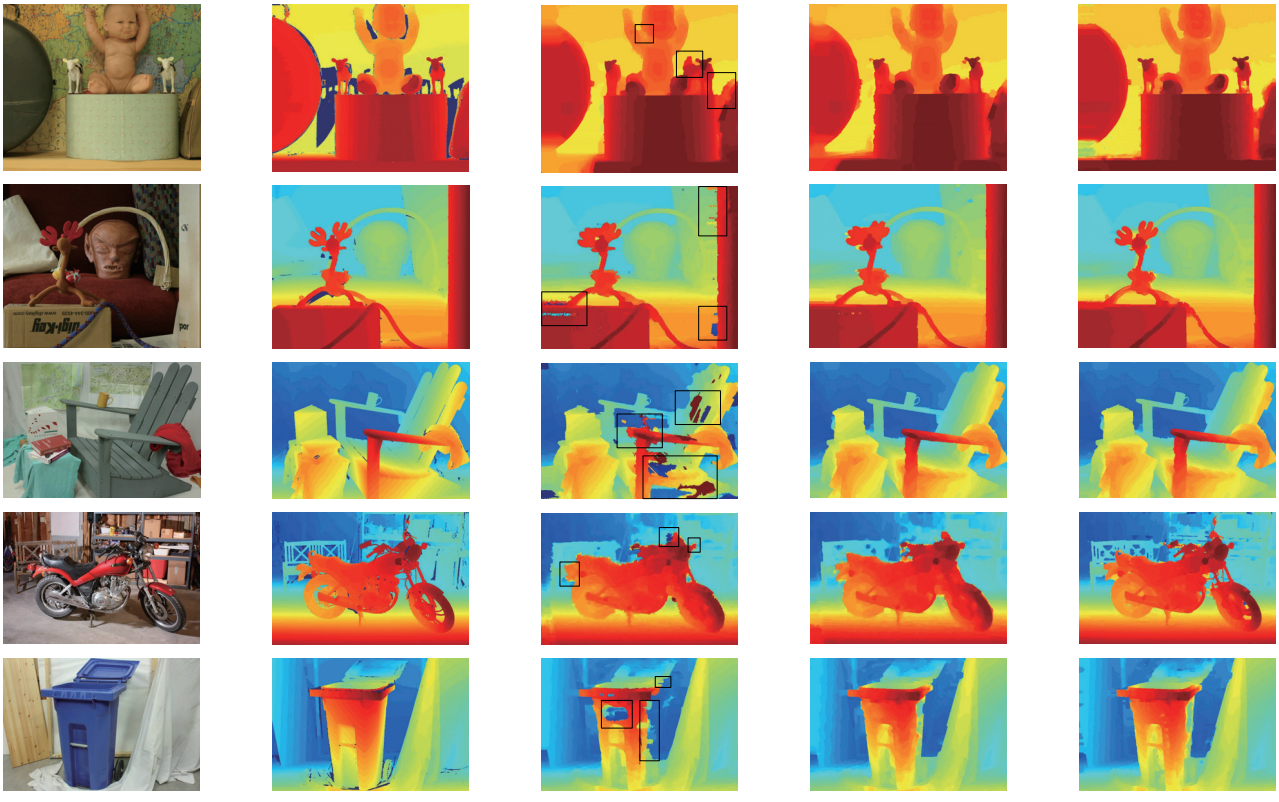


Fig. 10 Stereo result on Middlebury 2.0 *Middle2.0 and Middlebury 3.0 Middle3.0 benchmarks (Baby3, Reindeer, Adirondack, Motorcycle and Recycle) (a) Left image; (b) ground truth; (c) other approaches: Classic+NLP for the Baby3 case , ARW for the and Motorcycle case and CostFilter for the rest. The failed regions are highlighted by black boxes; (d) the refined version; (e) ours*

Table 4 Bad pixel rates evaluated at 0.5-pixel error threshold on Middleburry datasets [42, 43]

	Classic+NLP	CostFilter	ARW	Classic+NLP (R)	CostFilter (R)	ARW (R)	Proposed
Baby2	29.99	32.06	22.44	24.82	18.62	20.43	**18.46**
Baby3	17.12	18.43	21.30	16.73	15.64	19.46	**12.69**
Bowling2	31.83	27.66	24.30	25.80	21.80	22.40	**21.48**
Dolls	33.39	33.12	29.27	27.66	29.10	28.90	**22.60**
Reindeer	43.22	25.84	28.85	28.79	20.52	27.72	**16.59**
Teddy	21.91	17.72	25.90	21.62	17.03	23.31	**16.50**
Cones	29.57	12.65	24.50	16.70	12.27	21.33	**11.54**
Adiron.	35.70	38.29	36.70	28.51	25.23	34.48	**16.45**
Motor.	43.25	30.46	40.17	29.61	27.26	38.93	**17.37**
Pipes	46.79	36.48	43.38	38.13	32.13	42.05	**26.77**
Recycle	31.61	30.22	35.54	29.98	25.12	34.83	**18.11**

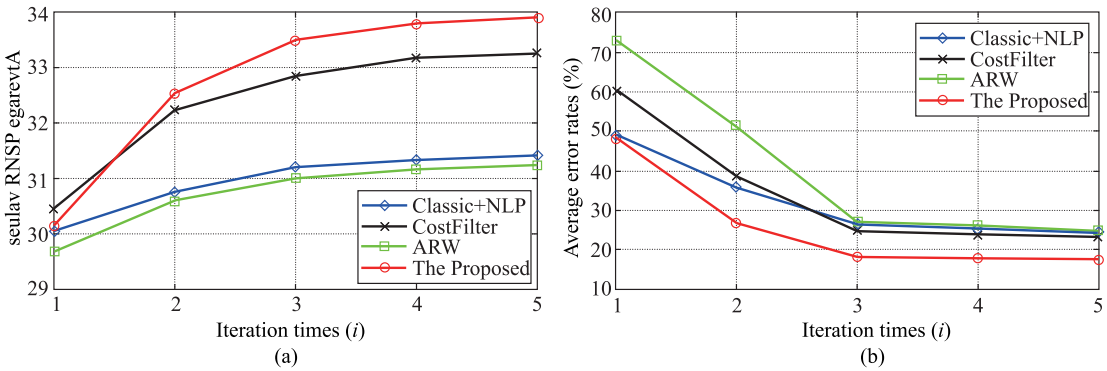


Fig. 11 Interpolation and stereo results yielded by each method versus the iteration time: (a) the average PSNR values of interpolation, and (b) the average error rates at 0.5-pixel threshold of stereo

In addition, we show the quality of the virtual views and the stereo results against iteration time in Fig. 11 (results are averaged on all the cases). The initial virtual view $I_S^{(0)}$ is 180 pixels in image width and is updated to the full resolution (same as the input image) at the 3rd iteration. In the last two iterations, the quality of the virtual views and the stereo results are tending to be stable.

5.3 Robustness testing

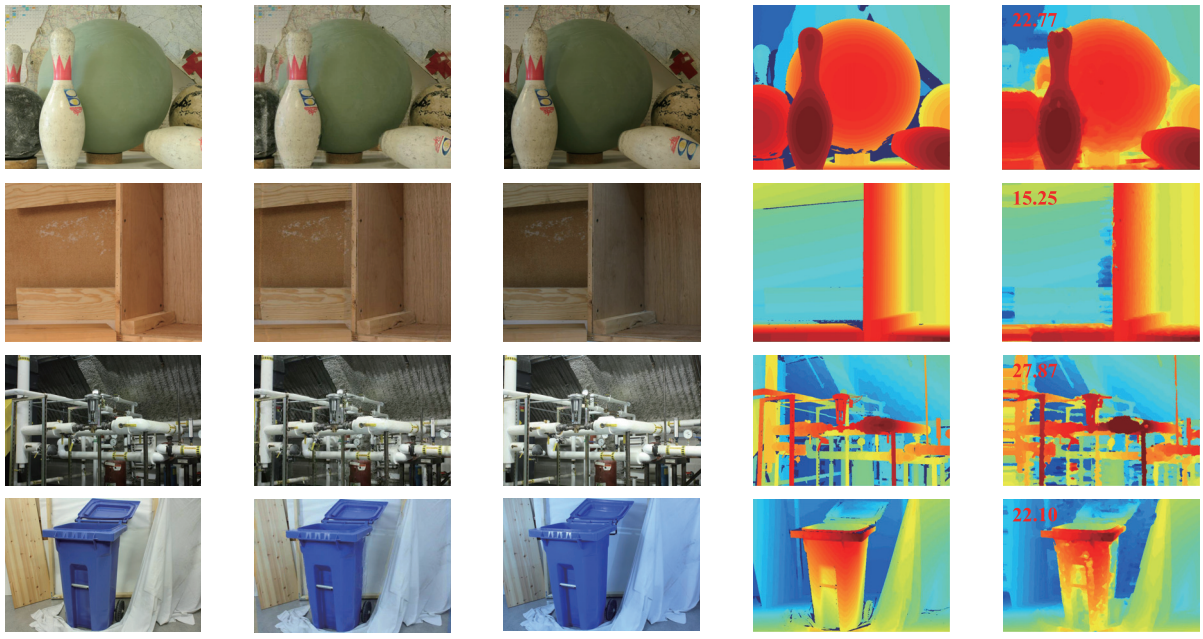
In the real world environment, the intensity value of corresponding pixels of the captured left and right views can be distort by illumination and camera exposure fluctuations. In this section, we test the proposed model using datasets [42, 43] under different illuminations or exposures to further demonstrate the robustness. The results are shown in Fig. 12, where the quantitative evaluation results at 0.5-pixel error threshold are given at top left of the disparity maps. The experiment results show that the proposed patch-based stereo matching method is robust to radiometric variations.

5.4 Limitations

The proposed framework is intended to be designed for addressing large parallax between the input views by introducing a synthesized view. When the input stereo pair has a small baseline, e.g., adopting *view1* and *view2* in the Middleburry 2.0 datasets as the input stereo pair, it will be very difficult to show its advantage for the designed framework. Tabel 5 shows the results using *view1* and *view2* / *view3* as the input. The results show that the proposed framework fail to improve the quality of the resulting disparity maps. Therefore, an appropriate baseline between the views is one of the conditions for the convergence of our iterative refinement framework. Besides, the proposed stereo matching algorithm is a local method, thus, suffers from textureless regions and repetitive patterns as other local methods. Fortunately, when applying a global method as the stereo kernel, such as Classic+NLP [9] or ARW [8], our framework is still able to converge.

6 Discussions and conclusions

In this paper, we have presented a novel iterative refinement model for joint view synthesis and disparity refinement. The main contribution of the paper is developing a framework that combines virtual view refinement and stereo matching.



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

Fig. 12 Stereo result using images with varying illuminations or exposures. (*Bowling2*, *Wood1*, *Pipes*, *Recycle*): (a) left image; (b) our synthesized view; (c) right image; (d) ours stereo result (0.5-pixel) error rate is given at top left of each image; (e) Ground truth disparity map

Table 5 Failure cases using stereo pairs with small parallax at 0.5-pixel

	Views' index	Classic+NLP	Classic+NLP(R)
Flower pots	*view1-2*	10.71	11.76
	view1-3	17.60	18.10
Cloth1	*view1-2*	1.67	8.44
	view1-3	2.91	4.45
Rock2	*view1-2*	5.21	6.80
	view1-3	6.23	7.00

To realize the mutual promotion between the interpolated view and the disparity map, we have proposed a disparity maps fusion and a disparity-assisted plane sweep-based rendering strategy. The former strategy is designed to eliminate the disparity error from interpolation artifacts by performing error detection and interpolation. And the latter strategy focuses on interpolation robustness to the bad pixels in the disparity maps. We have demonstrated that the proposed model is able to generate a synthesized view with high visual coherency as well as high quality disparity maps. We also show the general applicability of the model by employing other stereo matching approaches to the model. In the future work, we will extend the idea to optical flow estimation and light field stereo. In addition, there is potential to apply a learning-based system for joint view synthesis and disparity estimation to replace our iterative refinement framework.

Acknowledgements This work was supported by the National key foundation for exploring scientific instrument (2013YQ140517), the National Natural Science Foundation of China (Grant No. 61522111) and the Shenzhen Peacock Plan (KQTD20140630115140843).

References

1. Bleyer M, Rother C, Kohli P. Surface stereo with soft segmentation. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. 2010, 1570–1577
2. Fickel G P, Jung C R, Malzbender T, Samadani R, Culbertson B. Stereo matching and view interpolation based on image domain triangulation. IEEE Transactions on Image Processing, 2013, 22(9): 3353–3365
3. Vogel C, Schindler K, Roth S. 3D scene flow estimation with a piecewise rigid scene model. International Journal of Computer Vision, 2015, 115(1): 1–28
4. Wu G, Masia B, Jarabo A, Zhang Y, Wang L, Dai Q, Chai T, Liu Y. Light field image processing: an overview. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(7): 926–954
5. Wu G, Liu Y, Fang L, Dai Q, Chai T. Light field reconstruction using deep convolutional network on EPI and extended applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, DOI:10.1109/TPAML.2018.2845393
6. Fehn C. Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV. In: Proceedings of Stereoscopic Displays and Virtual Reality Systems XI. 2004, 93–104
7. Hosni A, Rhemann C, Bleyer M, Rother C, Gelautz M. Fast cost-volume filtering for visual correspondence and beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(2): 504–511
8. Lee S, Lee J H, Lim J, Suh I H. Robust stereo matching using adaptive random walk with restart algorithm. Image and Vision Computing,

2015, 37(C): 1–11

9. Sun D, Roth S, Black M J. A quantitative analysis of current practices in optical flow estimation and the principles behind them. International Journal of Computer Vision, 2014, 106(2): 115–137
10. Kubota A, Smolic A, Magnor M, Tanimoto M, Chen T, Zhang C. Multiview imaging and 3dtv. IEEE signal processing magazine, 2007, 24(6): 10–21
11. Stefanoski N, Wang O, Lang M, Greisen P, Heinzle S, Smolic A. Automatic view synthesis by image-domain-warping. IEEE Transactions on Image Processing, 2013, 22(9): 3329–3341
12. Tian D, Lai P L, Lopez P, Gomila C. View synthesis techniques for 3D video. In: Proceedings of Internatlbnal Society for Optics and Phontonics, Applications of Digital Image Processing XXXII. 2009, 74430T
13. Mori Y, Fukushima N, Yendo T, Fujii T, Tanimoto M. View generation with 3D warping using depth information for FTV. Signal Processing: Image Communication, 2009, 24(1-2): 65–72
14. Wanner S, Goldluecke B. Variational light field analysis for disparity estimation and super-resolution. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(3): 606–619
15. Zhang L, Zhang Y H, Huang H. Efficient variational light field view synthesis for making stereoscopic 3D images. Computer Graphics Forum. 2015, 7(34): 183–191
16. Zhang C, Li Z, Cheng Y, Cai R, Chao H, Rui Y. Meshstereo: a global stereo model with mesh alignment regularization for view interpolation. In: Proceedings of the IEEE International Conference on Computer Vision. 2015, 2057–2065
17. Kalantari N K, Wang T C, Ramamoorthi R. Learning-based view synthesis for light field cameras. ACM Transactions on Graphics, 2016, 35(6): 193
18. Scharstein D, Szeliski R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision, 2002, 47(1-3): 7–42
19. Kordelas G A, Alexiadis D S, Daras P, Izquierdo E. Enhanced disparity estimation in stereo images. Image and Vision Computing, 2015, 35: 31–49
20. Peng Y, Li G, Wang R, Wang W. Stereo matching with space-constrained cost aggregation and segmentation-based disparity refinement. In: Proceedings of International Society for Optics and Phoritonics, Three-Dimensional Image Processing, Measurement, and Applications. 2015, 939309
21. Zbontar J, LeCun Y. Stereo matching by training a convolutional neural network to compare image patches. Journal of Machine Learning Research, 2016, 17(1): 2287–2318
22. Luo W, Schwing A G, Urtasun R. Efficient deep learning for stereo matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, 5695–5703
23. Mattoccia S, Viti M, Ries F. Near real-time fast bilateral stereo on the GPU. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2011, 136–143
24. Mattoccia S, Giardino S, Gambini A. Accurate and efficient cost aggregation strategy for stereo correspondence based on approximated joint bilateral filtering. In: Proceedings of Asian Conference on Computer Vision. 2009, 371–380
25. Yoon K J, Kweon I S. Adaptive support-weight approach for correspondence search. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(4): 650–656
26. Kolmogorov V, Zabih R. Computing visual correspondence with occlusions using graph cuts. In: Proceedings of the 8 th IEEE International Conference on Computer Vision. 2001, 508–515
27. Taniai T, Matsushita Y, Naemura T. Graph cut based continuous stereo matching using locally shared labels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014, 1613–1620
28. Yang Q, Wang L, Yang R, Stewénius H, Nistér D. Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(3): 492–504
29. Guney F, Geiger A. Displets: resolving stereo ambiguities using object knowledge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015, 4165–4175
30. Psota E T, Kowalczuk J, Mittek M, Perez L C. Map disparity estimation using hidden markov trees. In: Proceedings of the IEEE International Conference on Computer Vision. 2015, 2219–2227
31. Mozerov M G, Van d W J. Accurate stereo matching by two-step energy minimization. the IEEE Transactions on Image Processing, 2015, 24(3): 1153–1163
32. Mei X, Sun X, Zhou M, Jiao S, Wang H, Zhang X. On building an accurate stereo matching system on graphics hardware. In: Proceedings of the IEEE Conference on Computer Vision Workshops. 2011, 467–474
33. Veksler O. Fast variable window for stereo correspondence using integral images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2003, 556–561
34. Klaus A, Sormann M, Karner K. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In: Proceedings of the 18th International Conference on Pattern Recognition. 2006, 15–18
35. Vogel C, Schindler K, Roth S. 3D scene flow estimation with a rigid motion prior. In: Proceedings of the IEEE International Conference on Computer Vision. 2011, 1291–1298
36. Boykov Y, Veksler O, Zabih R. Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(11): 1222–1239
37. Yang Q, Yang R, Davis J, Nistér D. Spatial-depth super resolution for range images. In: Proceedings of THE IEEE Conference on Computer Vision and Pattern Recognition. 2007, 1–8
38. Boominathan V, Mitra K, Veeraraghavan A. Improving resolution and depth-of-field of light field cameras using a hybrid imaging system. In: Proceedings of THE IEEE International Conference on Computational Photography. 2014, 1–10
39. Ma Z, He K, Wei Y, Sun J, Wu E. Constant time weighted median filtering for stereo matching and beyond. In: Proceedings of THE IEEE International Conference on Computer Vision. 2013, 49–56
40. Zhang Q, Xu L, Jia J. 100+ times faster weighted median filter (WMF). In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014, 2830–2837
41. Lin Z, Shum H Y. A geometric analysis of light field rendering. International Journal of Computer vision 2004, 58(2): 121–138
42. Scharstein D, Szeliski R. High-accuracy stereo depth maps using structured light. In: Proceedings of the IEEE Conference on Computer Vi-

sion and Pattern Recognition. 2003, 195–202

43. Scharstein D, Hirschmüller H, Kitajima Y, Krathwohl G, Nešić N, Wang X, Westling P. High-resolution stereo datasets with subpixel-accurate ground truth. In: Proceedings of German Conference on Pattern Recognition. 2014, 31–42

Gaochang Wu received the B S and M S degrees in mechanical engineering from Northeastern University, Shenyang, China in 2013 and 2015, respectively. He is currently working toward Ph D degree in control theory and control engineering in Northeastern University, China. His current research interests include data mining, signal analysis, image processing and computational photography.

Yipeng Li received the B S and M S degrees in electronic engineering from the Harbin Institute of Technology, Harbin, China in 2003 and 2005, respectively, and the Ph D degree in electronic engineering from Tsinghua University, China in 2011. Since 2011, he has been a Research Associate with the Department of Automation, Tsinghua University. His current research interests include computer vision and data mining by using deep architecture networks.

Yuanhao Huang received the BS degree from the Peking University, China in 2002, and the PhD degree from the City University of Hong Kong, China, in 2009. He was a research fellow in the Singapore-MIT Alliance for Research and Technology (SMART), in 2011. He is the expert of the Thousand Talents Plan(China) , and the CEO of ORBBEC CO.,LTD. His research interests include optical measurement, computer vision, infrared technology.

Yebin Liu received the BE degree from Beijing University of Posts and Telecommunications, China in 2002, and the PhD degree from the Automation Department, Tsinghua University, China in 2009. He has been working as a research fellow at the computer graphics group of the Max Planck Institute for Informatik, Germany in 2010. He is currently an associate professor in Tsinghua University. His research areas include computer vision and computer graphics.