Computer Vision and Image Understanding
journal homepage: www.elsevier.com

As-Planar-As-Possible Depth Map Estimation

Min-Gyu Park[a], Kuk-Jin Yoon[b,**]

[a] *Korea Electronics Technology Institute (KETI), Seongnam-si, South Korea*
[b] *Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea*

ABSTRACT

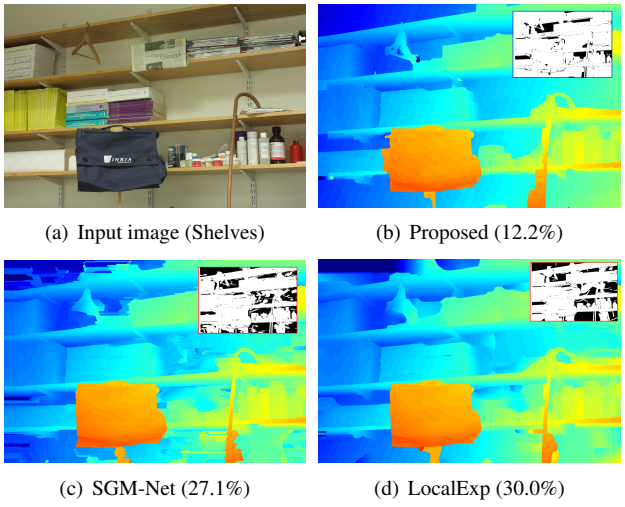
We propose an approach to compute dense disparity maps that takes the characteristics of man-made environments into account. The key contribution is to generate a piecewise planar disparity map while preventing the oversimplification problem in non-planar regions. To achieve this, we decompose the stereo matching problem into three sequential subproblems: initial disparity map estimation, plane hypotheses generation, and global optimization with plane hypotheses. After finding an initial disparity map, we find local and global plane hypotheses from the disparity map through segmentation-based local plane fitting, agglomerative hierarchical clustering, and energy-based multi-model fitting techniques. We then estimate a disparity map that is a mixture of over-parameterized and scalar disparity values while identifying unreliable pixels in an energy minimization framework; disparity values in planar regions are parameterized as a plane, disparity values in non-planar regions are represented as scalar, and unreliable pixels are marked as outliers. As a post-processing step, we perturb assigned plane parameters as well as scalar disparity values. We experimentally verify the proposed method using publicly available benchmarks and various stereo matching algorithms.

1. Introduction

The 3D reconstruction of man-made environments is an important computer vision research topic, owing to its wide range of applications. Man-made environments, especially indoor and architectural scenes, usually exhibit a high degree of structural regularity (Furukawa et al. (2009); Schindler and Dellaert (2004); Gallup et al. (2007); Straub et al. (2014)) owing to their axis-aligned geometry and planar scene structures; accordingly, these characteristics have been used to constrain the stereo matching problem (Furukawa et al. (2009); Gallup et al. (2007)). However, these assumptions can be easily violated in more complicated environments, and can lead to oversimplified results. Ironically, however, if we do not explicitly constrain the stereo problem, the characteristics of man-made environments, such as large homogeneous regions and non-fronto-parallel surfaces, may cause various stereo ambiguities. Therefore, a number of algorithms address these two problems in the context of matching cost computation (Bleyer et al. (2011)) and global optimization (Hirschmüller (2008); Taniai et al. (2017); Li et al. (2015)) to estimate accurate disparity maps while preserving detailed scene structures. However, existing algorithms begin to fail as the degree of ambiguity increases, *e.g.*, nearly constant intensities over a large region. For example, two state-of-the-art methods SGM-Net (Seki and Pollefeys (2017)) and LocalExp (Taniai et al. (2017)) generate erroneous results for a challenging dataset (Scharstein et al. (2014)), as shown in Figs. 1(c) and 1(d).

In this regard, we primarily focus on estimating a disparity map in a highly ambiguous man-made environment. Our approach is straightforward—because depth maps captured in man-made environments usually contain a large number of planar regions, the estimation of plane hypotheses can effectively improve stereo matching performance as long as oversimplification is avoided in non-planar regions. In particular, to handle largely ambiguous regions, we extract dominant planar structures in the scene and exploit them as global reconstruction cues (Gallup et al. (2007); Hadfield and Bowden (2015)) instead of having strong priors on the scene, *e.g.*, axis-aligned geometry (Furukawa et al. (2009)). At the same time, we consider non-planar regions and occluded pixels in the global energy minimization framework, in order to selectively approximate disparity values as planes and to identify pixels that cannot

**Corresponding author:
e-mail: kjyoon@kaist.ac.kr (Kuk-Jin Yoon)



(a) Input image (Shelves) (b) Proposed (12.2%)

(c) SGM-Net (27.1%) (d) LocalExp (30.0%)

Fig. 1. A comparison of disparity maps for Shelves dataset. Although three methods use deep learning-based matching costs to compare image patches, (c) and (d) suffer from poorly textured regions. However, with the aid of global planes, the proposed method generates a significantly better result than the state-of-the-art methods. The percentages indicate the bad pixel rates of the disparity maps at the threshold value of 2.0px. Error maps are shown in the top-right corner of each disparity map.

be matched reliably.

To be more specific, we treat the stereo matching problem as a series of subproblems, as described in Fig. 2. First, we estimate an initial disparity map; in our framework, any existing stereo algorithm can be adopted to perform this task. We utilize state-of-the-art methods as well as popularly used methods to compute initial disparity maps, in order to show the proposed methods robustness to different initial methods. Second, we detect local and global plane hypotheses through segment-based local plane fitting, agglomerative hierarchical clustering (Mllner (2013)), and energy-based multi-model fitting (Isack and Boykov (2012)). Hierarchical clustering reduces redundant planes from initially extracted plane hypotheses and the energy-based fitting gives plane hypotheses that best describe the scene with the minimum number of planes. This set of minimum planes are used as global plane hypotheses. Third, we estimate a disparity map as a mixture of over-parameterized disparity, scalar disparity, and outlier pixels in a pairwise Markov random field using graph cuts (Boykov et al. (2001); Delong et al. (2012)) with local and global reconstruction cues. Finally, we perturb assigned plane hypotheses to better align plane hypotheses with scene structures.

The remainder of the paper is organized as follows. A literature review is given in Sec. 2 and the proposed method is explained in Sec. 3. In Sec. 4, we experimentally verified the performance of the proposed method from various aspects, *e.g.*, the dependency on initial disparity maps and conclude the paper in Sec. 5.

2. Related Work

We review various stereo algorithms that handle the difficulties in man-made environments, *e.g.*, non-fronto-surfaces and homogeneity, and discuss recent trends in stereo matching.

To handle the planarity and non-fronto-surfaces found in many scenes, numerous stereo algorithms have been developed. We divide previous studies into two groups; the first group explicitly finds plane parameters for each pixel or region and the second group implicitly handle slanted surfaces in the global minimization framework. The first group, can be subdivided into three approaches: segmentation-based approach, PatchMatch-based approach, and plane-sweeping approach, though more than two approaches can be combined together, *e.g.*, PatchMatch-based matching costs with segmentation-based global optimization (Li et al. (2017b)). The segmentation-based stereo matching approach has long been studied since the early work of Birchfield and Tomasi (1999). In general, the segmentation-based approach segments images into multiple non-overlapping regions and assign plane parameters to each region (Birchfield and Tomasi (1999); Hong and Chen (2004); Klaus et al. (2006); Yamaguchi et al. (2014); Wang and Zheng (2008)), and then, segment and depth information is iteratively merged or refined through various optimization techniques such as graph cuts (Hong and Chen (2004)), belief propagation (Klaus et al. (2006)), and cooperative optimization (Wang and Zheng (2008)). One crucial drawback of the segmentation-based approach is that it depends on the quality of initial disparity maps and segmentation. To avoid the dependency on the initial disparity map, Muninder et al. (2014) proposed to assign plane hypotheses to each pixel rather than each region, assuming that the initial set of planes is the superset of the actual set of planes that describe the scene. The initial plane set is extracted from an oversegmented disparity map, and then, one of the planes is assigned to each pixel through cost volume filtering. This approach, however, initially extracts a large number of plane hypotheses owing to oversegmentation so that it is not suitable for high-resolution images.

Similarly, Bleyer et al. (2011) proposed a PatchMatch stereo algorithm that overparamerizes each pixel with a local disparity plane. Instead of extracting planes fromt the disparity map, they utilized the PatchMatch algorithm (Barnes et al. (2009)) to effectively search optimal planes. Afterward, several studies, Besse et al. (2012); Li et al. (2015, 2017b), tried to link the PatchMatch stereo algorithm to a global optimization algorithm, *i.e.*, belief propagation (Sun et al. (2003)), mainly focusing on handling a continuous label space. A recent work of Taniai et al. (2017) applies graph cuts to a continuous labeling problem by employing multiple local expansion moves to small grid regions. They differenciated candidate α-labels for each grid and propagated assigned labels for nearby regions to handle a large number of labels effectively. In addition, Li et al. (2017a) extended the conventional minimum spanning tree (MST)-based cost aggregation scheme to PatchMatch-based continuous 3D labels by introducing multiple MST structures and tree-level random search. These two studies show state-of-the-art results in the Middlebury benchmark (Scharstein et al. (2014)).

The plane-sweeping stereo of Gallup et al. (2007) aimed at reconstructing an urban environment in real-time assuming the captured scene consists of planes parallel to three orthogonal planes. Based on this assumption, they proposed a method to

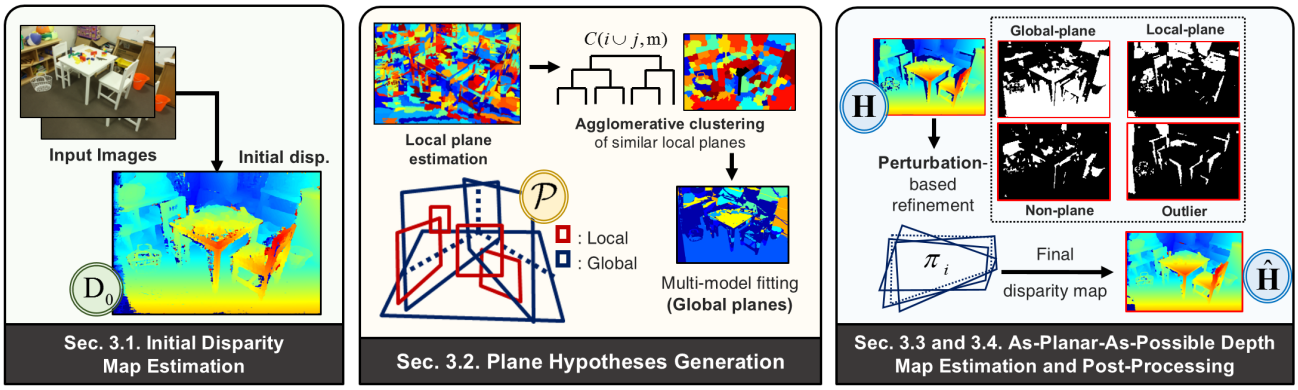


Fig. 2. Overview of the proposed method.

extract orthogonal planes from sparse correspondences. After inducing a set of plane hypothese that are parallel to orthogonal planes, they warped the target image by using the plane hypotheses. Finally, the disparity value of each pixel is determined by using one of the plane hypotheses, after comparing the reference image and warped images. Afterward, Gallup et al. (2010) extended this idea by handling non-planar regions such as bushes and trees from planar regions to avoid oversimplification. To acheive this, they trained a supervised classifier to distinguish planar and non-planar regions. A more recent work of Sinha et al. (2014) computes hundreds of local sweeping directions from a set of sparse correspondences, in order to handle more complex scenarios. Häne et al. (2014) modified conventional plane-sweeping stereo for general cameras such as fisheye and omnidirectional cameras through an adaptation of camera projection models.

Similar to plane-sweeping stereo, Manhattan world stereo (Furukawa et al. (2009)) assumes axis-alighed geometry (Coughlan and Yuille (2000)). They found axis-aligned plane hypotheses from an oriented point cloud, and then, assigned one of the planes to each pixel through graph cuts. The Atlanta world assumption (Schindler and Dellaert (2004)) and the mixture of Manhattan frames (Straub et al. (2014)) extend the Manhattan world assumption to more general scnearios, nevertheless, these assumptions were rarely used for scene reconstruction.

The fourth group, in contrast, does not compute plane parameters explicitly. The semi-global matching (SGM) method (Hirschmüller (2008)) is widely used in driving environments, owing to the methods simplicity and accuracy. To preserve slanted surfaces, SGM gives a small penalty to pixels having small disparity differences with its neighbors, otherwise, it gives a large penalty to pixels. Several studies advanced the conventional SGM method by reducing memory usage (Hirschmüller et al. (2012); Lee et al. (2018)) or by adopting deep learning techniques (Seki and Pollefeys (2017)). From the global optimization point of view, Woodford et al. (2009); Zhang et al. (2014a) tried to impose second-order smoothness priors to better preserve non-fronto-surfaces. Because the second order model is non-submodular, Woodford et al. (2009) solved the problem using the quadratic pseudo-Boolean optimization (QPBO) algorithm (Kolmogorov and Rother (2007)). Zhang et al. (2014a) utilized the Laplacian operator to impose pixel-wise second-order smoothness. From the local optimization point of view, Einecke and Eggert (2014) handled slanted surfaces in the matching cost aggregation step by using multiple local windows. They confirmed that a simple aggregation method can outperform complicated methods as long as the matching cost is carefully aggregated.

Because many recent studies employed deep learning techniques to compute disparity maps, we briefly review related stuides. Notably, Žbontar and LeCun (2015)'s work, also known as MC-CNN, confirmed the significance of a convolutional neural network (CNN). They trained a CNN to predict the similarity between two patches. Although they used existing aggregation methods to compute the disparity map, they reported accurate results in both indoor (Scharstein et al. (2014)) and outdoor benchmarks (Geiger et al. (2013)). Chen et al. (2015) simplified MC-CNN by replacing fully connected layers with a dot product operation that provides significantly faster performance than the original method. Luo et al. (2016) designed a dot product layer instead of computing feature vectors from the network. Park and Lee (2016) employed an additional pooling layer to take large image patches, *e.g.*, 37×37, as input. On the other hand, many researchers investigated end-to-end networks to predict disparity maps especially in driving environments. Notably, Mayer et al. (2016) introduced an end-to-end architecture to train disparity (DispNet) and optical flow (FlowNet). They initially trained counvolutional neural networks (CNNs) using large synthetic datasets and the networks were fine-tuned by using real data. Knöbelreiter et al. (2017) designed a network that combines CNN and CRF in a unified network based on the structured output support vector machine. Interestingly, Kendall et al. (2017); Liang et al. (2018) designed end-to-end networks while mimicking a conventional stereo matching pipeline. Kendall et al. (2017) constructed a 3D cost volume from deep unary features, and then, 3D convolution and soft argmin operation are carried out to compute a disparity map. Whereas Liang et al. (2018) decomposed the entire procedure into multi-scale feature extraction, disparity estimation, and disparity refinement steps. Chang and Chen (2018) extended the idea of Kendall et al. (2017) by adding the spa-

tial pooling module to compute deep unary features and by replacing the 3D convolution network with the stacked hourglass model (Newell et al. (2016)). Pang et al. (2017) proposed a two-stage network called cascade residual learning (CRL) where the first step computes a disparity map with DispNet with extra up-convolutions and the second step refines the disparity map based on residual signals across multiple scales.

Besides deep learning techniques, which mostly focus on computing accurate initial matching costs, several studies address the stereo matching problem from a different perspective. Güney and Geiger (2015) addressed transparent and reflective surfaces with specific object knowledge. To selectively refine disparity values of cars in the post-processing step, they aligned car CAD models after computing the disparity map. Hadfield and Bowden (2015) employed high-level scene cues (such as common configurations of surfaces and edge classes) in order to leverage stereo matching.

3. Proposed Method

We regard the stereo matching problem as a sequence of subproblems. First, we compute an initial disparity map $\mathbf{D}_0$ using an existing algorithm. We then generate two sets of plane hypotheses $\mathcal{P}$ from the disparity map. Afterward, we estimate the final disparity map $\mathbf{H}$. Here, $\mathbf{H}$ is the disparity map, which consists of over-parameterized disparity values (Besse et al. (2012); Klaus et al. (2006)), scalar disparity values, and outlier pixels. For example, the disparity value of a pixel can be defined by using plane parameters, *e.g.*, $d = ap_x + bp_y + c$, or can be defined as a scalar value, to avoid oversimplified results in non-planar regions.

3.1. Initial disparity map estimation

The primary goal of computing the initial disparity map is to extract plane hypotheses from the scene, assuming that the captured image contains a number of planar regions, *e.g.*, walls and desks. Because different initial disparity maps can yield different plane hypotheses, one may argue that this approach has dependency on the initial disparity map estimation algorithm. However, we also claim that the proposed method consistently improves the quality of initial disparity maps (even poor-quality maps) because of two reasons. First, the proposed method robustly extracts plane hypotheses in the global optimization framework. Second, once plane hypotheses are computed, we do not simply approximate the initial disparity map as a set of piecewise planar regions. Instead, we examine image patches with the aid of the computed plane hypotheses. Therefore, if the initial disparity map is accurate, only a few erroneous pixels are replaced with new disparity values, whereas a large number of pixels will be changed as the number of mismatched pixels increases.

To verify the independence of the initial disparity estimation step, we employ various stereo matching algorithms, including state-of-the-art algorithms (Žbontar and LeCun (2016); Taniai et al. (2017); Seki and Pollefeys (2017)) as well as popularly used algorithms (Hirschmüller (2008)), as the front-end step of the proposed method. Note that we do not specifically consider the characteristics of the employed stereo matching algorithm.

3.2. Plane hypotheses generation

Given an initial disparity map, we generate two sets of plane hypotheses, $\mathcal{P} = \{\mathcal{P}_{\text{local}}, \mathcal{P}_{\text{global}}\}$, where the subscripts indicate that they are local and global plane hypotheses. First, we segment the input image into a set of superpixels (Achanta et al. (2012)) $\mathcal{S} = \{\mathbf{s}_1, ..., \mathbf{s}_n\}$ and find n local plane hypotheses from each superpixel, where n is the number of superpixels. To compute a local plane hypothesis, we find inlier pixels using the RANSAC technique, and then fit a plane to inlier pixels for each superpixel in a least square manner. Therefore, each pixel in a superpixel has its corresponding local plane hypothesis, *i.e.*, pixels in a superpixel $\mathbf{s}_i$ share the same local plane hypotheses π_i. In general, local plane hypotheses do not describe the scene structure accurately, because the pixels in a segment do not always lie on a coplanar surface. Moreover, incorrect plane parameters can be estimated from a largely erroneous region, as shown in Fig. 1.

In this sense, we find global plane hypotheses in order to use them as a global reconstruction cue, and to handle largely ambiguous regions by minimizing the following energy function:

$$E(\mathcal{P}_{\text{global}}|\hat{\mathcal{P}}_{\text{local}}, \mathbf{D}_0) = \sum_{\mathbf{p}} U(\pi_i^\top \mathbf{p}, \mathbf{D}_0(\mathbf{p})) + \beta|\mathcal{L}_{\mathcal{P}}|, \quad (1)$$

which consists of the unary term and the label cost term (Isack and Boykov (2012)). β balances two terms. $\pi_i^\top \mathbf{p}$ is a disparity value in which $\pi_i = [a_i\ b_i\ c_i]^\top$ and $\mathbf{p} = [p_x\ p_y\ 1]^\top$. $D_0(\mathbf{p})$ is an initial disparity value at $\mathbf{p}$. Here, we use the clustered local plane hypotheses $\hat{\mathcal{P}}_{\text{local}}$ instead of $\mathcal{P}_{\text{local}}$, which is described at the end of this subsection. The unary term is defined as follows:

$$U(\pi_i^\top \mathbf{p}, \mathbf{D}_0(\mathbf{p})) = \begin{cases} |\pi_i^\top \mathbf{p} - \mathbf{D}_0(\mathbf{p})| & \text{if } \pi_i \neq \pi_\emptyset, \\ \gamma_d & \text{otherwise,} \end{cases} \quad (2)$$

where the unary term measures the discrepancy between the initial disparity value and the reconstructed disparity value $\pi_i^\top \mathbf{p}$. Here, we employ the null hypothesis $\pi_\emptyset$ to assign this label to pixels having noisy disparity values or in small planar regions, instead of assigning one of the plane hypotheses in $\hat{\mathcal{P}}_{\text{local}}$. The second term is the label count penalty (Isack and Boykov (2012)) for using the minimum number of plane hypotheses to describe the initial disparity map. $\mathcal{L}_{\mathcal{P}}$ is the set of distinct labels, *i.e.*, plane hypotheses, assigned to all pixels. Together with the null hypotheses and the label count penalty, the minimization of Eq. (1) assigns a set of plane hypotheses that describe the scene with the minimum number of plane hypotheses. Finally, we set $\mathcal{P}_{\text{global}}$ to $\mathcal{L}_{\mathcal{P}} \backslash \pi_\emptyset$. We solve this energy function using graph cuts that supports label costs (Delong et al. (2012)).

Local plane clustering: Because $\mathcal{P}_{\text{local}}$ contains a large number of redundant plane hypotheses, we merge similar local plane hypotheses through agglomerative hierarchical clustering (Mllner (2013)) based on the linear combination of two distance metrics,

$$\mathbf{C}(i, j) = \alpha \mathbf{C}_p(i, j) + (1 - \alpha)\mathbf{C}_c(i, j), \quad (3)$$

where $\mathbf{C}(i, j)$ measures the distance between two planes, *e.g.*, π_i and π_j. The first distance metric is the average squared difference of reconstructed disparity values,

$$\mathbf{C}_p(i, j) = \frac{1}{|\mathbf{s}_i \cup \mathbf{s}_j|} \sum_{\mathbf{p} \in \{\mathbf{s}_i \cup \mathbf{s}_j\}} (\pi_i^\top \mathbf{p} - \pi_j^\top \mathbf{p})^2, \quad (4)$$

where $\pi_i^\top \mathbf{p}$ and $\pi_j^\top \mathbf{p}$ are reconstructed disparity values using two plane parameters at the position of $\mathbf{p}$. $\mathbf{s}_i$ and $\mathbf{s}_j$ indicate superpixels in which pixels in $\mathbf{s}_i$ and $\mathbf{s}_j$ were used to compute π_i and π_j, respectively. Instead of directly comparing plane parameters, *e.g.*, an angle between normal vectors and the distance from the camera, we employ the residual of the disparity values, because this residual describes the geometric relationship between local plane hypotheses effectively with a single value. For example, the residual increases proportionally to the distance between two superpixels in the image coordinates (unless they are perfectly coplanar). The second metric utilizes the color or intensity difference between superpixels and is defined as

$$\mathbf{C}_c(i,j) = 1 - \sum_{c\in\{\mathrm{R,G,B}\}} \left(\frac{\min(h_c(i), h_c(j))}{\max(h_c(i), h_c(j))} \right), \tag{5}$$

where $h_c(i)$ is the normalized color or intensity histogram for pixels within superpixel $\mathbf{s}_i$. Based on the distance matrix, similar planes that have the smallest distance are merged into the same plane, and the distance matrix (*e.g.*, the distance between the merged plane and the other segments) is updated by the average formula (Mllner (2013)),

$$\mathbf{C}(i \cup j, m) = \frac{n_i \mathbf{C}(i,m) + n_j \mathbf{C}(j,m)}{n_i + n_j}, \tag{6}$$

where n_i is the number of planes that have been merged to the i^{th} segment and m is an index of other segments. This sequential merging procedure is repeated until the minimum distance value exceeds the cutoff value. The use of a cutoff value allows a varying number of segments depending on the structure of the scene, rather than fixing the number of clustered segments. After clustering, we estimate plane parameters again using RANSAC for each clustered segment and denote estimated plane hypotheses as $\hat{\mathcal{P}}_{\mathrm{local}}$.

3.3. Disparity map estimation with plane hypotheses

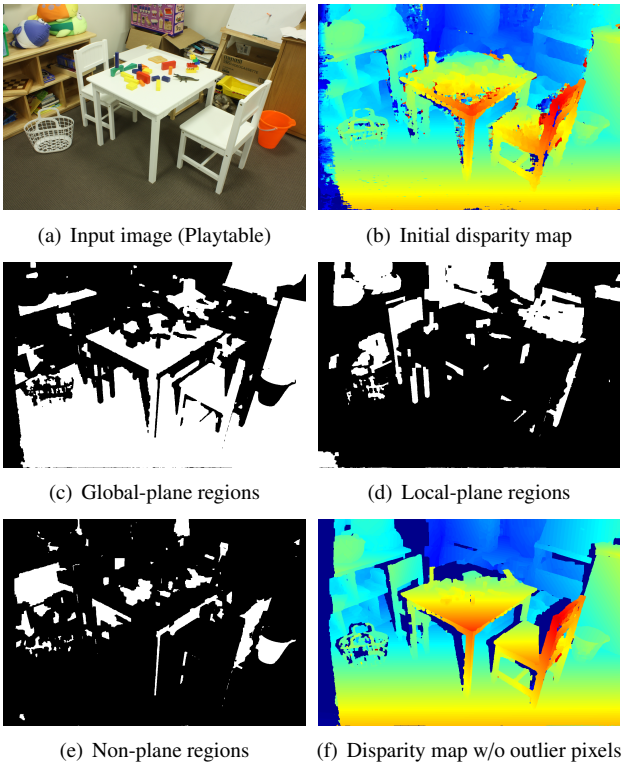
In this step, we estimate the disparity map $\mathbf{H}$ in which the disparity value of a pixel can be 1) over-parameterized, *e.g.*, $\mathbf{H}(\mathbf{p}) = \pi_i^\top \mathbf{p}$, 2) scalar, *e.g.*, $\mathbf{H}(\mathbf{p}) = \mathbf{D}_0(\mathbf{p})$, and 3) labeled as an outlier, *e.g.*, $\mathbf{H}(\mathbf{p}) = \phi$, as described in Fig. 3. Here, ϕ means that the pixel has an empty value that is refined through post-processing. To compute $\mathbf{H}$, we formulate the following energy function:

$$E(\mathbf{H}|\mathcal{P}, \mathbf{I}, \mathbf{D}_0, \mathcal{S}) = \sum_{\mathbf{p}} U(\mathbf{p}, \pi_i) + \lambda_{\mathrm{aps}} \sum_{(\mathbf{p},\mathbf{q})\in\mathcal{N}} V(\mathbf{p},\mathbf{q}), \tag{7}$$

where the unary term is defined as

$$U(\mathbf{p},\pi_i) = \begin{cases} s(\mathbf{p},\pi_i) & \text{if } \pi_i \in \mathcal{P}_{\mathrm{global}}, \\ s(\mathbf{p},\pi_i) + \epsilon_1 & \text{if } \pi_i \in \mathcal{P}_{\mathrm{local}} \quad \text{and} \quad \mathbf{p} \in \mathbf{s}_i, \\ s(\mathbf{p}, \mathbf{D}_0(\mathbf{p})) + \epsilon_2 & \text{if } \pi_i = \pi_{\text{non-plane}}, \\ \gamma_p & \text{if } \pi_i = \pi_{\mathrm{outlier}}, \end{cases} \tag{8}$$
$$\text{s.t. } \epsilon_1 < \epsilon_2.$$

Here, the dissimilarity function $s(\mathbf{p},\pi_i)$ compares two image patches from the reference image, *e.g.*, the left image, centered at $\mathbf{p}$, and from the target image at the shifted position of $\mathbf{p}$ by



(a) Input image (Playtable) (b) Initial disparity map

(c) Global-plane regions (d) Local-plane regions

(e) Non-plane regions (f) Disparity map w/o outlier pixels

Fig. 3. Disparity map estimation with local and global plane hypotheses. In (c)-(e), white indicates pixels that have been labeled as one of the following: global plane hypotheses (c), local plane hypotheses (d), or non-plane regions (e). The estimated disparity map (f) is the result of minimizing Eq. 7, in which the bad pixel rate in non-occluded regions is reduced from 15.69% (Einecke and Eggert (2015)) (b) to 8.60%. Moreover, unreliable pixels do not have disparity values because of the π_{outlier} label. These unreliable pixels are handled through the post-processing step.

the amount of $\pi_i^T \mathbf{p}$ or $D_0(\mathbf{p})$ along the scanline. We adopt MC-CNN (Žbontar and LeCun (2016)) to compare image patches; most recent algorithms adopt this approach because of its high performance. Moreover, we use the initial disparity value to consider pixels in non-planar regions. The pairwise term enforces the smoothness of labels between neighboring pixels using the Potts interaction model

$$V(\mathbf{p},\mathbf{q}) = w_{\mathbf{p},\mathbf{q}} \cdot T(\pi_i \neq \pi_j), \tag{9}$$

where $w_{\mathbf{p},\mathbf{q}}$ is a non-negative weight between two pixels

$$w_{\mathbf{p},\mathbf{q}} = \begin{cases} P_1 & \text{if } |\mathbf{I}(\mathbf{p}) - \mathbf{I}(\mathbf{q})| < \tau_{\mathrm{grad}}, \\ P_2 & \text{otherwise}, \end{cases} \tag{10}$$

that increases the strength of smoothing if the intensity or color difference between neighboring pixels is less than a predefined value τ_{grad}. $T(\pi_i \neq \pi_j)$ is a binary penalty function that returns one if two pixels have the different labels, and zero otherwise.

In contrast to previous studies (Besse et al. (2012); Bleyer et al. (2011)), our objective function does not assign plane parameters to all pixels. Instead, we exploit two additional labels, $\pi_{\text{non-plane}}$ and π_{outlier}, to consider non-plane regions as well as pixels that cannot be matched reliably. This energy function is the key contribution of the study; it relaxes the resultant disparity map as a mixture of over-parameterized disparity values, scalar disparity values, and outliers, with the aid of

global and local plane hypotheses. To distinguish different labels and planes, we examine the dissimilarity between pixels and employ two bias values, ϵ_1 and ϵ_2, to avoid the following situation. If the initial disparity map has smooth disparity values in a planar region, matching costs for different labels, $\mathcal{P}_{\text{local}}$, $\mathcal{P}_{\text{global}}$, and $\pi_{\text{non-plane}}$, are likely to be similar to each other because $\mathbf{D}_0(\mathbf{p}) \approx \pi_i^\top \mathbf{p}$. In this case, we assign the highest priority to the global plane, and the second highest priority to the local plane by simply assigning a small bias value.

On the other hand, Eq. (7) extends the idea of plane sweeping stereo Gallup et al. (2007); Sinha et al. (2014) in consideration of various reconstruction cues from the scene. The plane sweeping approach is highly efficient compared to conventional approaches, especially in cases in which the range of disparity values is large, *e.g.*, high-resolution images or wide baseline cameras, because the number of plane hypotheses is generally less than the range of disparity values. In the case shown in Fig. 3, the number of labels is 15 whereas the size of the original disparity range was 145. Among 15 labels, 12 labels were used to consider global plane hypotheses and the other three labels were used for the remaining hypotheses. Here, $\mathcal{P}_{\text{local}}$ and $\pi_{\text{non-plane}}$ each require only one label, because a pixel has neither more than one local hypothesis nor multiple initial disparity values. Note that if an important plane is missing, our approach does not degrade the quality of disparity maps significantly, owing to the non-plane label.

3.4. Post-processing

The disparity map obtained in the previous step significantly improves the quality of the initial disparity map. However, the scene structure may not be perfectly planar in practical situations, and plane parameters are prone to small errors caused by disparity errors. To further improve the disparity map $\mathbf{H}$, we perturb plane parameters assigned to each pixel. To this end, we define a new energy function for computing a refined disparity map $\hat{\mathbf{H}}$,

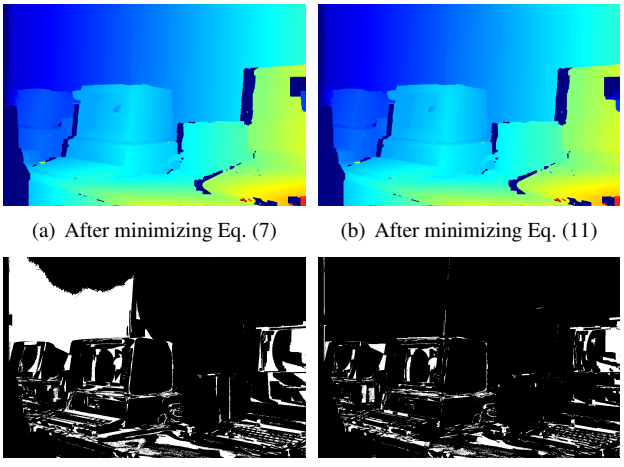
$$E(\hat{\mathbf{H}}|\mathbf{H}, \mathbf{I}) = \sum_{\pi_i \in \{\mathcal{P} \cup \pi_{\text{non-plane}}\}} E(\hat{\mathbf{H}}|\mathbf{H} = \pi_i, \mathbf{I}), \tag{11}$$

where $\hat{\mathbf{H}}$ is the refined disparity map. Here, $\mathbf{H} = \pi_i$ indicates the pixels that have the same label, *i.e.*, plane parameters. Therefore, we slightly change the disparity value by perturbing an assigned plane hypothesis or a disparity value, rather than assigning different plane hypotheses. Moreover, we do not refine pixels having the ϕ value at this moment, because these pixels usually do not have correspondences, *e.g.*, pixels in the leftmost columns or in the occluded regions as described in Fig. 3(f), such that correct disparity values cannot be recovered by examining image patches. Then, we define each subproblem as

$$E(\hat{\mathbf{H}}|\mathbf{H} = \pi_i, \mathbf{I}) = \sum_{\mathbf{p}} U(\mathbf{p}, \pi_i^{(k)}) + \lambda_{\text{per}} V_L(\mathbf{p}, \mathbf{q}). \tag{12}$$

Here, we define the unary term in a similar manner as Eq. (8):

$$U(\mathbf{p}, \pi_i^{(k)}) = \begin{cases} s(\mathbf{p}, \pi_i^{(k)}) & \text{if } \pi_i \in \mathcal{P}, \\ s(\mathbf{p}, \mathbf{D}^{(k)_0(\mathbf{p})}) & \text{if } \pi_i = \pi_{\text{non-plane}}, \end{cases} \tag{13}$$



(a) After minimizing Eq. (7) (b) After minimizing Eq. (11)

(c) Error map for (a) (d) Error map for (b)

Fig. 4. Perturbation-based disparity map refinement. Disparity map (b) is refined from (a). Although (a) and (b) appear identical, bad pixels are significantly reduced after the refinement procedure, as shown in (d).

where $\pi_i^{(k)}$ indicates a perturbed plane of π_i, *e.g.*, $\pi_i^{(k)} = \pi_i + [0\ 0\ z_k]^\top$ such that z_k is additive noise. We only perturb the disparity along the z-direction because of two reasons. The first reason is for the sake of simplicity if we perturb other elements, the size of the configuration space increases either quadratically or cubically. The second reason is that changing the last element was sufficient to refine the disparity map as shown in Fig. 3(b). If the non-plane label is assigned, we perturb the initial disparity value $\mathbf{D}_0(\mathbf{p})$ with z_k, *e.g.*, $\mathbf{D}_0^{(k)}(\mathbf{p}) = \mathbf{D}_0(\mathbf{p}) + z_k$. For the pairwise term, we slightly change Eq. (9), *e.g.*, $T(\pi_i^{(k)} \neq \pi_i^{(j)})$, in which the smoothness constraint is enforced between perturbed planes or disparity values.

On the other hand, two energy functions, Eqs. (7) and (11), can be combined into a single energy function by considering more plane hypotheses, *e.g.*, perturbed planes, in Eq. (7). However, this increases the size of the label space proportional to the number of perturbed planes. Moreover, Eq. (12) can be easily parallelized.

After the perturbation-based refinement step, we further improve the quality of the disparity map through conventional techniques. First, we estimate parabola-based subpixel displacements of pixels in non-plane regions, because they have discrete disparity values. Second, we propagate plane parameters or disparity values along the scanline from left to right Bleyer et al. (2011) to interpolate disparity values for pixels having the ϕ value. Third, we perform the median filter with a 5×5 size kernel. Afterward, we run the fast weighted median filter Zhang et al. (2014b) and employ filtered disparity values if the difference between the input disparity value and the filter response is larger than a predefined constant value τ_{disc}, in order to refine disparity values along the depth discontinuity; otherwise, the weighted median filter can significantly degrade the quality of disparity maps, owing to blocking artifacts or texture copying.

Table 1. Quantitative evaluation for the Middlebury 2014 benchmark at half resolution. Different algorithms are compared in terms of bad pixel rates in non-occluded regions, using a threshold value of 2.0px. Here, the average error indicates the weighted average bad pixel rates that were computed by assigning a small weight (0.5) to challenging datasets, including PianoL, Playroom, Shevles, Playtable, Vintage, Australia, ClassroomE, DjembL, Hoops, Livingroom, and Staircase, to decrease the influence of ill-conditioned datasets. The tables are written in the ascending order of average errors. Results are written in bold if the difference between the initial disparity map and the final disparity map is larger than 3%.

	Adiron	ArtL	Jade	Motor	MotorE	Piano	PianoL	Pipes	Playrm	Playt	PlaytP	Recyc	Shelevs	Teddy	Vintage	Avg.
LocalExp	1.20	3.53	8.95	3.38	3.64	9.11	14.7	3.97	9.07	6.45	5.85	6.50	**30.0**	2.64	5.24	6.52
LocalExp-apap	1.39	4.69	10.1	3.69	3.83	9.17	13.0	4.13	8.46	6.85	5.23	6.70	**17.2**	3.16	5.56	6.21
SGM-Net	2.77	4.86	11.9	3.30	3.57	8.71	13.4	3.45	8.66	6.50	6.10	6.55	**27.1**	2.89	**11.5**	7.01
SGM-Net-apap	1.88	5.57	9.09	3.71	4.15	7.93	13.4	4.10	7.65	8.21	5.88	6.63	**17.9**	3.27	**6.41**	6.32
3DMST	1.53	4.66	10.7	3.96	4.35	10.0	15.6	4.99	9.86	5.73	5.25	6.39	**29.9**	2.68	6.92	7.08
3DMST-apap	2.04	4.66	10.4	4.03	4.50	8.63	13.8	4.82	9.39	6.41	5.34	6.11	**14.8**	3.22	6.86	6.35
LW-CNN	2.81	4.86	13.0	3.10	3.29	11.7	**17.4**	3.66	**11.9**	**10.4**	9.63	6.97	**30.5**	2.68	**14.3**	8.31
LW-CNN-apap	2.20	4.71	10.2	3.83	3.94	10.0	**14.3**	4.07	**8.41**	**6.91**	7.17	6.98	**16.2**	2.98	**6.01**	6.56
MeshExt	3.53	6.76	**18.1**	5.30	5.88	8.80	13.8	8.10	11.1	8.87	8.33	10.5	**31.2**	4.96	**12.2**	9.51
MeshExt-apap	2.33	5.19	**14.3**	4.25	4.55	8.48	12.0	5.84	9.23	6.32	6.24	7.59	**20.4**	3.40	**6.36**	7.14
MC-CNN	3.33	8.04	16.1	3.66	3.76	12.5	**18.5**	4.22	**14.6**	**15.1**	**13.3**	6.92	**30.5**	4.65	**24.8**	10.3
MC-CNN-apap	2.39	5.77	17.0	3.98	4.23	11.4	**14.7**	4.52	**10.1**	**10.8**	**9.42**	7.38	**18.6**	3.29	**8.19**	8.05
MBM	**8.2**	8.9	**17.5**	5.45	5.49	**16.5**	**25.2**	6.09	**18.5**	**15.7**	**15.5**	10.6	**38.5**	5.0	**29.5**	**13.0**
MBM-apap	**3.04**	7.22	**13.5**	4.39	4.68	**10.7**	**16.1**	5.35	**10.1**	**8.60**	**8.11**	7.70	**12.2**	5.16	**7.97**	**7.78**
SGM	**15.3**	7.69	**18.1**	**10.9**	**8.90**	**16.4**	**29.1**	**11.5**	**21.7**	**52.5**	**15.8**	**14.6**	**46.4**	6.52	**39.3**	**17.6**
SGM-apap	**4.23**	6.10	**12.2**	**4.63**	**4.78**	**12.4**	**18.2**	**6.05**	**13.8**	**21.3**	**10.0**	**8.01**	**25.7**	3.66	**8.36**	**9.26**

	Austr	AustrP	Bicyc2	Class	ClassE	Compu	Crusa	CrusaP	Djemb	DjembL	Hoops	Livgrm	Nkuba	Plants	Stairs	Avg.
LocalExp	3.65	2.87	2.98	1.99	5.99	3.37	3.48	3.35	2.05	10.3	9.75	8.57	14.4	5.40	9.55	5.43
3DMST	3.71	2.78	4.75	2.72	7.36	4.28	3.44	3.76	2.35	12.6	11.5	8.56	14.0	5.35	8.87	5.92
LW-CNN	4.65	3.95	5.30	2.63	11.2	5.41	4.32	4.22	2.43	12.2	13.4	13.6	14.8	4.72	12.0	7.04
MeshExt	4.41	3.98	5.40	3.17	10.0	8.89	4.62	4.77	3.49	12.7	12.4	10.4	14.5	7.80	8.85	7.29
MBM-apap	5.43	4.91	5.11	5.17	21.6	6.99	4.31	4.23	3.24	14.3	9.78	7.32	13.4	6.30	8.46	7.26
SGM-Net	4.71	3.69	4.93	3.18	11.1	5.37	5.57	5.81	2.65	14.5	13.2	13.1	14.8	5.63	11.2	7.37
MC-CNN	5.59	4.55	5.96	2.83	11.4	8.44	8.32	8.89	2.71	16.3	14.1	13.2	13.0	6.40	11.1	8.29
HybridCNN-CRF	4.09	3.97	8.44	6.93	11.1	13.8	19.5	19.0	3.66	17.0	18.2	18.0	21.0	7.29	17.8	12.5

4. Experimental Results

To evaluate the proposed method, we used two popular datasets: the Middlebury 2014 benchmark (Scharstein et al. (2014)) and the KITTI 2015 benchmark (Geiger et al. (2013)). In particular, the Middlebury 2014 benchmark provides various challenging scenarios containing large homogeneous regions, illumination changes, occlusions, and rectification errors. We analyzed characteristics of the proposed method from various aspects, including failure cases and the sensitivity of our method against changes to important parameters.

Parameter setting: To compute initial disparity maps, we employed various state-of-the-art methods and popularly used algorithms as the first step; these include MC-CNN (Žbontar and LeCun (2016)), MeshStereo (Zhang et al. (2015)), MBM (Einecke and Eggert (2015)), SGM (Hirschmüller (2008)), LocalExp (Taniai et al. (2017)), and SGM-Net (Seki and Pollefeys (2017)). Therefore, we inserted "-apap" at the end of the algorithms' names to indicate that the proposed method utilized a specific algorithm to acquire the initial disparity maps. To generate local plane hypotheses, we set the size of a superpixel Achanta et al. (2012) to 50, the regularization parameter to 20, α to 0.5, the cutoff threshold to 50, β to 1000, and γ_d to 10. For RANSAC, we set the inlier threshold to 1.5px and the number of iterations to 500. To estimate $\mathbf{H}$, we set λ_{aps} to 28, ϵ_1 to 0.05, ϵ_2 to 0.10, γ_p to 0.55, τ_{grad} to 9, P_1 to 1, P_2 to 3, and λ_{per} to 10. For perturbation-based refinement, we added a noise vector $[0\,0\,z_k]^\top$ to the plane hypotheses such that z_k is an integer value in $-2 \leq z_k \leq 2$, and similarly, added z_k to scalar disparity values. For the weighted median filter Zhang et al. (2014b), we set the radius of a window to 5, the regularization parameter to 1.5, and τ_{disc} to 4. For all energy minimization procedures, we used graph cuts Delong et al. (2012).

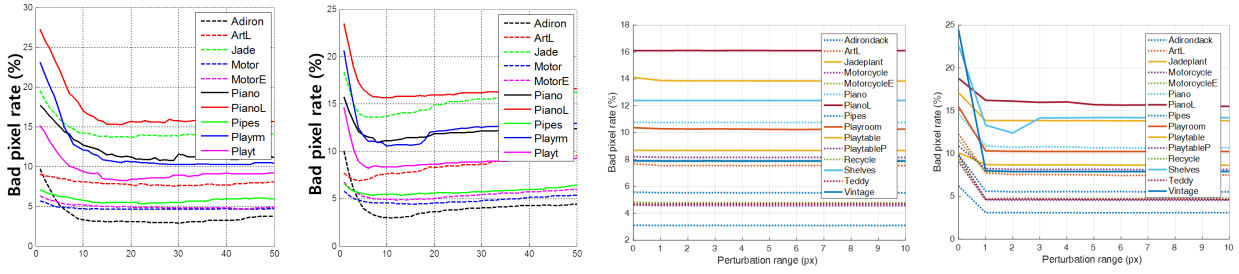
4.1. Middlebury 2014 benchmark

First of all, the proposed method was shown to be effective when the captured image contains largely homogeneous regions such as walls and ceilings. For example, Shelves and Vintage datasets contain large homogeneous regions; in this case, the global plane hypotheses played an important role in recovering underlying structures because the homogeneous regions were planar regions as shown in Fig. 7. Interestingly, even when the state-of-the-art methods were used as input, the proposed method reduced the number of bad pixels for Shelves and Vintage as described in Tab. 1. This improvement verifies that the proposed approach is desirable for man-made environments, especially when the captured image contains large textureless regions. In addition, the proposed method showed accurate results for highly ambiguous regions resulting from different lighting conditions and reflective surfaces, *e.g.*, PianoL and Playroom, for which existing algorithms frequently fail to estimate accurate structures. For the testing dataset, we could not evaluate various methods, because the Middlebury benchmark does not allow multiple submissions. Therefore, we only uploaded MBM-apap, to show that a simple and efficient approach Einecke and Eggert (2015) can achieve state-of-the-art performance if it is coupled with the proposed method. Here, MBM refers to multi-block matching that aggregates match-

(a) PianoL (b) Playroom (c) Playtable (d) Shelves

(e) Vintage (f) Hoops (g) Livingroom (h) Staircase

Fig. 5. Challenging images from the Middlebury 2014 benchmark. (a)-(e) are from the training dataset and (f)-(h) are from the test dataset. Frequently mismatched regions are marked with red rectangles. Here, the left and right sides of PianoL have different lighting conditions, and Playroom has a reflective region at the top-right of the image. The remaining datasets contain largely homogeneous regions.



(a) Sensitivity to λ_{aps} (b) Sensitivity to λ_{per} (c) Sensitivity to z_k (scalar) (d) Sensitivity to z_k (plane)

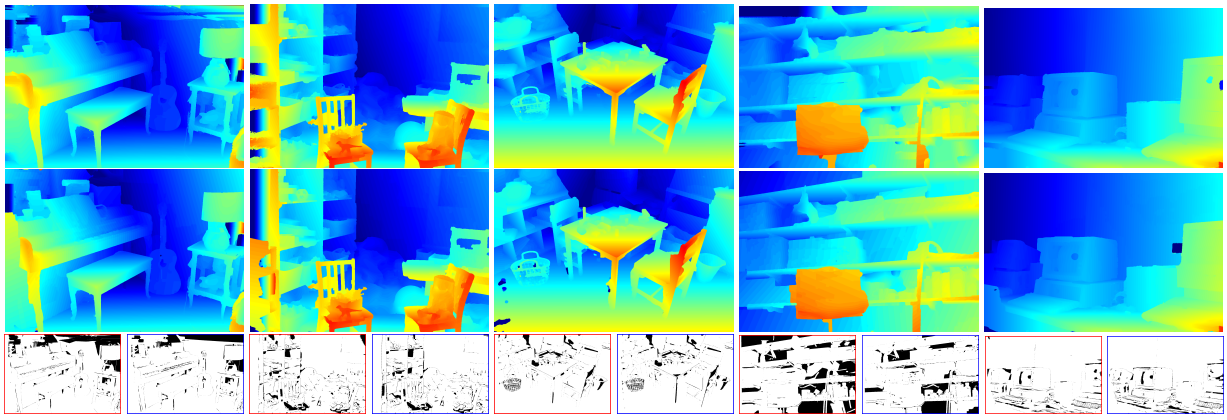
Fig. 6. Sensitivity analysis. Each figure shows bad pixel rates resulting from changing for parameters.

ing costs with multiple box filters. In addition, it was interesting to see that there is no method that employ an end-to-end network for the Middlebury benchmark among top-performing methods; most of the state-of-the-art methods utilize MC-CNN (Žbontar and LeCun (2016)) to compute initial matching costs. To find a reason for this, we evaluated the performance of HybridCNN-CRF (Knöbelreiter et al. (2017)) which computes disparity maps through an end-to-end network. As shown in Table 1, their results show poorer results compared to other methos that utilize MC-CNN to compute matching costs. This is because the Middlebury dataset contains high-resolution images captured at diverse view points, training an end-to-end network for such a dataset is a difficult task. In other words, it seems more practical to train a similarity function to deal with diverse images. Žbontar and LeCun (2016) showed that the trained similarity function using the Middlbury dataset also shows similar performance for the KITTI dataset, though two datasets have different characteristics and contents.

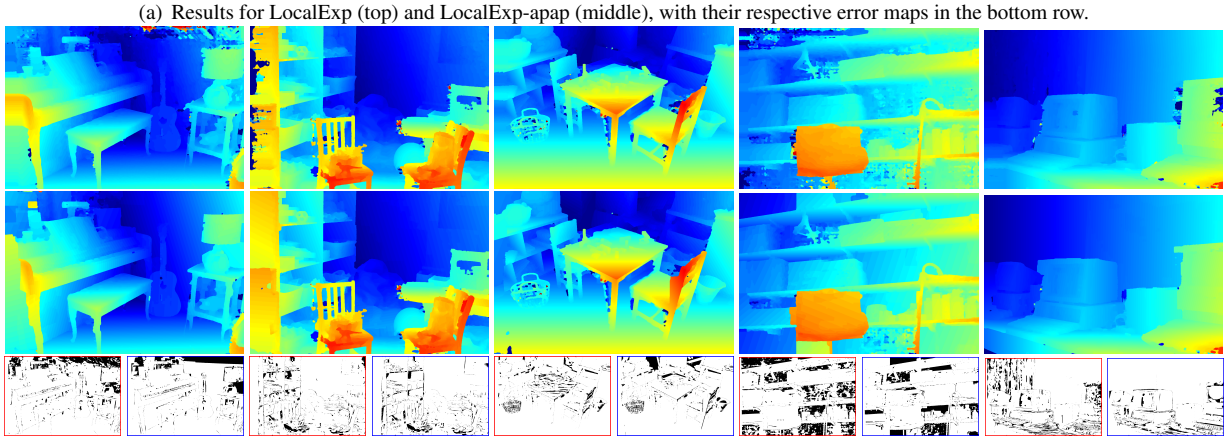
For qualitative evaluation, we compared disparity maps for the selected images from the training and test datasets as shown in Fig. 7. These results verify the necessity of global plane hypotheses in computing disparity maps, where existing methods frequently failed to estimate correct disparity values in ambiguous regions; these regions are described in Fig. 5.

Sensitivity analysis: The quantitative evaluation in Tab. 1 verifies that the proposed method is not limited to a specific stereo matching algorithm. Even the state-of-the-art method can be further improved with the proposed method. In addition, we analyzed the sensitivity of two important parameters, λ_{aps} and λ_{per}, which were used in Eqs. (7) and (12). We changed the value of λ_{aps} and λ_{per} from 1 to 50 in intervals of 1, and computed the bad pixel rate of the disparity maps generated for various input images. As shown in Fig. 6, the proposed method shows consistent bad pixel rates as long as the values of λ_{aps} and λ_{per} are higher than a certain value, *e.g.*, 10. We also varied z_k for perturbing planes and disparity values where zero on the x-axis indicates that we did not perturb them. In case of scalar disparity values, the perturbation slightly affected the quality of resultant disparity maps and the optimal value was about 6 pixels. On the contrary, the perturbation of plane hypotheses significantly decreases error rates as long as the range was greater than or equal to 1 pixel. Here, among 6.6% of average improvement, Eq. (7) improved 1.5%, Eq. (11) improved 4.9%, and the additional refinment improved 0.2% of errors, when the MBM method was used to compute initial disparity maps.

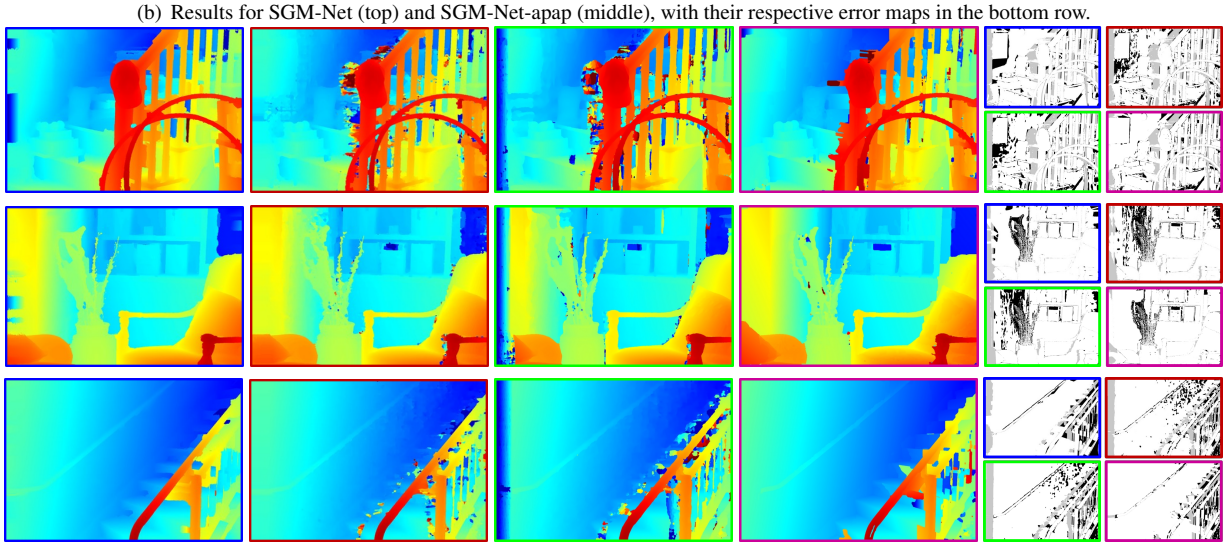
In Table 2, we showed the change of bad pixel rates after removing each term in Eq. (8). Here, the global plane hypotheses affected the quality of disparity maps most significantly, be-



(a) Results for LocalExp (top) and LocalExp-apap (middle), with their respective error maps in the bottom row.



(b) Results for SGM-Net (top) and SGM-Net-apap (middle), with their respective error maps in the bottom row.



(c) LocalExp (d) SGM-Nets (e) MC-CNN (f) MBM-apap (g) Error maps

Fig. 7. Qualitative evaluation for the Middlebury 2014 dataset. We compared disparity maps for the most challenging images from the Middelbury benchmark, which are shown in Fig. 5. These images primarily consist of homogeneous and planar regions, which are also typical characteristics of man-made environments. In particular, among all the benchmark datasets, recent algorithms generate the largest number of errors for the Shelves dataset. Erroneous pixels are shown in black; in addition, error maps for (c)-(f) are described in (g) with the same boundary colors.

Table 2. Ablation study. Numbers indicate increase in average bad pixel rates when each term or procedure is removed.

	$\mathcal{P}_{\text{global}}$	$\mathcal{P}_{\text{local}}$	$\pi_{\text{non-plane}}$	π_{outlier}	Add. ref.
MBM	4.72	0.44	1.05	1.18	0.22
MC-CNN	4.02	0.02	0.28	1.28	0.11
SGM-Net	1.47	0.002	2.01	0.89	0.17
LocalExp	0.58	0.01	1.57	0.48	0.14

(a) Input stereo pair (b) Our result (c) Error map for (b)

Fig. 8. Results for the Classroom2E dataset. The proposed method shows a large number of errors (especially in chair regions), because these pixels were classified as outliers and then post-processed through hole-filling.

cause failure cases of existing algorithms usually occurr in large homogeneous regions. The improvement tends to decrease as the quality of input increases as the quality of input increases, in this case, the non-plane term becomes more important. Local plane hypotheses were meaningful when the input image contains a large amount of errors, otherwise, local plane hypotheses were assigned to a small number of pixels. Moreover, we showed the increase in bad pixel rates after removing additional post-processng steps that were applied after the plane perturbation step.

Failure case analysis: The proposed method showed the highest error rate when processing the Classroom2E dataset, in which input images were taken under different exposures as shown in Fig. 10(a). Therefore, the matching costs for this image were higher than for the other images. Consequently, many pixels in chair regions were labeled as outliers, including correct initial disparity values. Because we refine outlier pixels through the hole filling technique, the disparity values of these pixels were interpolated from neighboring pixels. As shown in the right-bottom region, many pixels in the chair region were erased and filled with incorrect disparity values. Increasing the value of γ_p can prevent this problem, but we used the same parameters to ensure a fair comparison. However, similar phenomenon did not occur with other images.

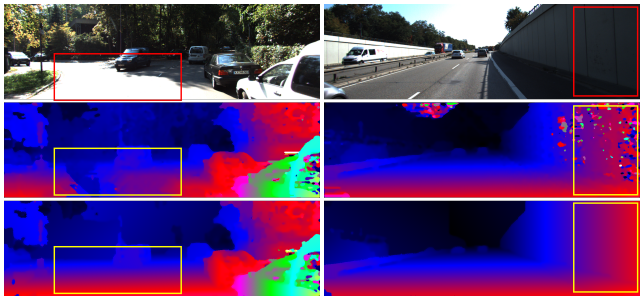
Time complexity analysis: On average, the proposed method took 120.1 seconds to compute half-resolution disparity maps, excluding the computation of per-pixel matching costs. We provide a detailed description of running time for the Playtable dataset in Table 3 in Table 3. Here, essential procedures took approximately 50 seconds, including three energy minimization procedures that were written in C++. The overhead includes memory allocation, RANSAC, likelihood computation, etc., which were written in MATLAB. We ran this test on a PC with 3.5GHz CPU and 16.0GB RAM.

Table 3. Time complexity analysis for Playtable image (926×1360 with 145 disparity levels). We used MBM to compute the initial disparity map.

Initial	clustering	Eq. (1)	Eq. (7)	Eq. (11)	Refine	Overhead	All
5.20	1.06	4.27	26.59	14.03	2.54	61.78	107.7

Table 4. Quantitative evaluation for the KITTI 2015 test dataset. Algorithms are sorted in ascending order of D1-bg errors. SGM Hirschmüller (2008) is the result of our implementation.

	D1-bg	**D1-fg**	**D1-All**
PSM-Net	1.71 %	4.31 %	2.14 %
PSM-Net-apap	1.83 %	4.71 %	2.30 %
MC-CNN-acrt	2.89 %	8.88 %	3.89 %
Displets v2	3.00 %	5.56 %	3.43 %
3DMST	3.36 %	13.03 %	4.97 %
SGM-apap	3.66 %	12.20 %	5.08 %
Content-CNN	3.73 %	8.58 %	4.54 %
SPS-St	3.84 %	12.67 %	5.31 %
SGM	4.11 %	15.29 %	5.97 %
DispNetC	4.32 %	4.41 %	4.34 %



(a) 23rd test frame (b) 50th test frame

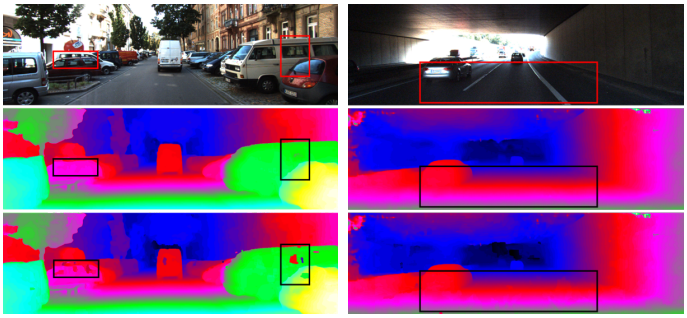
Fig. 9. A qualitative comparison. Left input images, intial disparity maps (SGM), and our results are shown from the top. Frequently mismatched regions are marked with rectangles.

4.2. KITTI 2015 benchmark

We evaluated the proposed method using the KITTI 2015 dataset which is captured in various driving environments. We selected a popularly used algorithm, semi-global matching (SGM) of Hirschmüller (2008), and a state-of-the-art method, a pyramid stereo matching network (PSM-Net) of Chang and Chen (2018) to compute initial disparity maps. Moreover, we used MC-CNN to compute similarity between patches.

Figure 9 compares disparity maps obtained in challenging scenarios *e.g.*, when the pixels are saturated owing to reflected sunlight or darkened because of low dynamic range. In these cases, the proposed method effectively recovered mismatched pixels, with the aid of global plane hypotheses. Otherwise, the improvement was negligible because most pixels were matched correctly through the SGM method. Quantitatively, the average improvement was approximately 0.5% and 0.9% for background and all regions, respectively, as shown in Table 4. Furthermore, when the highly accurate disparity maps are given as input, *i.e.* more than 98% pixels are correctly matched, most pixels do not change thier disparity values after applying the proposed method. Although the average bad pixel rate increased for the PSM-Net, nearly half of the images, 98 out of 200, were improved slightly, when the same experiment was carried out using the training dataset.

Failure case analysis: The proposed method does not always improve the quality of disparity maps, especially when a glossy or transparent surface is presented in the scene. For example,



(a) 7th test frame (b) 51th test frame

Fig. 10. Failure cases for the KITTI 2015 benchmark. Left input images, intial disparity maps (PSM-Net), and our results are shown from the top. Degraged regions are regions are marked with rectangles.

in Fig. 9(b), mismatched pixels in the car region remain unchanged after applying the proposed method. Moreover, if an input disparity map has accurate disparity values in these regions, the proposed method can turn correct disparity values into errors as shown in Fig. 10(a). One reason for this is because of the lack of training data for glossy and transparent pixels, which is also difficult to acquire ground truth disparity labels. We observed one more case of quality degradation from images capturing a high-contrast scene, *e.g.* at the end of a tunnel, as shown in Fig. 10(b). The proposed method tends to fluctuate disparity values in homogeneous regions, as the signal-to-noise ratio decreases.

5. Conclusion

We have presented a stereo matching algorithm designed for man-made environments such as indoor scenes and driving environments. After computing an initial disparity map, we found local and global plane hypotheses to exploit them, in order to recover accurate structures in highly ambiguous regions, *e.g.*, walls and roads. The key idea was to avoid the oversimplification problem by employing two additional labels (non-plane and outlier) in the energy minimization framework. We demonstrated that the proposed method effectively handles largely ambiguous regions where existing stereo algorithms fail to estimate correct depth maps.

Acknowledgments

This work was supported by 'The Cross-Ministry Giga KOREA Project' grant funded by the Korea government (MSIT) (No.GK17P0300, Real-time 4D reconstruction of dynamic objects for ultra-realistic service), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2015R1A2A1A01005455), and Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-TC1603-05.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Sšstrunk, S., 2012. Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans. on Pattern Analysis and Machine Intelligence 34.

Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B., 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. ACM Trans. on Graphics 28.

Besse, F., Rother, C., Fitzgibbon, A., Kautz, J., 2012. Pmbp: Patchmatch belief propagation for correspondence field estimation, in: British Machine Vision Conference (BMVC).

Birchfield, S., Tomasi, C., 1999. Multiway cut for stereo and motion with slanted surfaces, in: IEEE International Conference on Computer Vision (ICCV), pp. 489–495. doi:10.1109/ICCV.1999.791261.

Bleyer, M., Rhemann, C., Rother, C., 2011. Patchmatch stereo - stereo matching with slanted support windows, in: British Machine Vision Conference (BMVC).

Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 1222–1239. doi:10.1109/34.969114.

Chang, J.R., Chen, Y.S., 2018. Pyramid stereo matching network, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5410–5418.

Chen, Z., Sun, X., Yu, Y., Wang, L., Huang, C., 2015. A deep visual correspondence embedding model for stereo matching costs. IEEE International Conference on Computer Vision (ICCV) .

Coughlan, J., Yuille, A.L., 2000. The manhattan world assumption: Regularities in scene statistics which enable bayesian inference, in: Conf. on Neural Information Processing Systems (NIPS), p. 845–851.

Delong, A., Osokin, A., Isack, H.N., Boykov, Y., 2012. Fast approximate energy minimization with label costs. Int. J. Comput. Vision 96, 1–27.

Einecke, N., Eggert, J., 2014. Block-matching stereo with relaxed fronto-parallel assumption, in: IEEE Intelligent Vehicles Symposium (IV), pp. 700–705.

Einecke, N., Eggert, J., 2015. A multi-block-matching approach for stereo, in: IEEE Intelligent Vehicles Symposium (IV), pp. 585–592. doi:10.1109/IVS.2015.7225748.

Furukawa, Y., Curless, B., Seitz, S., Szeliski, R., 2009. Manhattan-world stereo, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1422–1429.

Gallup, D., Frahm, J.M., Mordohai, P., Yang, Q., Pollefeys, M., 2007. Real-time plane-sweeping stereo with multiple sweeping directions, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1–8.

Gallup, D., Frahm, J.M., Pollefeys, M., 2010. Piecewise planar and non-planar stereo for urban scene reconstruction, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1418–1425.

Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR) .

Güney, F., Geiger, A., 2015. Displets: Resolving stereo ambiguities using object knowledge, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

Hadfield, S., Bowden, R., 2015. Exploiting high level scene cues in stereo reconstruction, in: IEEE International Conference on Computer Vision (ICCV), pp. 783 – 791.

Häne, C., Heng, L., Lee, G.H., Sizov, A., Pollefeys, M., 2014. Real-time direct dense matching on fisheye images using plane-sweeping stereo., in: 3DV, IEEE Computer Society. pp. 57–64.

Hirschmüller, H., 2008. Stereo processing by semiglobal matching and mutual information. IEEE Trans. on Pattern Analysis and Machine Intelligence 30, 328–341.

Hirschmüller, H., Buder, M., Ernst, I., 2012. Memory efficient semi-global matching, in: The XXII Congress of the International Society for Photogrammetry and Remote Sensing. URL: https://elib.dlr.de/78804/.

Hong, L., Chen, G., 2004. Segment-based stereo matching using graph cuts, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Isack, H., Boykov, Y., 2012. Energy-based geometric multi-model fitting. Int. J. Comput. Vision 97, 123–147.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A., 2017. End-to-end learning of geometry and context for deep stereo regression, in: Proceedings of the International Conference on Computer Vision (ICCV).

Klaus, A., Sormann, M., Karner, K., 2006. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure, in: Proceedings of the 18th International Conference on Pattern Recognition - Volume 03, pp. 15–18.

Knöbelreiter, P., Reinbacher, C., Shekhovtsov, A., Pock, T., 2017. End-to-end

training of hybrid CNN-CRF models for stereo, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society. pp. 1456–1465.

Kolmogorov, V., Rother, C., 2007. Minimizing non-submodular functions with graph cuts - a review. IEEE Trans. on Pattern Analysis and Machine Intelligence 29.

Lee, Y., Park, M., Hwang, Y., Shin, Y., Kyung, C., 2018. Memory-efficient parametric semiglobal matching. IEEE Signal Processing Letters 25, 194–198. doi:10.1109/LSP.2017.2778306.

Li, L., Yu, X., Zhang, S., Zhao, X., Zhang, L., 2017a. 3d cost aggregation with multiple minimum spanning trees for stereo matching. Applied Optics 56, 3411–3420.

Li, L., Zhang, S., Yu, X., Zhang, L., 2017b. Pmsc: Patchmatch-based superpixel cut for accurate stereo matching. IEEE Transactions on Circuits and Systems for Video Technology PP, 1–1.

Li, Y., Min, D., Brown, M.S., Do, M.N., Lu, J., 2015. Spm-bp: Sped-up patchmatch belief propagation for continuous mrfs, in: IEEE International Conference on Computer Vision (ICCV).

Liang, Z., Feng, Y., Guo, Y., Liu, H., 2018. Learning for disparity estimation through feature constancy, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

Luo, W., Schwing, A., Urtasun, R., 2016. Efficient deep learning for stereo matching, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T., 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

Mllner, D., 2013. fastcluster: Fast hierarchical, agglomerative clustering routines for r and python.

Muninder, V., Soumik, U., Krishna, G., 2014. Robust segment-based stereo using cost aggregation, in: British Machine Vision Conference (BMVC).

Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation, in: European Conference on Computer Vision (ECCV), Springer. pp. 483–499.

Pang, J., Sun, W., Ren, J.S., Yang, C., Yan, Q., 2017. Cascade residual learning: A two-stage convolutional neural network for stereo matching, in: ICCV Workshop on Geometry Meets Deep Learning.

Park, H., Lee, K.M., 2016. Look wider to match image patches with convolutional neural networks. IEEE Signal Processing Letters PP, 1–1. doi:10.1109/LSP.2016.2637355.

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

Schindler, G., Dellaert, F., 2004. Atlanta world: an expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

Seki, A., Pollefeys, M., 2017. Sgm-nets: Semi-global matching with neural networks, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

Sinha, S.N., Scharstein, D., Szeliski, R., 2014. Efficient high-resolution stereo matching using local plane sweeps, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

Straub, J., Rosman, G., Freifeld, O., Leonard, J.J., Fisher III, J.W., 2014. A Mixture of Manhattan Frames: Beyond the Manhattan World, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

Sun, J., yeung Shum, H., ning Zheng, N., 2003. Stereo matching using belief propagation, pp. 787–800.

Taniai, T., Matsushita, Y., Sato, Y., Naemura, T., 2017. Continuous 3d label stereo matching using local expansion moves. IEEE Transactions on Pattern Analysis and Machine Intelligence , 1–1doi:10.1109/tpami.2017.2766072.

Wang, Z.F., Zheng, Z.G., 2008. A region based stereo matching algorithm using cooperative optimization, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8. doi:10.1109/CVPR.2008.4587456.

Woodford, O., Torr, P., Reid, I., Fitzgibbon, A., 2009. Global stereo reconstruction under second-order smoothness priors. IEEE Trans. on Pattern Analysis and Machine Intelligence 31, 2115–2128.

Yamaguchi, K., McAllester, D., Urtasun, R., 2014. Efficient joint segmentation, occlusion labeling, stereo and flow estimation, in: European Conference on Computer Vision (ECCV).

Žbontar, J., LeCun, Y., 2015. Computing the stereo matching cost with a convolutional neural network, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

Žbontar, J., LeCun, Y., 2016. Stereo matching by training a convolutional neural network to compare image patches. Journal of Machine Learning Research 17, 1–32.

Zhang, C., Li, Z., Cai, R., Chao, H., Rui, Y., 2014a. As-rigid-as-possible stereo under second order smoothness priors, in: European Conference on Computer Vision (ECCV), pp. 112–126.

Zhang, C., Li, Z., Cheng, Y., Cai, R., Chao, H., Rui, Y., 2015. Meshstereo: A global stereo model with mesh alignment regularization for view interpolation, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, pp. 2057–2065. URL: https://doi.org/10.1109/ICCV.2015.238, doi:10.1109/ICCV.2015.238.

Zhang, Q., Xu, L., Jia, J., 2014b. 100+ times faster weghted median filter, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).