Sensors **2015**, *15*, 20894-20924; doi:10.3390/s150820894

OPEN ACCESS
sensors
ISSN 1424-8220
www.mdpi.com/journal/sensors

Article

Reliable Fusion of Stereo Matching and Depth Sensor for High Quality Dense Depth Maps

Jing Liu [1,2,*], Chunpeng Li [1], Xuefeng Fan [1,2] and Zhaoqi Wang [1]

[1] The Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, No.6 Kexueyuan South Road Zhongguancun, Haidian District, Beijing 100190, China; E-Mails: cpli@ict.ac.cn (C.L.); fanxuefeng@ict.ac.cn (X.F.); meifeng@ict.ac.cn (Z.W.)

[2] University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing 100049, China

* Author to whom correspondence should be addressed; E-Mail: liujing01@ict.ac.cn; Tel.: +86-10-6260-0874.

Academic Editor: Gonzalo Pajares Martinsanz

Received: 23 June 2015 / Accepted: 17 August 2015 / Published: 21 August 2015

Abstract: Depth estimation is a classical problem in computer vision, which typically relies on either a depth sensor or stereo matching alone. The depth sensor provides real-time estimates in repetitive and textureless regions where stereo matching is not effective. However, stereo matching can obtain more accurate results in rich texture regions and object boundaries where the depth sensor often fails. We fuse stereo matching and the depth sensor using their complementary characteristics to improve the depth estimation. Here, texture information is incorporated as a constraint to restrict the pixel's scope of potential disparities and to reduce noise in repetitive and textureless regions. Furthermore, a novel pseudo-two-layer model is used to represent the relationship between disparities in different pixels and segments. It is more robust to luminance variation by treating information obtained from a depth sensor as prior knowledge. Segmentation is viewed as a soft constraint to reduce ambiguities caused by under- or over-segmentation. Compared to the average error rate 3.27% of the previous state-of-the-art methods, our method provides an average error rate of 2.61% on the Middlebury datasets, which shows that our method performs almost 20% better than other "fused" algorithms in the aspect of precision.

Keywords: stereo matching; depth sensor; multiscale pseudo-two-layer model; segmentation; texture constraint; fusion move

1. Introduction

Depth estimation is one of the most fundamental and challenging problems in computer vision. For decades, it has been important for many advanced applications, such as 3D reconstruction [1], robotic navigation [2], object recognition [3] and free viewpoint television [4]. Approaches for obtaining 3D depth estimation can be distinguished into two categories: passive and active. The goal of passive methods like stereo matching is to estimate a high-resolution dense disparity map by finding corresponding pixels in image sequences [5]. However, these methods heavily rely on how the scene is presented and contain error matchings caused by the luminance variation. Passive methods fail in textureless and repetitive regions where there is not enough visual information to obtain the correspondence. On the contrary, active methods, like depth sensors (ASUS Xtion [6] and Microsoft Kinect [7]), do not suffer from ambiguities in textureless and repetitive regions, because they emit an infrared signal. Unfortunately, sensor errors and the properties of the object surfaces mean that depth maps from a depth sensor are often noisy [8]. Additionally, their resolution is at least an order of magnitude lower than common digital single-lens reflex (DSLR) cameras, which limits many applications. Moreover, they cannot satisfactorily deal with object boundaries and a wide range of distances. Therefore, fusing different kinds of methods using their complementary characteristics undoubtedly makes the obtained depth map more robust and improves the quality. Commonly-used consumer DSLR cameras have higher resolution and can record better texture information than depth sensors. Therefore, it is reasonable to fuse the depth sensor with DSLR cameras to yield a high resolution depth map. Note that for notational clarity, all values mentioned here are disparities (considering that depth values are inversely proportional to disparities).

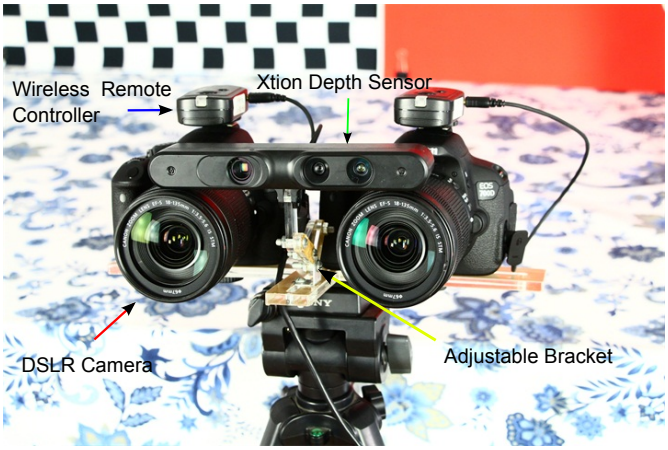

Figure 1. The system used in our method. It consists of two Cannon EOS 700D digital single-lens reflex (DSLR) cameras and one Xtion depth sensor. All DSLR cameras are controlled by the wireless remote controller. We used an adjustable bracket to change the angle and height of the Xtion depth sensor.

In this paper, we propose a novel disparity estimation method for the system shown in Figure 1. It fuses the complementary characteristics of high resolution DSLR cameras and the Xtion depth sensor to obtain an accurate disparity estimate. Compared to the average error rate 3.27% of the previous state-of-the-art methods, our method provides an average error rate of 2.61% on the Middlebury datasets.

It is clear that our method performs almost 20% better than other "fused" algorithms in the aspect of precision. The proposed method views a scene with complex geometric characteristics as a set of segments in the disparity space. It assumes that the disparities of each segment have a compact distribution, which strengthens the smooth variance of the disparities in each segment. Additionally, we assume that each segment is biased towards being a 3D planar surface. The major contributions are as follows:

1. We incorporated texture information as a constraint. The texture variance and gradient is used to restrict the range of the potential disparities for each pixel. In textureless and repetitive regions (which often cause ambiguities when stereo matching), we restrict the possible disparities for a neighborhood centered on each pixel to a limited range around the values suggested by the Xtion. This reduces the errors and strengthens the compact distribution of the disparities in a segment.
2. We propose the multiscale pseudo-two-layer image model (MPTL; Figure 2) to represent the relationships between disparities at different pixels and segments. We consider the disparities from the Xtion as the prior knowledge and use it to increase the robustness to luminance variance and to strengthen the 3D planar surface bias. Furthermore, considering the spatial structures of segments obtained from the depth sensor, we treat the segmentation as a soft constraint to reduce matching ambiguities caused by under- and over-segmentation. Here, pixels with similar colors, but on different objects are grouped into one segment, and pixels with different colors, but on the same object are partitioned into different segments. Additionally, we only retain the disparity discontinuities that align with object boundaries from geometrically-smooth, but strong color gradient regions.

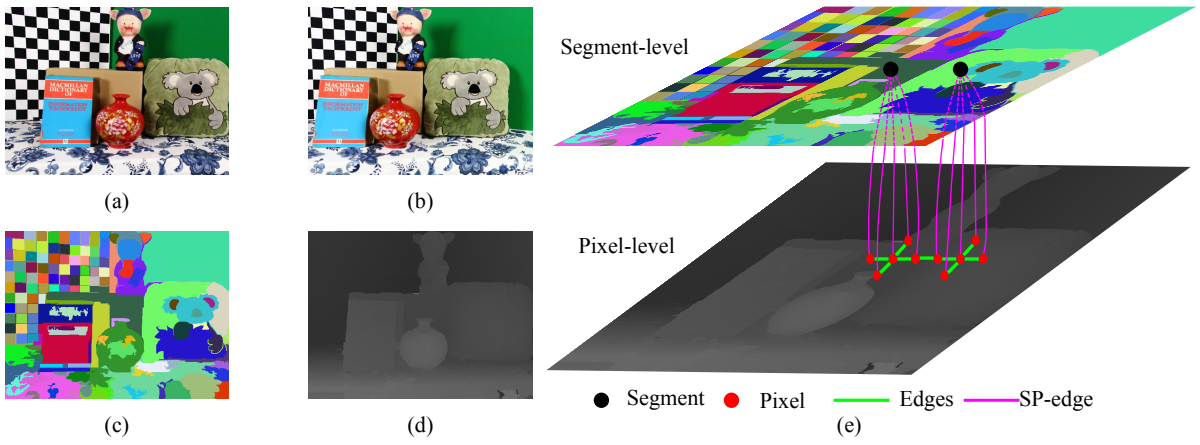

Figure 2. The illustration of the multiscale pseudo-two-layer (MPTL) model. The rectified left (**a**) and right (**b**) images from DLSRs, the segmentation; (**c**) as well as the depth map (**d**) from the depth sensor are taken as our inputs; (**e**) The conceptual structure of our MPTL. The MPTL captures the complementary characteristics of active and passive methods by allowing interactions between them. All interactions are defined to act in the segment-level component, the pixel-level component and the edges that connect them (Segment Pixel-edge, SP-edge). See Section 3.2 for the full details of the MPTL model and Section 4 for further results.

The remainder of this paper is organized as follows. Section 2 gives a summary of various methods used for disparity estimation. We present a pre-processing and some important notations of our model in Section 3.1. We discuss the details of the MPTL image model in Section 3.2, the optimization in Section 3.7 and the post-processing in Section 3.8. Section 4 contains our experiments, and Section 5 presents some conclusions with suggestions for future work.

2. Previous Work

There are many approaches to obtaining disparity estimation. They can generally be categorized into two major classes: passive and active. A passive method indirectly obtains the disparity map using image sequences captured by cameras from different viewpoints. Among the plethora of passive methods, stereo matching is probably the most well known and widely applied. Stereo matching algorithms can be divided into two categories [9]: local and global methods. Local methods [10] estimate disparity using color or intensity values in a support window centered on each pixel. However, they often fail around disparity discontinuities and low-texture regions. Global methods [11] use a Markov random field model to formulate the stereo matching as a maximum *a posteriori* probability energy function with explicit smoothness priors. They can significantly minimize matching ambiguities compared to local methods. However, the biggest disadvantage of them is the low computational efficiency. Segmentation-based global approaches [12,13] encode the scene as a set of non-overlapping homogeneous color segments. They are based on the hypothesis that the variance of the disparity in each segment is smooth. In other words, the segment boundaries are forced to coincide with object boundaries. Recently, the ground control point (GCP)-based methods [14] were used as prior knowledge to encode rich information on the spatial structure of the scene. Although a significant number of stereo matching methods have been proposed for obtaining dense disparity estimation, they heavily rely on radiometric variations and assumptions regarding the presentation of the scene. This means that stereo matching often fails in textureless and repetitive regions, where there is not enough visual information to obtain a correspondence. Furthermore, their accuracy is relatively low. Passive methods heavily rely on the luminance condition and how the scene is presented. They often fail in textureless and repetitive regions where there is not enough visual information to obtain the correspondence.

On the contrary, active methods like depth sensors, do not suffer from ambiguities in textureless and repetitive regions, because they emit an infrared signal. Three different kinds of equipments are used in active methods: a laser scanner device, a time-of-flight (ToF) sensor and an infrared single-based device (such as ASUS Xtion [6] and Microsoft Kinect [7]). The laser scanner device [15] can provide extremely accurate and dense depth estimation, but it is too slow to use in real time and too expensive for many applications. The ToF sensor and infrared single-based device can obtain real-time depth estimation and have recently become available from companies, such as 3DV [16] and PMD [17]. However, sensor errors and the properties of the object surfaces mean that depth maps from them are often noisy [8]. Additionally, their resolution is at least an order of magnitude lower than commonly-used DSLR cameras [18], which limits many applications. Moreover, they cannot satisfactorily deal with object boundaries.

It is clear that each disparity acquisition method is limited in some aspects where other approaches may be effective. Joint optimization methods that combine active and passive sensors have been used to make the obtained depth map more robust and to improve the quality. Zhu *et al.* please check throughout [19,20] fused a ToF sensor and stereo cameras to obtain better disparity maps. They improved the quality of the estimated maps for dynamic scenes by extending their fusion technology to the temporal domain. Yang *et al.* [21] presented a fast depth sensing system that combined the complementary properties of passive and active sensors in a synergistic framework. It relied on stereo matching in rich textured regions, while using data from depth sensors in textureless regions. Zhang *et al.* [22] proposed a system that addresses high resolution and high quality depth estimation by fusing stereo matching and a Kinect. A pixel-wise weighted function was used to reflect the reliabilities of the stereo camera and the Kinect. Wang *et al.* [23] presented a novel method that combined the initial stereo matching result and the depth data from a Kinect. Their method also considers the visibilities and pixel-wise noise of the depth data from a Kinect. Gowri *et al.* [24] proposed a global optimization scheme that defines the data and smoothness costs using sensor confidences and the low resolution geometry from a Kinect. They used a spatial search range to limit the scope of the potential disparities at each pixel. The smoothness prior was based on the available low resolution depth data from the Kinect, rather than the image color gradients.

Although existing disparity estimation methods have achieved remarkable results, they are typically performed using pixel-level cues, such as the smoothness of neighboring pixels, and do not consider the regional information (regarding, for example, 3D spatial structure, segmentation and texture) as a cue for the disparity estimation, which is the largest distinction between their method and ours. For example, occlusion cannot be precisely estimated using a single pixel, but a fitted plane-based filling occlusion in a segment can give good results. Additionally, if the spatial structure of neighboring segments is not known, matching ambiguities can arise at the boundaries of neighboring segments that physically belong to the same object, but have different appearances. Without texture information, we cannot be sure if the disparity from stereo matching is more confident than that from the depth sensor in textureless and repetitive regions (where stereo matching usually fails and the depth sensor performs well).

3. Method

The proposed method can be partitioned into four phases: pre-processing, problem definition, optimization and post-processing. Each phase will be discussed in detail later.

3.1. Pre-Processing

There are three camera coordinates involved in our system (Figure 1): the Xtion coordinate, the coordinates of the two DSLR cameras before the epipolar rectification and the DSLR camera coordinates after the epipolar rectification. During the pre-processing step, in order to combine the data from the Xtion and DSLR cameras, as shown in Figure 3, we firstly calibrated two DSLR cameras using the checkerboard-based method [25] and calibrated the DSLR camera pair with the Xtion sensor using the planar surfaces-based method [26], respectively. After the calibration, the depth image obtained from

the Xtion is first transformed from the Xtion coordinate to the original DSLR cameras' coordinates, then rotated and up-sampled, so that it registers with the unrectified left image. Furthermore, according to the theory of epipolar geometric constraints, the registered depth image and original left image, as well as the original right image are rectified to be row-aligned, which means there are only horizontal disparities in the row direction. We denote the seed image (Π) as the map with disparities transferred from the rectified depth map. Each pixel $p \in \Pi$ is defined as a seed pixel when it is assigned a non-zero disparity. The initial disparity maps (D_L and D_R) of the rectified left and right images (I_L and I_R) are computed using a local stereo matching method [27]. I_L is partitioned into a set of segments using the edge-aware filter-based segmentation algorithm [28].

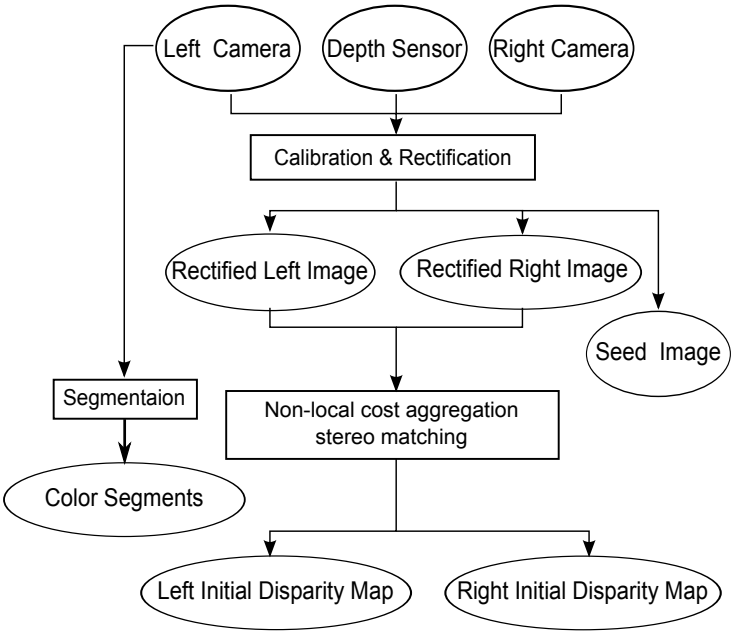

Figure 3. Conceptual flow diagram for the calibration and rectification phase.

In addition, as shown in Figure 4, all pixels and segments are divided into different categories. The occlusion judgment is used to find the occluded pixels with initial disparity maps (D_L and D_R of I_L and I_R, respectively) and to classify pixels into different categories: reliable and occluded. As we known, how to find occluded pixels accurately is always the challenging problem, because it often leads to error results that matching points might not even exist at all, especially in depth discontinuities. Pixels are defined as occluded when they are only visible from the left rectified view (I_L), but not from the right rectified view (I_R). Since image pairs have been rectified, we assume that occlusion only occurs in the horizontal direction. In early algorithms, cross-consistency checking is often applied to identify occluded pixels by enforcing a one-to-one correspondence between pixels. It is written as:

$$O(p) = \begin{cases} 0 & |D_L(p) - D_R(q)| < 1 \\ 1 & \text{otherwise} \end{cases} \qquad p \in I_L, q \in I_R \tag{1}$$

$D_L(p)$ and $D_R(q)$ are the disparity of p and q, and q is the corresponding matching point of p. If p does not meet the cross-consistency checking, then it will be regard as an occluded pixel ($O(p) = 0$); otherwise, p is a reliable pixel ($O(p) = 1$). The cross-consistency checking states that a pixel of one image corresponds to at most one pixel of the other image. However, because of different sampling, the

projection of a horizontal slant or a curved surface shows various lengths in the image pairs. Therefore, conventional cross-consistency checking that often identifies occluded pixels by enforcing a one-to-one correspondence is only suitable for a frontal parallel surface and cannot be true for a horizontal slant or curved surfaces. Considering the different sampling of image pairs, Bleyer *et al.* [29] proposed a new visibility constraint by extending the asymmetric occlusion model [30] that allows a one-to-many correspondence between pixels. Let p_0 and p_1 be neighboring pixels in the same horizontal line of I_L. Then, p_0 will be occluded by p_1 when they meet three conditions:

- p_0 and p_1 have the same matching point in I_R under their current disparity value;
- $D_L(p_0) \leq D_L(p_1)$;
- p_0 and p_1 belong to different segments.

In this paper, for each pixel p of I_L, if there is only one matching point in I_R, the conventional cross-checking is applied to obtain the occlusion Equation (1). Otherwise, if there are more than two matching points in I_R, pixels in I_L are marked as either reliable ($O(p_0) = 0$) or occluded ($O(p_0) = 1$), which satisfy or do not satisfy the Bleyer's asymmetric occlusion model. As shown in Figure 4, each segment belongs to the reliable segment (R) if it contains a sufficient amount of reliable pixels; otherwise, it belongs to the unreliable segment ($\bar{R}$). Furthermore, each segment $f_i \in R$ is denoted as a stable segment (S) when it contains a sufficient number of seed pixels. Otherwise, f_i belongs to the unstable segment ($\bar{S}$). We apply a RANSAC-based algorithm to approximate each stable segment $s_i \in S$ as a fitted plane Ψ_{s_i} using the image coordinates and known disparities of all seed pixels belonging to s_i. Table 1 lists important notifications used in this paper.

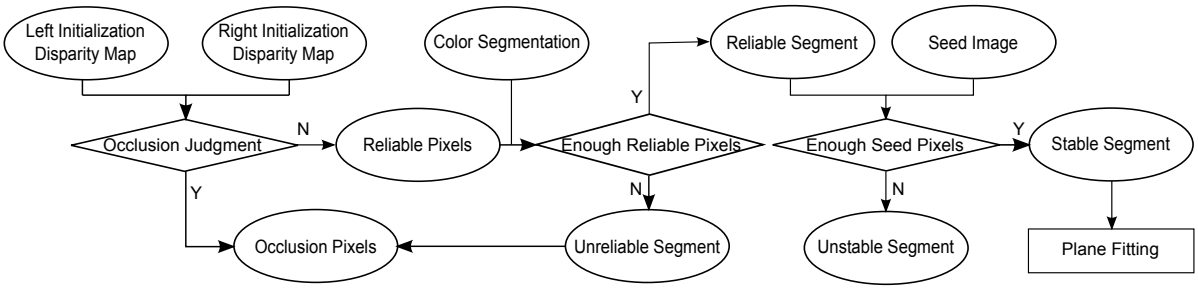

Figure 4. Conceptual flow diagram for the classification phase.

Table 1. Notations.

D:	Disparity map	$D(p)$:	Disparity value of pixel p	Π:	Seed image
D_L:	Initial disparity map of rectified left image	D_R:	Initial disparity map of rectified right image	I_L:	Rectified left DSLR image
I_R:	Rectified right DSLR image	R:	Reliable segment	$\bar{R}$:	Unreliable segment
S:	Stable segment	$\bar{S}$:	Unstable segment	f_i:	i-th segment
s_i:	i-th stable segment	Ψ_{s_i}:	Fitted plane of stable segment s_i	$f(p)$:	Segment that contains pixel p
f_i:	i-th segment	d_l:	Minimum disparity	d_u:	maximum disparity
ϖ:	Segment boundary pixels	Λ_l^p:	Pixel's potential minimum disparity	Λ_u^p:	Pixel's potential maximum disparity
Ψ_i^l:	Minimum fitted disparity of the i-th stable segment	Ψ_i^u:	Minimum fitted disparity of the i-th stable segment		

3.2. Problem Formulation

In the problem formulation phase, we propose the MPTL model, which combines the complementary

characteristics of stereo matching and the Xtion sensor. As shown in Figure 2, the MPTL model consists of three components:

- The pixel-level component, which improves the robustness against the luminance variance (Section 3.3) and strengthens the smoothness of disparities between neighboring pixels and segments (Section 3.4). Nodes at this level represent reliable pixels from stable and unstable segments. The edges between reliable pixels represent different types of smoothness terms.
- The edge that connects two level components (the SP-edge), which uses the texture variance and gradient as a guide to restrict the scope of potential disparities (Section 3.5).
- The segment-level component, which incorporates the information from the Xtion as prior knowledge to capture the spatial structure of each stable segment and to maintain the relationship between neighboring stable segments (Section 3.6). Each node at this level represents a stable segment.

Existing global methods have achieved remarkable results, but the capability of the traditional Markov random field stereo model remains limited. To lessen the matching ambiguities, additional information is required to formulate an accurate model. In this paper, the pixel-level improved luminance consistency term (E_l), the pixel-level hybrid smoothness term (E_s) and the SP-edge texture term (E_t), as well as the segment-level 3D plane bias term (E_p) are integrated as additional regularization constraints to obtain a precise disparity estimation (D) for a scene with complex geometric characteristics. According to Bayes' rule, the posterior probability over D given l, s, t and p is:

$$p(D|l,s,t,p) = \frac{p(l,s,t,p|D)p(D)}{p(l,s,t,p)} \tag{2}$$

During each optimization process, $P(l,s,t,p|D)$ is only dependent on l, s, t and p. Therefore, $P(D|l,s,t,p)$ can be rewritten as:

$$p(D|l,s,t,p) \propto p(l|D)p(s|D)p(t|D)p(p|D)p(D) \tag{3}$$

Because maximizing this posterior is equivalent to minimizing its negative log likelihood, our goal is to obtain the disparity map (D) that minimizes the following energy function:

$$E(D) = E_l(D) + E_s(D) + E_t(D) + E_p(D) \tag{4}$$

Each term will be discussed in detail in the following sections.

3.3. Improved Luminance Consistency Term

The conventional luminance consistency hypothesis is used to penalize the appearance dissimilarity between corresponding pixels in I_L and I_R, based on the hypothesis that the surface of a 3D object is Lambertian. Because it refers to a perfectly-diffuse appearance in which pixels originating from the same 3D object have similar appearances in different views, its accuracy is heavily dependent on the

lighting condition for which colors change substantially depending on the viewpoint. Furthermore, an object may appear to have different colors because different views have different sensor characteristics. In contrast, the Xtion sensor is more robust to the light condition and can be used as prior knowledge to reduce ambiguities caused by the non-Lambertian surface. Thus, the improved luminance consistency term is denoted as:

$$E_l = \sum_{p \in I_L} \lambda_l \cdot (1 - O(p)) \cdot [w_p^l \cdot C(p,q) + w_p^x \cdot X(p,q)] + O(p) \cdot \lambda_o \tag{5}$$

where q is the matching pixel of p in the other image. $O(p)$ is the asymmetric occlusion function described in Section 3.1, and λ_o is a positive penalty used to avoid maximizing the number of occluded pixels. $C(p,q)$ is defined as the pixel-wise cost function from stereo matching to measure the color dissimilarity.

$$C(p,q) = \alpha \cdot (1 - \exp(-\frac{C_{ssd}(p,q)}{r_{ssd}})) + (1-\alpha) \cdot (1 - \exp(-\frac{C_g(p,q)}{r_g})) \tag{6}$$

where r_{ssd} and r_g are constant values defined by our experience. α is the scalar weight from zero to one. $C_{ssd}(p,q)$ and $C_g(p,q)$ are the color dissimilarity and gradient in three color channels as:

$$C_{ssd}(p,q) = \sqrt{\sum_{i=R,G,B} (I_L^i(p) - I_R^i(q))^2} \tag{7}$$

$$C_g(p,q) = \sqrt{\sum_{i=R,G,B} (\nabla I_L^i(p) - \nabla I_R^i(q))^2} \tag{8}$$

$X(p,q)$ is the components from the Xtion sensor, which are defined as:

$$X(p,q) = \min\{|D(p) - \Pi(p)|, T_\pi\} \tag{9}$$

T_π is the constant threshold, and $D(p)$ is the disparity value assigned to pixel p in each optimization. $\Pi(p)$ is the disparity of pixel $p \in \Pi$. w_p^x and w_p^l are pixel-wise confidence weights that are denoted as $w_p^x = 1 - w_p^l$. They are derived from the reliabilities of disparities obtained from stereo matching (m_p^l) and the Xtion (m_p^x) as:

$$w_p^l = \begin{cases} \frac{m_p^l}{m_p^l + m_p^x} & p \in \Pi \\ 1 & otherwise \end{cases} \tag{10}$$

where m_p^l is similar to the attainable maximum likelihood (AML) in [31], which models the cost for each pixel using a Gaussian distribution centered at the minimum actually achieved cost value for that pixel. The reliability of Xtion data m_p^x is the inverse of the normalized standard deviation of the random error [20]. The confidence of each depth value obtained from the depth sensor decreases with the increasing of the normalized standard deviation.

3.4. Hybrid Smoothness Term

The hybrid smoothness term strengthens the segmentation-based assumption that the disparity variance in each segment is smooth and reduces errors caused by under- and over-segmentation. It consists of four terms: the smoothness term for neighboring reliable pixels belonging to the unstable segment (E_{s_0}), the smoothness term for neighboring reliable pixels in the same stable segment (E_{s_1}), the smoothness term for neighboring reliable pixels in different stable segments (E_{s_2}) and the smoothness term for neighboring reliable pixels that belong to stable and unstable segments (E_{s_3}).

Because there is no prior knowledge about the spatial structure of unstable segments, we define the smoothness term E_{s_0} as the conventional second-order smoothness prior Equation (11), which can produce better estimates for a scene with complex geometric characteristics [11].

$$E_{s_0} = \sum_{\Phi_i^0 \in \Phi^0} \lambda_s^0 \cdot [1 - exp(-\frac{\nabla D(\Phi_i^0)}{\gamma_s})] \quad \{p_0, p_1, p_2\} \in \Phi_i^0 \tag{11}$$

where γ_s and λ_s^0 are the geometric proximity and the positive penalty. Φ^0 is the set of triple-cliques consisting of consecutive reliable pixels belonging to unstable segment. $\nabla D(\Phi_i^0)$ is the second derivative of the disparity map as:

$$\nabla D(\Phi_i^0) = D(p_0) - 2D(p_1) + D(p_2). \quad \{p_0, p_1, p_2\} \in \Phi_i^0 \tag{12}$$

E_{s_0} captures richer features of the local structure and permits planar surfaces without penalty by setting $\nabla D(\Phi_i^0) = 0$. However, E_{s_0} only considers disparity information when representing the smoothness of neighboring pixels. This means that, in several cases, error matching can result in different disparity assignments, which correspond to the same second derivatives in the disparity map (see Figure 5b–d). Meanwhile, each stable segment can be represented as a fitted plane using the disparity data from the Xtion, which contains prior knowledge about the spatial structure of each stable segment. We can incorporate the spatial similarity weight with the prior knowledge from the Xtion into a conventional second-order smoothness prior. This term encourages constant disparity gradients for pixels in a stable segment and local spatial structures that are similar to the fitted plane of the stable segment. The smoothness term for neighboring reliable pixels in the same stable segment is as follows.

$$E_{s_1} = \sum_{\Phi_i^1 \in \Phi^1} \lambda_s^1 \cdot [2 - \delta(\Phi_i^1) - \exp(-\frac{\nabla D(\Phi_i^1)}{\gamma_s})] \tag{13}$$

where Φ^1 is the set of triple-cliques defined by all 3×1 and 1×3 consecutive reliable pixels along the coordinate direction of the rectified image coordinate in each stable segment. λ_s^1 is a positive value penalty. As for the spatial 3D relationship shown in Figure 5, let s_i be the stable segment containing Φ_i^1 and Ψ_{s_i} be its corresponding fitted plane. Then, the spatial similarity weight $\delta(\Phi_i^1)$ is denoted as:

$$\delta(\Phi_i^1) = \begin{cases} 0 & \text{I: } \overline{p_0p_1} \neq \overline{p_1p_2} \\ 0.25 & \text{II: } \overline{p_0p_1} = \overline{p_1p_2} \quad \text{and} \quad \overline{p_0p_2} \cap \Psi_{s_i} \\ 0.5 & \text{III: } \overline{p_0p_1} = \overline{p_1p_2} \quad \text{and} \quad \overline{p_0p_2} // \Psi_{s_i} \\ 1 & \text{IV: } \overline{p_0p_1} = \overline{p_1p_2} \quad \text{and} \quad \overline{p_0p_2} \in \Psi_{s_i} \end{cases} \quad \{p_0, p_1, p_2\} \in \Phi_i^1 \tag{14}$$

- Case I: When $\overline{p_0p_1} \neq \overline{p_1p_2}$, the disparity gradients of pixels in Φ_i^1 are not constant ($\nabla D(\Phi_i^1) \neq 0$). This case violates the basic segmentation assumptions that the disparity variance of neighboring pixels is smooth, so a large penalty is added to prevent it from happening in our model (see Figure 5a).
- Cases II, III and IV: When $\overline{p_0p_1} = \overline{p_1p_2}$, the disparity gradients of pixels in Φ_i^1 are constant ($\nabla D(\Phi_i^1) = 0$). This means that the variance of the disparities is smooth. Furthermore, our model checks the relationship between all pixels in Φ_i^1 and Ψ_{s_i} (see Figure 5b–d). $\delta(\Phi_i^1)$ does not penalize the disparity assignment if all pixels in Φ_i^1 belong to Ψ_{s_i} (Case IV in Figure 5d), because it is reasonable to assume that the local structure of Φ_i^1 is the same as the spatial structure of Ψ_{s_i}. Note that we impose a larger penalty to Case II than to Case III to strengthen the similarity between the spatial structure of Φ_i^1 and Ψ_{s_i}.

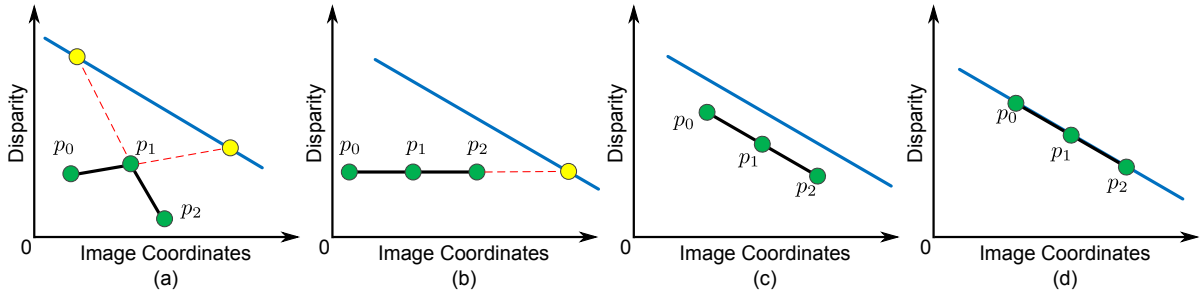

Figure 5. Smoothness term for pixels in the same stable segment (s_i) with the spatial similarity weight as the triple-clique $\Phi_i^1 = \{p_0, p_1, p_2\}$, under different disparity assignments. Given Ψ_{s_i} (blue line) as the fitted plane of s_i. Yellow nodes are the intersect points. (**a**) Case I, $D(p0) \neq D(p1) \neq D(p2)$ with $\nabla D(\Phi_i^1) \neq 0$; (**b**) Case II, $D(p0) = D(p1) = D(p2)$ with $\nabla D(\Phi_i^1) = 0$; (**c**) Case III, $D(p0) \neq D(p1) \neq D(p2)$ with $\nabla D(\Phi_i^1) = 0$; (**d**) Case IV, $D(p0) \neq D(p1) \neq D(p2)$ with $\nabla D(\Phi_i^1) = 0$. As shown in Cases II, III and IV, different disparity assignments correspond to the same second derivative ($\nabla D(\Phi_i^1) = 0$).

In some segmentation-based algorithms [32], the segmentation is implemented as a hard constraint by setting λ_s^0 and λ_s^1 to be positive infinity. This does not allow any large disparity variance within a segment. In other words, each segment can only be represented as a single plane model, and the boundaries of a 3D object must be exactly aligned with segment boundaries. Unfortunately, not all segments can be accurately represented as a fitted plane, and not all 3D object boundaries coincide with segment boundaries. The accuracy of the segmentation-based algorithms is easily affected by the initial segmentation. On the one hand, the initial segmentation typically contains some under-segmented regions (where pixels from different objects, but with similar colors are grouped into one segment). As a direct consequence of under-segmentation, foreground and background boundaries are blended if they have similar colors at disparity discontinuities. To avoid this, we use segmentation as a soft constraint by setting λ_s^0 and λ_s^1 to be positive finite, so that each segment can contain arbitrary fitted planes.

On the other hand, pixels with different colors, but on the same object are over-segmented into different segments in the initial segmentation, which causes computationally inefficiency and ambiguities on segment boundaries. In this paper, we considered the spatial structure of neighboring stable segments

using disparities from the Xtion. Therefore, we apply the smoothness term for neighboring pixels belonging to different stable segments (E_{s_2}) to avoid errors caused by the over-segmentation. Let p and q be neighboring pixels belonging to stable segments s_i and s_j, respectively, Then, E_{s_2} can be expressed as:

$$E_{s_2} = \begin{cases} 0 & \text{I:}\Psi_{s_i} \neq \Psi_{s_j} \quad and \quad D(p) \neq D(q) \\ \lambda_s^2 & \text{II::}\Psi_{s_i} \neq \Psi_{s_j} \quad and \quad D(p) = D(q) \\ \lambda_s^2 & \text{III::}\Psi_{s_i} = \Psi_{s_j} \quad and \quad D(p) \neq D(q) \\ 0 & \text{IV:}\Psi_{s_i} = \Psi_{s_j} \quad and \quad D(p) = D(q) \end{cases} \tag{15}$$

As shown in Equation (15), for Cases I and II, if Ψ_{s_i} is not equal to Ψ_{s_j}, this means that s_i and s_j have different spatial structures, and the 3D object boundary coincides with the boundary between them. The disparity variance between p and q is allowed without any penalty (Case I); otherwise, a constant penalty λ_s^2 is added (Case II). In contrast, for Cases III and IV, if Ψ_{s_i} is equal to Ψ_{s_j}, this means that s_i and s_j have different appearances, but have similar spatial structures and belong to the same 3D object. In these two cases, the disparity variance between p and q is not allowed by adding a penalty. E_{s_2} reduces the ambiguities caused by over-segmentation and retains only the disparity discontinuities that are aligned with object boundaries from geometrically-smooth, but strong color gradient regions, where pixels with different colors, but from the same object are partitioned into different segments.

Because unstable segments do not have sufficient disparity information from the Xtion to regard their spatial plane models, the smoothness term for neighboring pixels that belong to the stable and unstable segments (E_{s_3}) encourages neighboring pixels to take the same disparity assignment. It takes the form of a standard Potts model,

$$E_{s_3} = \begin{cases} 0 & D(p) = D(q) \\ \lambda_s^3 & D(p) \neq D(q) \end{cases} \quad p \in S, q \in \bar{S} \tag{16}$$

Thus, let ϖ be the set of pixels belonging to segment boundaries, the hybrid smoothness term is:

$$E_s = \begin{cases} E_{s_0} & \{p_0, p_1, p_2\} \in \Phi_i \text{ and } \Phi_i \in \bar{S} \text{ and } \Phi_i \cap \varpi = \emptyset \\ E_{s_1} & \{p_0, p_1, p_2\} \in \Phi_i \text{ and } \Phi_i \in S \text{ and } \Phi_i \cap \varpi = \emptyset \\ E_{s_2} & \{p_0, p_1, p_2\} \in \Phi_i \text{ and } \Phi_i \cap \varpi \neq \emptyset \text{ and } \{p_0, p_1, p_2\} \in S \\ E_{s_3} & \{p_0, p_1, p_2\} \in \Phi_i \text{ and } \Phi_i \cap \varpi \neq \emptyset \text{ and } \{p_0, p_1\} \in S, \{p_2\} \in \bar{S} \end{cases} \tag{17}$$

3.5. Texture Term

Stereo matching often fails in textureless and repetitive regions, because there is not enough visual information to obtain a correspondence. However, the Xtion does not suffer from ambiguities in these regions. Therefore, the disparities from the Xtion are more reliable than those obtained from stereo matching on textureless and repetitive regions and should be closer to the range of potential disparities for pixels in these regions. In contrast, the disparities from the Xtion are susceptible to noise and problems caused by rich texture regions and have poor performance in preserving object boundaries. Therefore, the disparities obtained from stereo matching are more reliable than that of the Xtion and should be used to define the scope of potential disparities of pixels in those regions. Considering the complementary

characteristics of stereo matching and the Xtion sensor, texture information can be used as a useful guide for disparities.

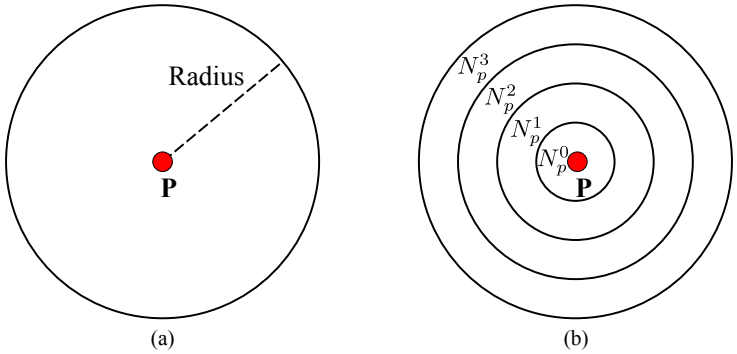

Figure 6. Surrounding neighborhood patch, N_p, for: (**a**) pixel p and (**b**) its corresponding sub-regions.

The texture variance and gradient are used as a cue to restrict the scope of potential disparities for pixels. This reduces errors caused by noise or outliers and makes the distribution of the disparity more compact. To do this, we first define a surrounding neighborhood patch N_p (with a radius of, for example, 20 pixels) centered at each pixel $p \in I_L$, as shown in Figure 6. Considering that the annular spatial histogram is translation and rotation invariant [33], N_p is evenly partitioned into four annular sub-regions. For each sub-region $N_p^i (i = 0 \cdots 3)$, we compute its normalized intensity 16-bin gray histogram $H_p^i = \{h_p^{(i,j)}, j = 0 \cdots 15\}$ to represent the annular distribution density of N_p as a 64-dimensional feature vector.

Finally, let L_p be a 1D line segment ranging from $(p - 10)$ to $(p + 10)$ in the same row of p in I_L. The texture variance and gradient of p is determined by the texture dissimilarity Γ_p Equation (18), using the Hamming distance Equation (19) between the annular distribution densities of p and its neighboring pixel q in L_p. That is,

$$\Gamma_p = min_{(q \in L_p, q \neq p)} \sum_{i=0}^{3} \sum_{j=0}^{15} H(h_p^{(i,j)}, h_q^{(i,j)}) \tag{18}$$

$$H(h_p^{(i,j)}, h_q^{(i,j)}) = \begin{cases} 1 & |h_p^{(i,j)} - h_q^{(i,j)}| \geq T_H \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

Each pixel's disparity variance buffer (Ω_p) can be denoted as:

$$\Omega_p = 1 + [1 - exp(-\frac{\Gamma_p}{\gamma_H})] \cdot \xi \qquad \xi = 0.2 \cdot (d_u - d_l) \tag{20}$$

d_l and d_u are the minimum and maximum disparities. Γ_p is small in the textureless and repetitive regions and is large in the rich texture regions or object boundaries. The scope of each pixel's potential disparities $[\Lambda_l^p, \Lambda_u^p]$ is denoted as:

$$\Upsilon_l = \max \{(\Psi_{f(p)}^l - \Omega_p), d_l)\} \tag{21}$$

$$\Upsilon_u = \min \{(\Psi_{f(p)}^l + \Omega_p), d_u)\} \tag{22}$$

$$\Lambda_l^p = \begin{cases} \max\{(\Theta_l^p - \Omega_p), \Upsilon_l\} & \chi_p \geq T_\Lambda \quad and \quad f(p) \in S \\ \Upsilon_l & \chi_p < T_\Lambda \quad and \quad f(p) \in S \\ d_l & f(p) \in \bar{S} \end{cases} \tag{23}$$

$$\Lambda_u^p = \begin{cases} \min\{(\Theta_u^p + \Omega_p), \Upsilon_u\} & \chi_p \geq T_\Lambda \quad and \quad f(p) \in S \\ \Upsilon_u & \chi_p < T_\Lambda \quad and \quad f(p) \in S \\ d_u & f(p) \in \bar{S} \end{cases} \tag{24}$$

where Θ_l^p and Θ_u^p are the minimum and maximum disparities from the Xtion in the region centered at p in I_L. $f(p)$ is the segment that contains p. $\Psi_{f(p)}^l$ and $\Psi_{f(p)}^u$ are the minimum and maximum fitted disparities of $f(p)$. χ_p is the number of seed pixels in the region centered at p. T_Λ is a positive value. As described in Equation (23), there are three cases for the definition of Λ_l^p:

- When $f(p)$ is a stable segment ($f(p) \in S$) and contains sufficient seed pixels ($\chi_p > T_\Lambda$), Λ_l^p is equal to $\max\{(\Theta_l^p - \Omega_p), \Upsilon_l\}$. In this case, there are enough seed pixels from the Xtion to denote a guide for the variance of disparities of p. If p is in the textureless or repetitive region, Ω_p is small. This indicates that stereo matching may fail in these regions, and a small search range should be used around disparities from the Xtion. In contrast, if p is in the rich textured region or object boundaries, Ω_p is large. This indicates that disparities from the Xtion may be susceptible to noise and problems caused by rich texture regions where disparities obtained from stereo matching are more reliable. Then, a broader search range should be used, so that we can extract better results not observed by the Xtion.
- When $f(p)$ is a stable segment ($f(p) \in S$), but there are not enough seed pixels around p ($\chi_p \leq T_\Lambda$), Λ_l^p is equal to Υ_l. In this case, although there are some seed pixels from the Xtion, they are not enough to represent the disparity variance around p. On the other hand, because each stable segment is viewed as a 3D fitted plane, the search range for the potential disparities is limited by the fitted disparity of $f(p)$ and the disparity variance buffer (Ω_p).
- When $f(p)$ is an unstable segment ($f(p) \in \bar{S}$), Λ_l^p is the minimum disparity (d_l).

Similarly, Λ_u^p can be obtained in the same way. Then, the SP-edge term (which defines the scope of pixel's potential disparities) is:

$$E_t = \begin{cases} 0 & \Lambda_l^p \leq D(p) \leq \Lambda_u^p \\ \lambda_t & \text{otherwise} \end{cases} \tag{25}$$

3.6. 3D Plane Bias Term

This 3D plane bias term focuses on strengthening the assumption that each stable segment has a 3D plane bias. It is denoted as:

$$E_p = \sum_{s_i \in S} \sum_{p \in s_i} \lambda_p \cdot \min\{|D(p) - \Psi_{s_i}(p)|, T_p\} \tag{26}$$

where $D(p)$ is the assigned value of pixel p in I_L. $\Psi_{s_i}(p)$ is the plane fitted value, and T_p is a threshold value. Note that for notation clarity, the traditional 3D bias assumption is a hard constraint that forbids

any distinctive between $D(p)$ and $\Psi_{s_i}(p)$ by setting λ_p to be infinite. On the contrary, our 3D plane bias term is a soft constraint that a certain distinctive between $D(p)$ and $\Psi_{s_i}(p)$ is allowed by setting λ_p to be a finite positive value.

3.7. Optimization

The energy function defined in Equation (4) is a function of the real discrete disparity map. In this section, we describe how to optimize Equation (4) using the fusion move algorithm to obtain the disparity map D^*:

$$D^* = argmin_D E(D) \tag{27}$$

The fusion move approach [34] is an extended approach of the $\alpha - expansion$ algorithm [35], which allows arbitrary values for each pixel in the proposed disparity map. It generates a new result by fusing the current and proposed disparity maps with the energy either decreasing or remaining constant. Let D^c and D^p be the current and proposed disparity maps of I_L. Our goal is to optimally "fuse" D^c and D^p to generate a new depth map D^n, so that the energy $E(D^n)$ is lower than $E(D^c)$. This fusion move is achieved by taking each pixel in D^n from either D^c or D^p, according to a binary indicator map B. B is the result of the graph cut-based fusion move Markov random field optimization technique. During each optimization, each pixel either keeps its current disparity value ($B(p) = 0$) or changes it to proposed disparity value ($B(p) = 1$). That is,

$$D^n = (1 - B) \cdot D^c + B \cdot D^p \tag{28}$$

However, the fusion move is limited to optimizing the submodular binary fusion-energy functions that consist of unary and pairwise potentials. Because of the hybrid smoothness term, our binary fusion-energy functions are not submodular and cannot be directly solved using the fusion move [36]. Using the quadratic pseudo-Boolean optimization (QPBO) algorithm [37], we can obtain a partial solution for the non-submodular binary fusion-energy function by assigning either zero or one to partial pixels, and leaving the rest unassigned. The partial solution is a part of the global minimum solution, and its energy is not higher than that of the original solution. Because of the given lowest average number of unlabeled pixels, we used Quadratic Pseudo Boolean Optimization with Probing (QPBO-P) [38] and Quadratic Pseudo Boolean Optimization with Improving (QPBO-I) [39] as our fusion strategies. During the optimization, the pixel-level improved luminance consistency term (E_l), the SP-edge texture term (E_t) and the segment-level 3D plane bias term (E_p) are expressed as unary terms, respectively. We tackle the transformation problem of the pixel-level hybrid smoothness term (E_s) that contains triple-cliques using the decomposition method called Excludable Local Configuration (ELC) [40]. The essence of the ELC method is a QPBO-based transformation of a general higher-order Markov random field with binary labels into a first-order one that has the same minima as the original. It combines a new reduction with the fusion move and QPBO to approximately minimize higher-order multi-label energies. Furthermore, the new reduction technique is along the lines of the Kolmogorov-Zabih reduction that can reduce any higher-order minimization problem of Markov random fields with binary labels into an equivalent first-order problem. Each triple

clique in E_s is decomposed into a set of unary or pairwise terms by ELC without introducing any new variables.

The choice of the proposed disparity maps in the fusion move approach is another crucial factor for the successful use and efficiency of the fusion move. Because there is not an algorithm that can be applied to all situations, our goal is to expect all proposed disparity maps to be correct in some parts and under some parameter setting. Here, we use the following schemes to obtain all proposed disparity maps:

- Proposal A: Uniform value-based proposal. All disparities in the proposal are assigned to a discrete disparity, in the range of d_l to d_u.
- Proposal B: The hierarchical belief propagation-based algorithm [41] is applied to generate proposals with different segmentation maps.
- Proposal C: The joint disparity map and color consistency estimation method [42], which combines mutual information, a SIFT descriptor and segment-based plane-fitting techniques.

During each optimization, the result of the current fusion move is used as the initial disparity map of the next iteration.

3.8. Post-Processing

The post-processing is composed of two steps: filling occlusions and refinement. Given that p is a occluded pixel in I_L, a two-step method is implemented to estimate disparities of occluded pixels. If $f(p) \in S$, the fitted plane value $\Psi_{f(p)}(p)$ is assigned as p's disparity. Otherwise, the disparity of p is the smaller disparity of its closet left and right seed pixels that belongs to the background.

After filling occlusions, in order to obtain an accurate disparity map and to remove ambiguities at object boundaries, the weighted joint bilateral filter with the slope depth compensation filter [43] is applied to refine the disparity map.

4. Results and Discussion

Here, a series of evaluations were performed to verify the effectiveness and accuracy of the proposed method. Results were composed of qualitative and quantitative analyses. The segmentation parameters for all experiment are the same: spatial bandwidth $= 7$, color bandwidth $= 6.5$, minimum region $= 20$. Other parameters are presented in Table 2. They were kept constant for all experiments and were typically empirically based.

Table 2. Parameter settings for all experiments.

T_H	T_Λ	T_π	T_p	γ_H	γ_s	λ_t	λ_l	λ_o	λ_s^0	λ_s^1	λ_s^2	λ_s^3	λ_p
0.1	50	3	5	35	1.5	200	12	200	40	40	15	15	10

4.1. Qualitative Evaluation Using the Real-World Datasets

We performed qualitative analyses of the proposed method using real-world datasets. In all evaluations, we captured the image pairs using the system in Figure 1 and regarded the left DSLR cameras as the target to be estimated by the disparity map. Notice that all scenes contain weakly-textured and repetitive regions, as well as a non-Lambertian surface.

In order to illustrate that our method combines the complementary characteristics of the various disparity estimation methods and outperforms using the conventional stereo matching or depth sensor alone, we evaluated the qualitative quality of the disparity estimates from three stereo matching methods, the Xtion depth sensor and the proposed method using several complex indoor scenes in Figure 7. As show in Figure 7c, although the local stereo matching with fast cost volume filtering (FCVF [44]) performed well by recovering the object boundaries using a color image as a guide, it was very fragile for noise and textureless regions (such as the uniformly-colored board in the yellow rectangle of Figure 7a). In contrast, the depth values obtained from the active sensor are more accurate (see the same regions in Figure 7f). Therefore, our method overcame these problems with the improved luminance consistency term and texture term by incorporating the prior depth information from the depth sensor. As shown in Figure 7d, segmentation-based global stereo matching with second order smoothness prior (SOSP [11]) overcame some of the problems caused by the noise and outliers, but did not solve the problems caused by the over-segmentation, which led to ambiguous matching when segment boundaries did not correspond to object boundaries (green rectangle in Figure 7a). Segment tree-based stereo matching (ST [13]) blended the foreground and background in the under-segmented region (as shown in Figure 7e), where different objects with similar appearances (such as the red rectangle in Figure 7b) were grouped into a segment. Comparing to our result and those of segmentation-based methods, it is clear that the proposed hybrid smoothness term helps reduce matching ambiguities causing by over-segmentation and under-segmentation with the indication of the depth sensor. The raw data from the depth sensor were noisy and had poor performance in preserving object boundaries (blue rectangle in Figure 7f); our result is more robust in this situation and can be used to improve the performance of the depth sensor by considering the color and segmentation information from stereo matching. Based on the above, we can safely draw the conclusion that the proposed method obtains accurate depth estimation by combining the complementary characteristics of stereo matching and the depth sensor. We also tested the proposed method on other real-world scenes to verify its robustness (see Figure 8).

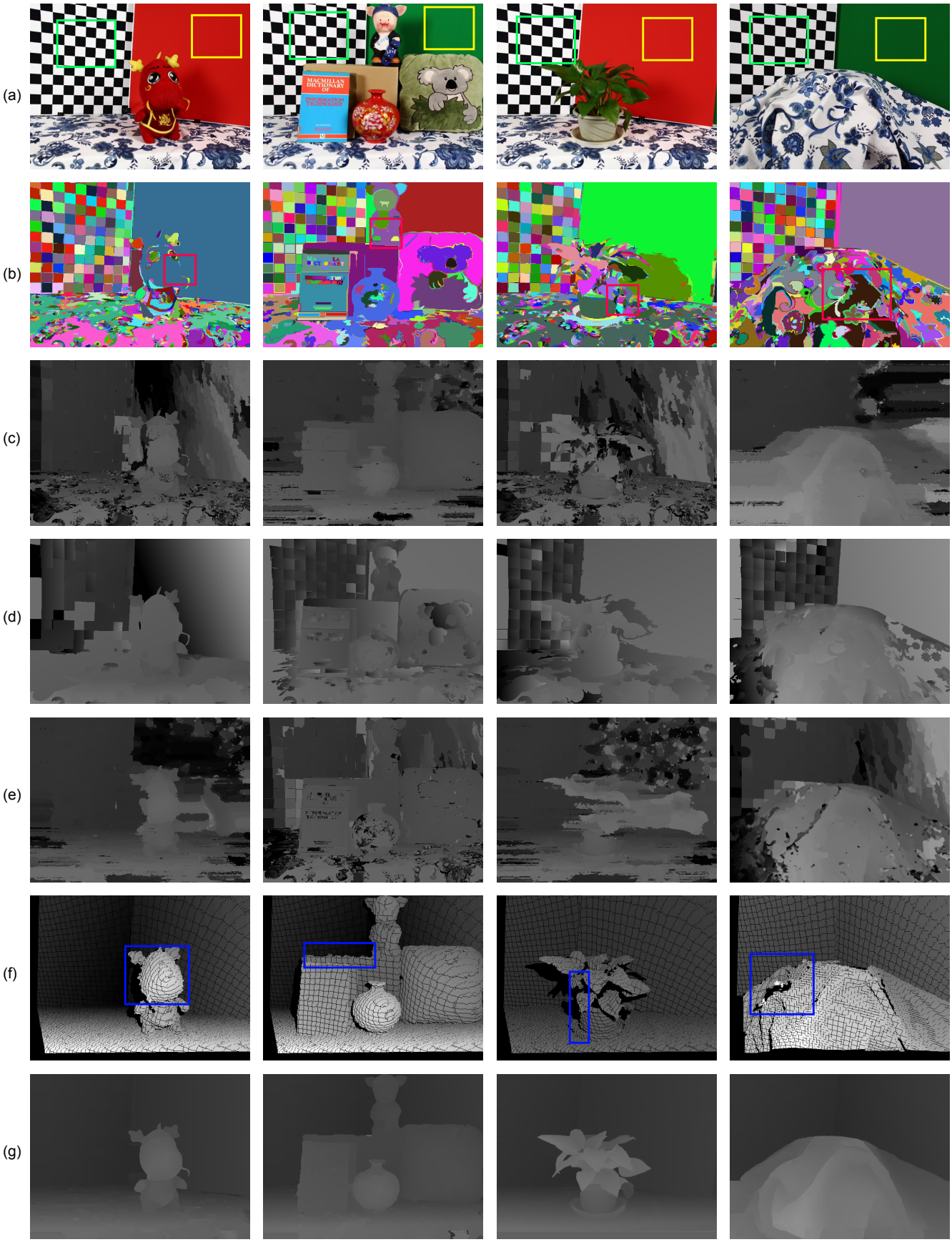

Figure 7. Results of the different methods applied to the real-world scenes: Dragon, Book, Plant and Tablecloth. Each column from up to down is: (**a**) the rectified left image; (**b**) the segmentation result, (**c**) the disparity map of FCVF [44]; (**d**) the disparity map of SOSP [11]; (**e**) the disparity map of segment tree (ST) [13]; (**f**) the seed image transformed from the Xtion data and (**g**) our result.

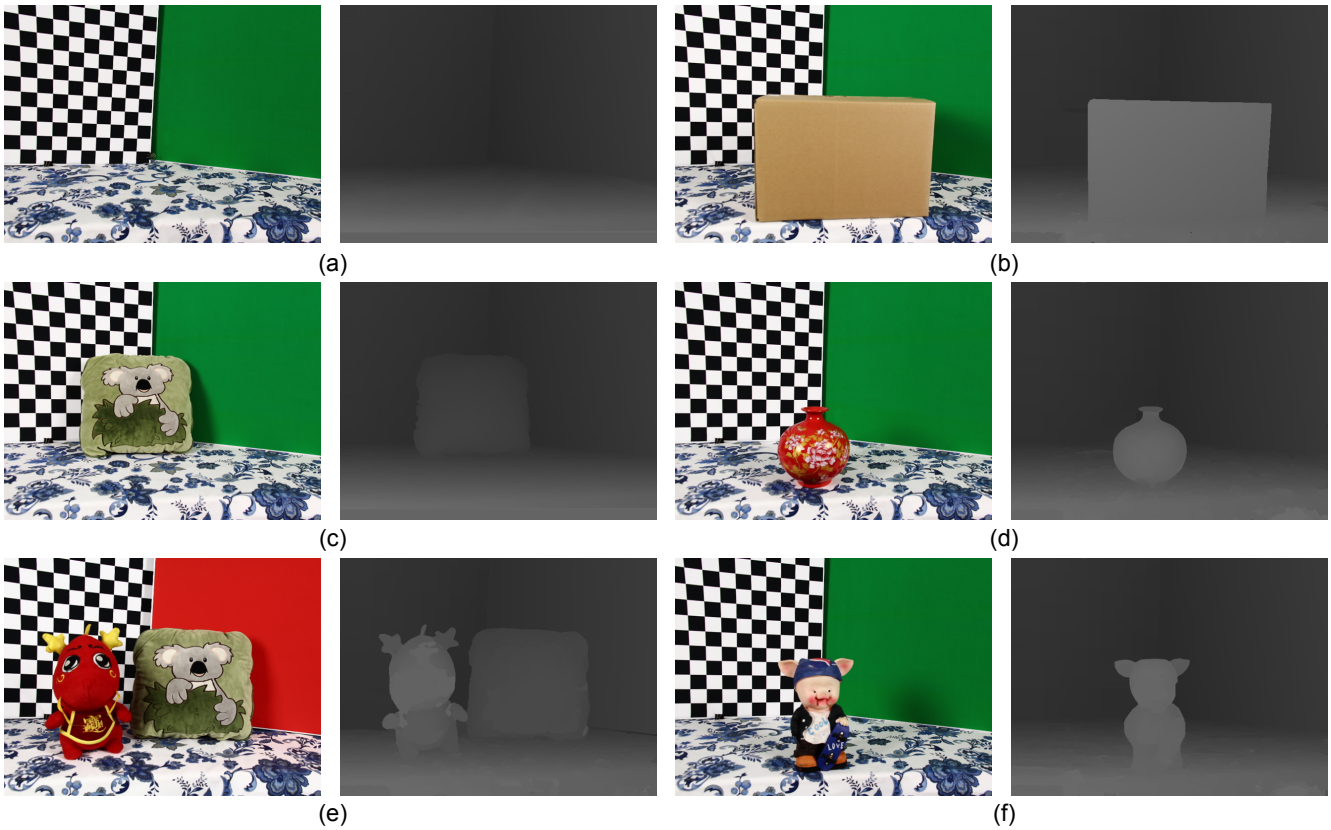

Figure 8. Comparative results for different real-world scenes: (**a**) Board; (**b**) Box; (**c**) Kola; (**d**) Vase; (**e**) Dragon and Kola; (**f**) Piggy. All scenes were approximately 0.5–1.5 m from the cameras, and the maximum disparity was 107 pixels. Each scene from left to right contains the rectified left image and its associated result of our method.

Furthermore, we implemented the post-processing processing introduced in Section 3.8 to assign valid disparities to pixels in the black regions of seed images. The seed image after assignment can be treated as the up-sampling disparity map of the target image captured by the Xtion alone. Then, we evaluated the quality of the 3D reconstruction from our method and that using only Xtion data (see Figure 9). The 3D point cloud reconstructions consist of the pixels' image coordinates and their associated disparities in disparity space. The blue rectangles highlight some regions where our method performed well. For example, the proposed method was more effective at retaining the boundaries of Piggy and Plant (Figure 9a,c) and correctly recovered the top of the head and beard of the dragon (Figure 9b). These comparisons illustrate that the stereo matching using the the depth sensor as the prior knowledge is more effective and accurate than using stereo matching or the depth sensor alone.

Figure 9. Comparative results for 3D reconstructions. (**a**) Book; (**b**) Dragon; (**c**) Plant; (**d**) Tablecloth; (**e**) Box; (**f**) Piggy; (**g**) Vase; (**h**) Dragon and Kola. Each scene from top to bottom contains the reconstruction using our method and the result using the up-sampled disparity map captured by the Xtion depth sensor.

4.2. Quantitative Evaluation Using the Middlebury Datasets

To quantitatively illustrate the validity of the proposed method, we also conducted evaluations on the Middlebury datasets [9,45] and focused on recovering the disparity map of the left image in each dataset. The evaluation is made by third-size resolution Views 1 and 5 of all image pairs. However, because there is nothing about the scanning depth information of this dataset, we used the method described in [14] to simulate the seed image transformed from the Xtion projected to View 1. This technique is based on a voting strategy and simply requires some disparity maps produced using several stereo methods [46–48]. Each pixel was labeled as a seed if its disparity in different maps was consistent (varied by less than a fixed threshold and was not near the intensity edge). Results on these datasets and their corresponding errors (compared with the ground truth) in non-occlusion regions are shown in Figure 10. As shown in Figure 11, our method ranks first among approximately 164 methods listed on the website [49]. It performs especially well on the Tsukuba image pairs, with minimum errors in non-occluded regions and near depth discontinuities.

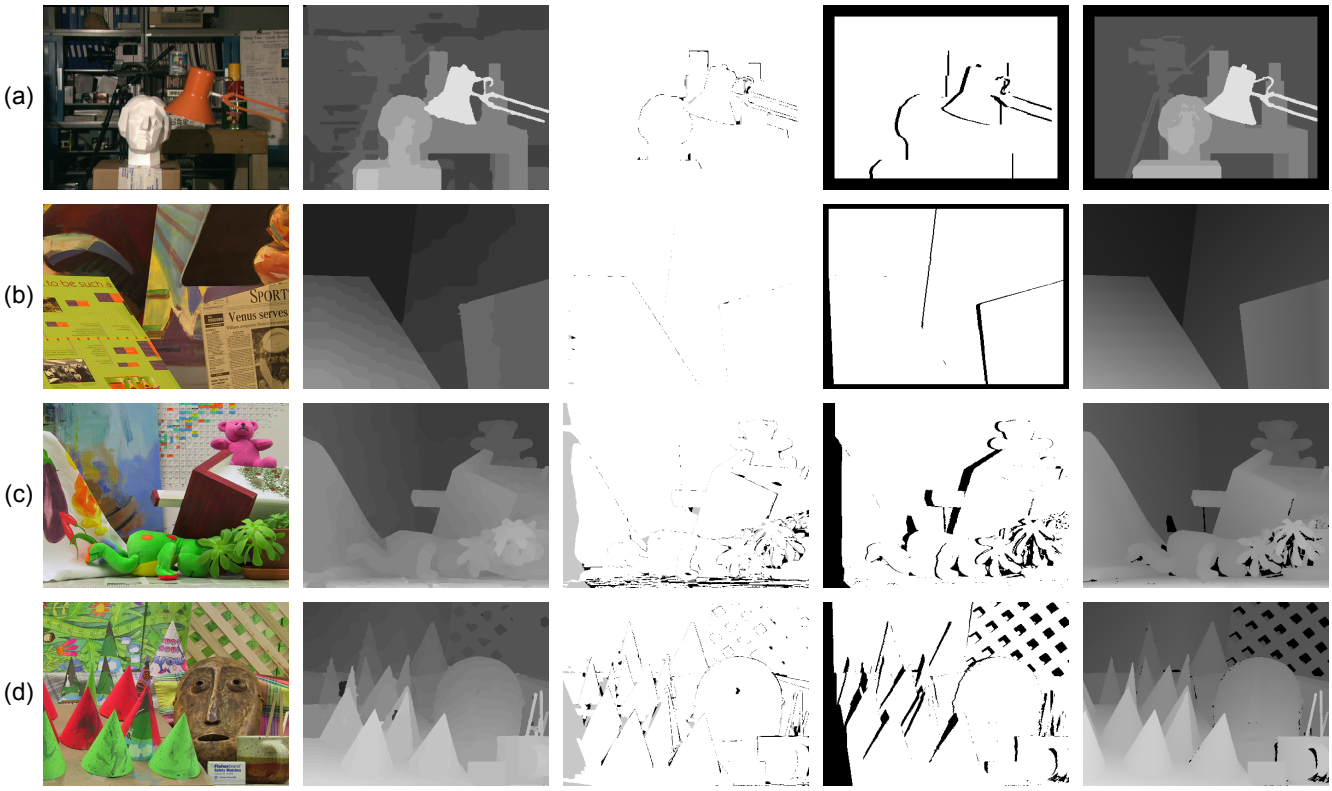

Figure 10. Evaluation results on the Middlebury standard data. (**a**) Tsukuba; (**b**) Venus; (**c**) Teddy; (**d**) Cones. Each row contains (from left to right): the left image, our results, the error map (error matching pixels whose absolute disparity errors are larger than one in non-occlusion and occlusion regions are marked in black and gray), the occlusion map (occluded pixels are marked black) and the ground truth map.

Error Threshold = 1 (Error Threshold...) — Sort by nonocc — Sort by all — Sort by disc

Algorithm	Avg. Rank	Tsukuba (ground truth) nonocc	Tsukuba all	Tsukuba disc	Venus (ground truth) nonocc	Venus all	Venus disc	Teddy (ground truth) nonocc	Teddy all	Teddy disc	Cones (ground truth) nonocc	Cones all	Cones disc	Average Percent Bad Pixels
YOUR METHOD	10.5	0.79 2	1.21 4	4.30 2	0.10 7	0.21 12	1.27 9	3.49 15	9.04 31	10.9 17	2.06 5	7.05 18	5.80 4	3.85
IGSM [158]	10.8	0.93 11	1.37 13	5.05 13	0.07 3	0.17 4	1.04 3	4.08 20	5.98 8	11.4 21	2.14 10	6.97 15	6.27 9	3.79
TSGO [142]	13.0	0.87 5	**1.13** 1	4.66 7	0.11 10	0.24 13	1.47 13	5.61 43	8.09 19	13.8 36	1.67 3	6.16 3	4.95 3	4.06
LCU [155]	13.0	1.06 20	1.34 9	5.50 18	0.07 2	0.26 19	1.03 2	3.68 17	9.95 37	10.4 15	1.63 2	6.87 13	4.82 2	3.89
JSOSP+GCP [150]	14.8	**0.74** 1	1.34 10	**3.98** 1	0.08 4	**0.16** 1	1.15 4	3.96 18	10.1 38	11.8 22	2.28 20	7.91 35	6.74 23	4.18
ADCensus [82]	18.0	1.07 23	1.48 20	5.73 26	0.09 5	0.25 16	1.15 4	4.10 21	6.22 9	10.9 18	2.42 26	7.25 21	6.95 27	3.97
CoopRegion [39]	21.5	0.87 7	1.16 2	4.61 6	0.11 9	0.21 8	1.54 16	5.16 35	8.31 23	13.0 30	2.79 47	7.18 20	8.01 55	4.41
AdaptingBP [16]	21.9	1.11 26	1.37 12	5.79 28	0.10 8	0.21 11	1.44 12	4.22 23	7.06 17	11.8 23	2.48 30	7.92 37	7.32 36	4.23
CCRADAR [151]	26.2	1.15 28	1.42 18	6.23 41	0.15 21	0.27 20	1.89 26	5.39 38	10.6 43	14.7 47	2.01 4	7.37 23	5.88 5	4.75

Figure 11. Middlebury results of our method. All numbers are the percentage of error pixels whose absolute disparity error is larger than one. The blue number is the ranking in every column. Our method outperforms the conventional stereo matching algorithms and ranks first among approximately 164 methods according to the average of the sum of the rankings in every column (up to 20 April 2015).

On the other hand, as shown in Figure 12, we presented some evaluation results of the Middlebury extension datasets [50,51] to illustrate the robustness of the proposed method. Meanwhile, we also show the quality of 3D reconstruction of Middlebury datasets using the pixels' image coordinate and their corresponding disparities in disparity space (see Figure 13). The evaluation results in Figure 12 and Figure 13 illustrate that our method is robust to different types of scenes and outperforms in slanted and highly curved surfaces.

Besides, we also compared our results with those produced by other "fused" schemes [20,23,52–56], and the compared results are listed in the Table 3. Our method provides an error rate of 2.61% on the Middlebury datasets, compared to the average error rate 3.27% of the previous state-of-the-art "fused" methods. It is clear that our method performs almost 20% better than other "fused" scheme-based algorithms in the aspect of precision. Furthermore, As shown in Figures 14 and 15, our method achieves comparable results in the following aspects:

- Noise and outliers are significantly reduced, mainly because of the improved luminance consistency term and the texture term.
- The method obtains precise disparities for slanted or highly-curved surfaces of objects with complex geometric characteristics, mainly because of the 3D plane bias term.
- Ambiguous matchings caused by over-segmentation or under-segmentation are overcome and disparity variances become smoother, mainly because of the hybrid smoothness term.

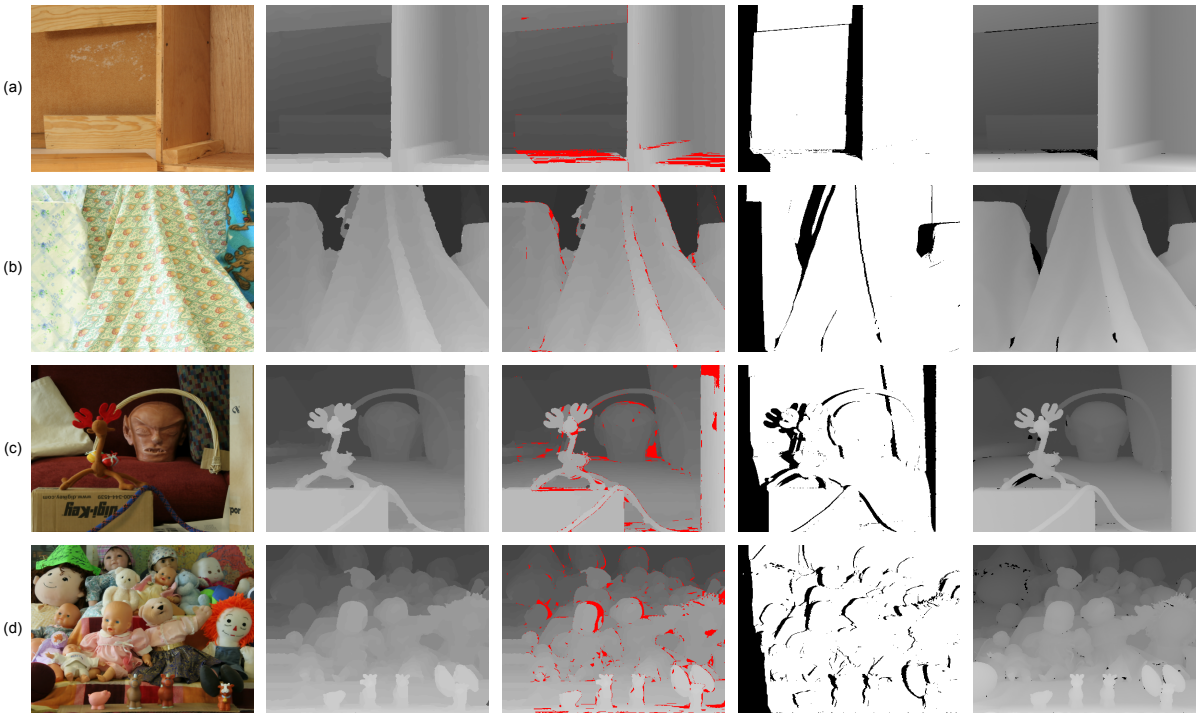

Figure 12. Evaluation results on the Middlebury extension datasets. (**a**) Wood1; (**b**) Cloth4; (**c**) Reindeer; (**d**) Dolls. Each row contains (left to right): the left image, our results, the error map (from top to bottom, the percentages of error pixels with absolute disparity error larger than one in non-occlusion regions are: 4.35%, 1.03%, 4.31%, 4.78%; error pixels are marked red), the occlusion map (occluded pixels are marked black) and the ground truth map.

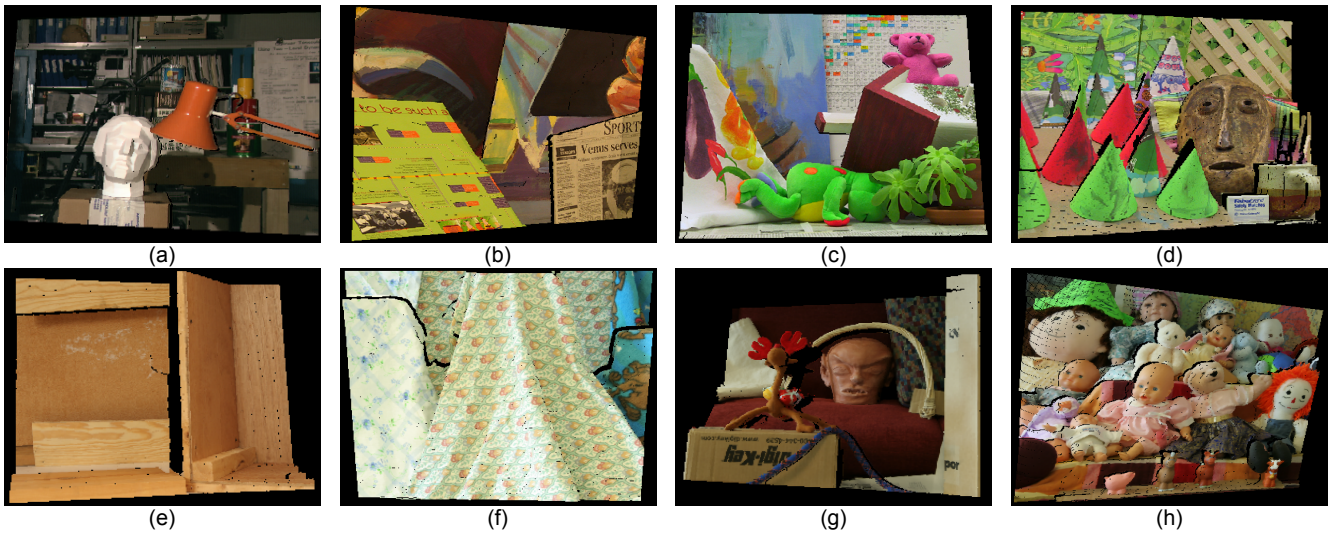

Figure 13. The results of 3D reconstructions. (**a**) Tsukuba; (**b**) Venus; (**c**) Teddy; (**d**) Cones; (**e**) Wood1; (**f**) Cloth4; (**g**) Reindeer; (**h**) Dolls.

Table 3. The percentages of error pixels (absolute disparity error larger than 1 in non-occlusion regions) of our method and other "fused" methods on the Middlebury datasets. "Averages" are the average percentages of error pixels over all images. Compared to the average error rate 3.27% of the previous state-of-the-art "fused" methods, our method provides a lower average error rate of 2.61% on the Middlebury datasets. It performs almost 20% better than other "fused" methods in the aspect of precision.

	The Percentages of Error Pixels (%)								
	Tsukuba	**Venus**	**Teddy**	**Conse**	**Wood1**	**Colth4**	**Reimdeor**	**Dools**	**Averages**
Zhu *et al.* [20]	1.16	0.14	2.83	3.47	5.38	3.74	5.83	5.46	3.50
Wang *et al.* [23]	0.89	0.12	6.39	2.14	4.05	3.81	3.55	2.71	2.96
Yang *et al.* [52]	0.94	0.26	5.65	7.18	1.76	2.60	4.43	4.13	3.37
Jaesik *et al.* [53,55]	2.38	0.56	5.59	6.28	3.72	2.88	4.04	4.69	3.77
James *et al.* [54]	2.90	0.29	2.12	2.83	2.74	2.32	5.02	4.02	2.78
Ours	0.79	0.10	3.49	2.06	4.35	1.03	4.31	4.78	2.61

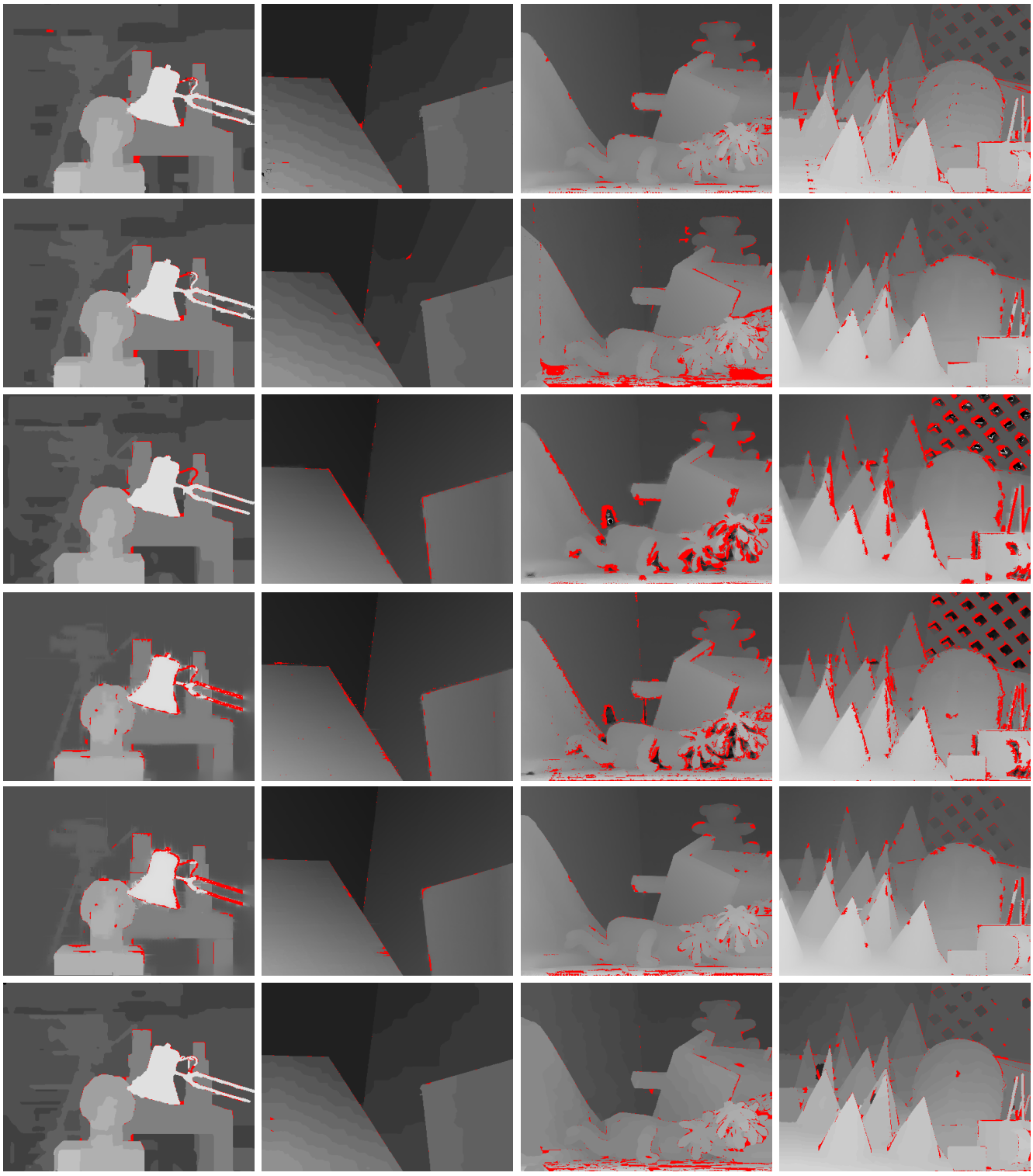

Figure 14. Evaluation results with the state-of-the-art "fused" scheme-based algorithms on the Middlebury datasets. (**a**) Tsukuba; (**b**) Venus; (**c**) Teddy; (**d**) Cones. Each column from top to bottom is the results obtained from: Zhu *et al.* [20], Wang *et al.* [23], Yang *et al.* [52], Jaesik *et al.* [53,55], James *et al.* [54] and our method. Error pixels with absolute disparity error larger than one in non-occlusion regions are marked red. The percentages of error pixels are listed in Table 3.

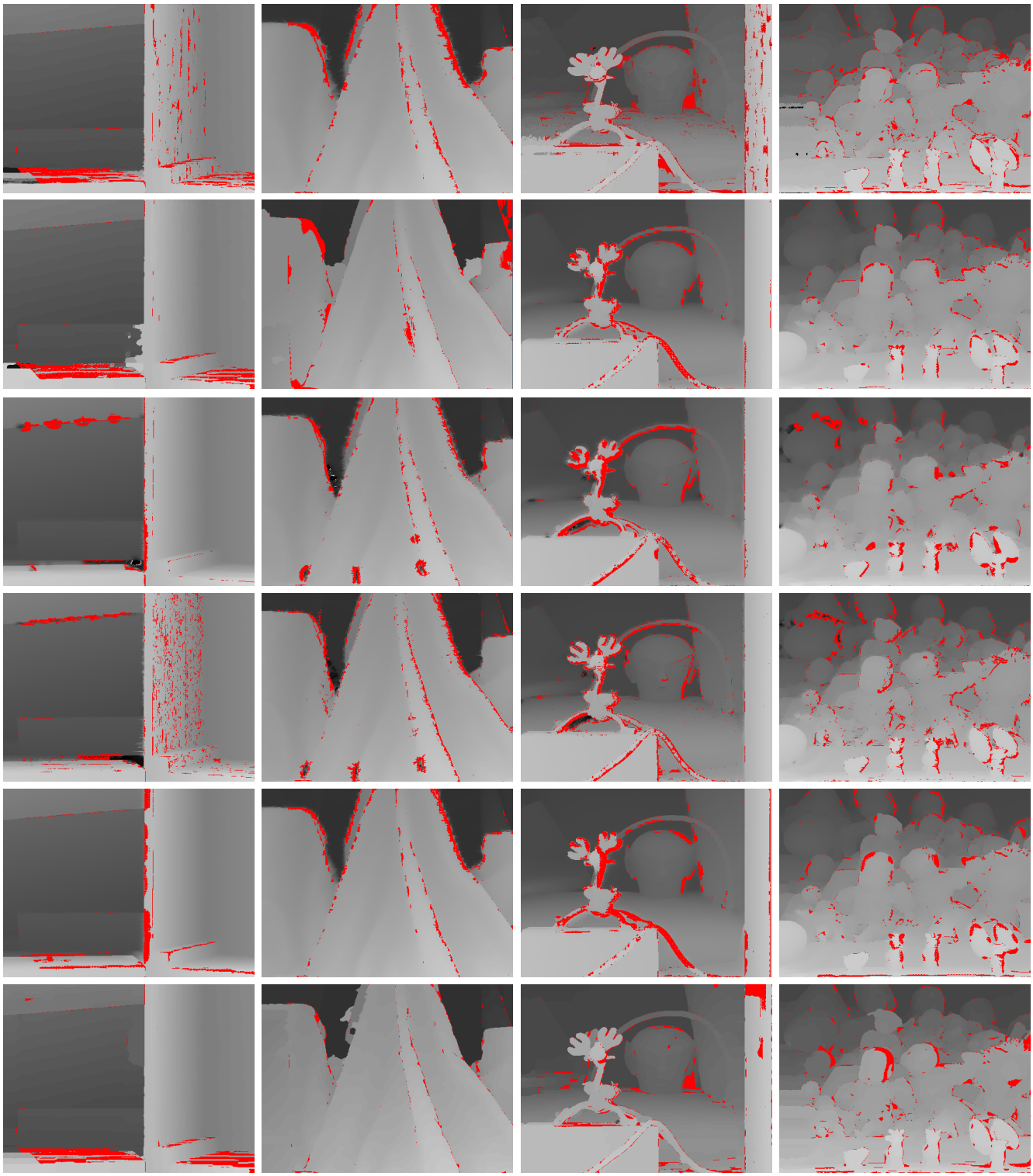

Figure 15. Evaluation results with the state-of-the-art "fused" scheme-based algorithms on the Middlebury extension datasets. (**a**) Wood1; (**b**) Cloth4; (**c**) Reindeer; (**d**) Dolls. Each column from top to bottom is the results obtained from: Zhu *et al.* [20], Wang *et al.* [23], Yang *et al.* [52], Jaesik *et al.* [53,55], James *et al.* [54] and our method. Error pixels with absolute disparity error larger than one in non-occlusion regions are marked red. The percentages of error pixels are listed in Table 3.

4.3. Evaluation Results for Each Term

We conducted evaluations to analyze the effect of the individual terms in Equation (4). In each experiment, one term was turned off and the others remained on. First, the texture term was turned off, which meant that the range of the potential disparities for each pixel was no longer restricted by the texture variance and gradient. Ambiguities occurred in textureless and repetitive texture regions without the prior restriction from the data of the depth sensor (see the yellow rectangle in Figure 16b). The average error rate of all images in non-occlusion regions sharply increased to 2.77%. Furthermore, the improved luminance consistency term was turned off by setting $w_p^x := 0$. Then, this term can be viewed as the conventional one that is easily affected by light variation and causes error matching on the non-Lambertian surface and rich texture regions (see the green rectangle regions in Figure 16c–e). The corresponding average error rate of all images in the non-occlusion regions is 2.37%. Thirdly, the hybrid smoothness term was turned out by replacing by the usual second-order smoothness term [11]. Some artifacts in the red rectangle in Figure 16f were caused by over-segmentation and under-segmentation. Its average error rate sharply increased to 3.02%. Finally, the 3D plane bias term was turned off by setting λ_p:=0. In that case, all 3D object surfaces are assumed as the frontal parallel ones, and the depth map is rather noisy, which makes it difficult to preserve the details at the boundary of objects (see the blue rectangle region in Figure 16g). Its average error rate is 2.14%. The corresponding error statistic analysis on the Middlebury datasets is listed in Table 4. It is clear that our method can obtain the lowest average error rate when all terms turn on (average error rate of 1.61% in non-occlusion regions).

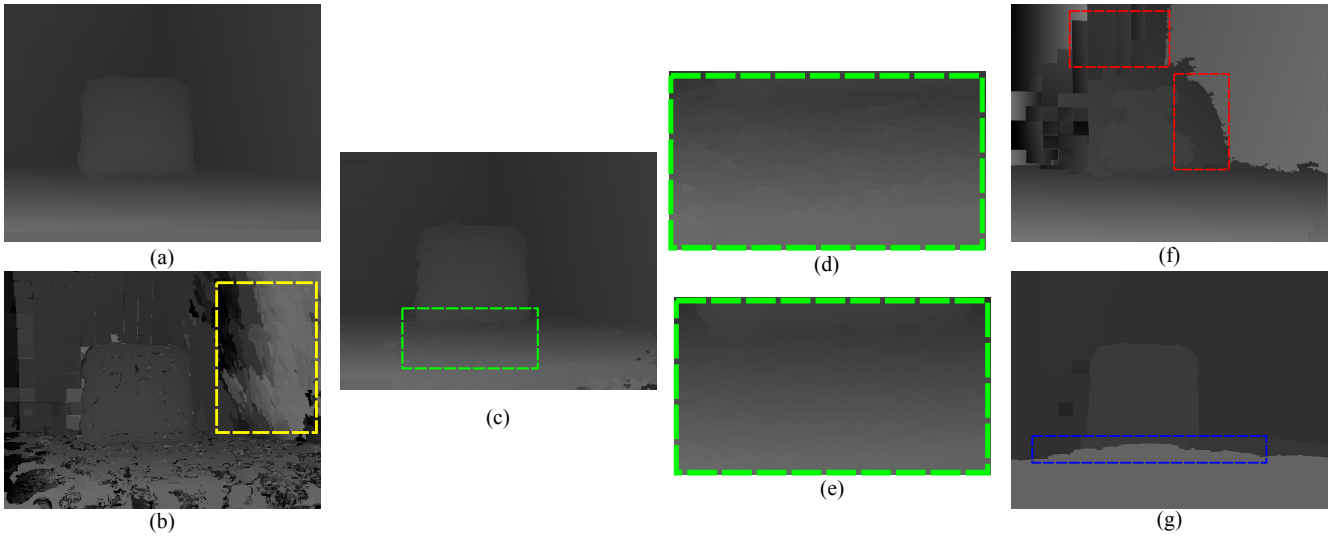

Figure 16. Evaluation results when turning off some terms. (**a**) Our result; (**b**) result without the texture term; (**c**) result without the improved luminance consistency term; (**d**) the detail with an enlarged scale in the green region of (c); (**e**) our result detail with enlarged scale in the same green region of (c); (**f**) result without the hybrid smoothness term; (**g**) result without the 3D plane bias term. Nonocc: non-occlusion regions.

Table 4. Error statistic for the Middlebury datasets with different constraint terms turned off. "Averages" are the average percentages of error pixels over all images in different regions.

	Tsukuba			Venus			Teddy			Cones			Averages		
	nonocc	**all**	**disc**	**nonocc**	**all**	**disc**	**nonocc**	**all**	**disc**	**nonocc**	**all**	**disc**	**nonocc**	**all**	**disc**
Texture term off	2.04	2.12	5.78	0.81	1.03	3.09	5.43	10.2	13.25	2.79	8.26	6.35	2.77	5.40	7.12
Luminance term off	1.01	1.65	4.78	0.11	0.25	1.54	5.49	11.20	14.92	2.88	8.47	7.74	2.37	5.39	7.25
Smoothness term off	1.39	2.14	4.94	0.85	0.93	2.02	6.57	13.01	14.80	3.28	7.50	7.13	3.02	5.90	7.22
Plane bias term off	0.88	1.49	4.86	0.23	0.65	2.27	4.53	9.30	10.63	2.90	7.96	8.96	2.14	4.76	6.68
All terms on	0.79	1.21	4.30	0.10	0.21	1.27	3.49	9.04	10.90	2.06	7.05	5.80	1.61	4.37	5.56

4.4. Computational Time Analyses

The proposed method was implemented on a PC with Core i5-2500 3.30 GHZ CPU and 4 GB RAM. Tables 5 and 6 list the running time of the proposed method for all experiments. It is obvious that the computational time is proportional to the image resolution and the scope of potential disparities. For example, it took approximately 1–9 mins to obtain results on Middlebury data and 19–25 mins on the real-world scene datasets. In the future, we aim to implement our method on a GPU to achieve a good balance between accuracy and efficiency.

Table 5. Running times for the real-world datasets. The disparity map resolution of all real-world datasets is 1024×960. The corresponding maximum disparity is 107.

	Dragon	**Book**	**Plant**	**Tablecloth**	**Board**	**Box**	**Kola**	**Vase**	**Piggy**	**Dragon and Piggy**
Running Time (m):	22.15	20.54	22.16	24.31	21.51	24.26	23.43	22.30	19.44	23.04

Table 6. Running times for the Middlebury datasets.

	Tsukuba	**Venus**	**Teddy**	**Cones**	**Wood1**	**Cloth4**	**Reindeer**	**Dolls**
Running Time (m):	1.03	1.19	4.08	4.57	8.18	7.44	7.02	8.31
Disparity Map Resolution:	384×288	434×383	450×375	450×375	457×370	433×375	447×370	463×375
Maximum Disparity:	15	19	59	59	71	69	67	73

5. Conclusions

In this paper, we present an accurate disparity estimation fusion model that "fused" the advantages of the complementary nature of active and passive sensors. Our main contributions are the texture information constraint and the multiscale pseudo two-layer image model. The comparison results show that our method can reduce the error estimate caused by under- or over- segmentation and has good performance in keeping object boundaries compared to using the conventional stereo matching or the depth sensor alone. Furthermore, the proposed method provides an error rate of 2.61% on the Middlebury datasets, compared to the average error rate 3.27% of the previous state-of-the-art "fused" methods. It is clear that our method performs almost 20% better than other "fused" scheme-based algorithms in

the aspect of precision. In the future, we will investigate a more accurate method for estimating the disparities of occluded pixels. We also intend to transform our method to a parallel GPU implementation.

Acknowledgments

This work is supported and funded by the National Natural Science Foundation of China (No. 61300131), the National Key Technology Research and Development Program of China (No. 2013BAK03B07) and the National High Technology Research and Development Program of China (863 Program) (No. 2013AA013902).

Author Contributions

All authors contributed to the understanding of the algorithm. Jing Liu and Zhaoqi Wang conceived of and designed the algorithm. Jing Liu and Xuefeng Fan performed the simulation analysis and the experiments. Jing Liu and Chunpeng Li wrote the paper. All authors read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Díaz-Vilariño, L.; Khoshelham, K.; Martínez-Sánchez, J.; Arias, P. 3D modeling of building indoor spaces and closed doors from imagery and point clouds. *Sensors* **2015**, *15*, 3491–3512.
2. Pan, S.; Shi, L.; Guo, S. A kinect-based real-time compressive tracking prototype system for amphibious spherical robots. *Sensors* **2015**, *15*, 8232–8252.
3. Yebes, J.J.; Bergasa, L.M.; García-Garrido, M. Visual object recognition with 3D-aware features in KITTI urban scenes. *Sensors* **2015**, *15*, 9228–9250.
4. Tanimoto, M.; Tehrani, M.P.; Fujii, T.; Yendo, T. Free-viewpoint TV. *IEEE Signal Process. Mag.* **2011**, *28*, 67–76.
5. Liu, J.; Li, C.; Mei, F.; Wang, Z. 3D entity-based stereo matching with ground control points and joint second-order smoothness prior. *Vis. Comput.* **2014**, *31*, 1–17.
6. ASUS Xtion. Available online: www.asus.com/Multimedia/Xtion/ (accessed on 19 August 2015).
7. Microsoft Kinect. Available online: www.microsoft.com/zh-cn/kinectforwindows/ (accessed on 19 August 2015).
8. Song, X.; Zhong, F.; Wang, Y.; Qin, X. Estimation of kinect depth confidence through self-training. *Vis. Comput.* **2014**, *30*, 855–865.
9. Scharstein, D.; Szeliski, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis.* **2002**, *47*, 7–42.
10. Yoon, K.J.; Kweon, I.S. Adaptive support-weight approach for correspondence search. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 650–656.

11. Woodford, O.; Torr, P.; Reid, I.; Fitzgibbon, A. Global stereo reconstruction under second-order smoothness priors. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 2115–2128.
12. Bleyer, M.; Rhemann, C.; Rother, C. Patch match stereo-stereo matching with slanted support windows. *BMVC* **2011**, *11*, 1–11.
13. Mei, X.; Sun, X.; Dong, W.; Wang, H.; Zhang, X. Segment-tree based Cost Aggregation for Stereo Matching. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 313–320.
14. Wang, L.; Yang, R. Global Stereo Matching Leveraged by Sparse Ground Control Points. In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, Providence, RI, USA, 20–25 June 2011; pp. 3033–3040.
15. LiDAR. Available online: www.lidarusa.com (accessed on 19 August 2015).
16. 3DV Systems. Available online: www.3dvsystems.com (accessed on 19 August 2015).
17. Photonix Mixer Device for Distance Measurement. Available online: www.pmdtec.com (accessed on 19 August 2015).
18. Khoshelham, K.; Elberink, S.O. Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors* **2012**, *12*, 1437–1454.
19. Zhu, J.; Wang, L.; Gao, J.; Yang, R. Spatial-temporal fusion for high accuracy depth maps using dynamic MRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 899–909.
20. Zhu, J.; Wang, L.; Yang, R.; Davis, J.E.; Pan, Z. Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1400–1414.
21. Yang, Q.; Tan, K.H.; Culbertson, B.; Apostolopoulos, J. Fusion of Active and Passive Sensors for Fast 3D Capture. In Proceedings of the IEEE International Workshop on Multimedia Signal Processing, Saint Malo, France, 4–6 October 2010; pp. 69–74.
22. Zhang, S.; Wang, C.; Chan, S. A New High Resolution Depth Map Estimation System Using Stereo Vision and Depth Sensing Device. In Proceedings of the IEEE 9th International Colloquium on Signal Processing and its Applications, Kuala Lumpur, Malaysia, 8–10 March 2013; pp. 49–53.
23. Wang, Y.; Jia, Y. A fusion framework of stereo vision and Kinect for high-quality dense depth maps. *Comput. Vis.* **2013**, *7729*, 109–120.
24. Somanath, G.; Cohen, S.; Price, B.; Kambhamettu, C. Stereo Kinect for High Resolution Stereo Correspondences. In Proceedings of the IEEE International Conference on 3DTV-Conference, Seattle, Washington, DC, USA, 29 June–1 July 2013; pp. 9–16.
25. Zhang, Z. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1330–1334.
26. Herrera, C.; Kannala, J.; Heikkilä, J. Joint depth and color camera calibration with distortion correction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2058–2064.
27. Yang, Q. A Non-Local Cost Aggregation Method for Stereo Matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1402–1409.
28. Christoudias, C.M.; Georgescu, B.; Meer, P. Synergism in low level vision. *IEEE Intern. Conf. Pattern Recognit.* **2002**, *4*, 150–155.

29. Bleyer, M.; Rother, C.; Kohli, P. Surface Stereo with Soft Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 1570–1577.
30. Wei, Y.; Quan, L. Asymmetrical occlusion handling using graph cut for multi-view stereo. *IEEE Intern. Conf. Pattern Recognit.* **2005**, *2*, 902–909.
31. Hu, X.; Mordohai, P. A quantitative evaluation of confidence measures for stereo vision. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2121–2133.
32. Liu, Z.; Han, Z.; Ye, Q.; Jiao, J. A New Segment-Based Algorithm for Stereo Matching. In Proceedings of the IEEE International Conference on Mechatronics and Automation, Changchun, China, 9–12 August 2009; pp. 999–1003.
33. Rao, A.; Srihari, R.K.; Zhang, Z. Spatial color histograms for content-based image retrieval. In Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence, Chicago, IL, USA, 9–11 November 1999; pp. 183–186.
34. Lempitsky, V.; Rother, C.; Blake, A. Logcut-Efficient Graph Cut Optimization for Markov Random Fields. In Proceedings of the IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8.
35. Boykov, Y.; Veksler, O.; Zabih, R. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 1222–1239.
36. Kolmogorov, V.; Zabin, R. What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 147–159.
37. Kolmogorov, V.; Rother, C. Minimizing nonsubmodular functions with graph cuts-a review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1274–1279.
38. Boros, E.; Hammer, P.L.; Tavares, G. *Preprocessing of Unconstrained Quadratic Binary Optimization*; Technical Report RRR 10-2006, RUTCOR Research Report; Rutgers University: Piscataway, NJ, USA, 2006.
39. Lempitsky, V.; Roth, S.; Rother, C. FusionFlow: Discrete-Continuous Optimization for Optical Flow Estimation. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
40. Ishikawa, H. Higher-Order Clique Reduction Without Auxiliary Variables. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1362–1369.
41. Yang, Q.; Wang, L.; Yang, R.; Stewénius, H.; Nistér, D. Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 492–504.
42. Heo, Y.S.; Lee, K.M.; Lee, S.U. Joint depth map and color consistency estimation for stereo images with different illuminations and cameras. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1094–1106.
43. Matsuo, T.; Fukushima, N.; Ishibashi, Y. Weighted Joint Bilateral Filter with Slope Depth Compensation Filter for Depth Map Refinement. *VISAPP* **2013**, *2*, 300–309.

44. Rhemann, C.; Hosni, A.; Bleyer, M.; Rother, C.; Gelautz, M. Fast Cost-volume Filtering for Visual Correspondence and Beyond. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 20–25 June 2011; pp. 3017–3024.
45. Scharstein, D.; Szeliski, R. High-accuracy stereo depth maps using structured light. *IEEE Comput. Vis. Pattern Recognit.* **2013**, *1*, 195–202.
46. Chakrabarti, A.; Xiong, Y.; Gortler, S.J.; Zickler, T. Low-level vision by consensus in a spatial hierarchy of regions. **2014**, arXiv:1411.4894.
47. Lee, S.; Lee, J.H.; Lim, J.; Suh, I.H. Robust stereo matching using adaptive random walk with restart algorithm. *Image Vis. Comput.* **2015**, *37*, 1–11.
48. Spangenberg, R.; Langner, T.; Adfeldt, S.; Rojas, R. Large Scale Semi-Global Matching on the CPU. In Proceedings of the IEEE Intelligent Vehicles Symposium Proceedings, Dearborn, MI, USA, 8–11 June 2014; pp. 195–201.
49. Middlebury Benchmark. Available online: vision.middlebury.edu/stereo/ (accessed on 19 August 2015).
50. Hirschmuller, H.; Scharstein, D. Evaluation of Cost Functions for Stereo Matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
51. Scharstein, D.; Pal, C. Learning Conditional Random Fields for Stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
52. 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, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
53. Park, J.; Kim, H.; Tai, Y.W.; Brown, M.S.; Kweon, I.S. High-Quality Depth Map Upsampling and Completion for RGB-D Cameras. *IEEE Trans Image Process.* **2014**, *23*, 5559–5572.
54. Diebel, J.; Thrun, S. An application of markov random fields to range sensing. *NIPS* **2005**, *5*, 291–298.
55. Park, J.; Kim, H.; Tai, Y.W.; Brown, M.S.; Kweon, I. High Quality Depth Map Upsampling for 3D-TOF Cameras. In Proceedings of the IEEE International Conference on Computer Vision, Barcelona , Spain, 6–13 November 2011; pp. 1623–1630.
56. Huhle, B.; Schairer, T.; Jenke, P.; Straßer, W. Fusion of range and color images for denoising and resolution enhancement with a non-local filter. *Comput. Vis. image Underst.* **2010**, *114*, 1336–1345.