Coordinating Multiple Disparity Proposals for Stereo Computation

Ang Li, Dapeng Chen, Yuanliu Liu, Zejian Yuan
Xi'an Jiaotong University, China
{bennie.522,chendapeng1988}@stu.xjtu.edu.cn, {liuyuanliu88,yzejian}@gmail.com

Abstract

While great progress has been made in stereo computation over the last decades, large textureless regions remain challenging. Segment-based methods can tackle this problem properly, but their performances are sensitive to the segmentation results. In this paper, we alleviate the sensitivity by generating multiple proposals on absolute and relative disparities from multi-segmentations. These proposals supply rich descriptions of surface structures. Especially, the relative disparity between distant pixels can encode the large structure, which is critical to handle the large textureless regions. The proposals are coordinated by point-wise competition and pairwise collaboration within a MRF model. During inference, a dynamic programming is performed in different directions with various step sizes, so the long-range connections are better preserved. In the experiments, we carefully analyzed the effectiveness of the major components. Results on the 2014 Middlebury and KITTI 2015 stereo benchmark show that our method is comparable to state-of-the-art.

1. Introduction

Stereo computation is a fundamental task in computer vision with many applications like 3D reconstruction [7, 25], autonomous driving [10], view synthesis [20], etc. Meanwhile it always suffers from matching ambiguities due to noise, low or repetitive textures, and occlusions. To alleviate the matching ambiguities, many efforts focus on making a prior on the disparity or surface. Cost filtering methods [22, 36, 15] assume that a local support region should possess the same disparity. While most energy based methods [33, 32, 11] impose smoothness prior that adjacent pixels tend to have similar disparities or lie on the same surface. Notably, segmentation of color image [38, 29, 16] has been adopted to build the disparity prior. Segmentation provides clues about depth discontinuities and potentially groups the points belonging to an identical regular surface. Building a regular 3D surface model over a superpixel can help preserve the depth discontinuity, and in particular can recover the disparities of large textureless regions.

A major challenge of segment-based methods is how to segment the images properly, such that both the regularity of the surface fragments and the depth discontinuities are well preserved. In the case of under-segmentation, a superpixel cannot be explained by a single surface model. In the case of over-segmentation, superpixels may get too small to estimate a reliable surface model. It remains an open problem to select a suitable segmentation scale in general situations. We consider to make use of multi-segmentation, so that we can obtain a rich set of overlapping surface fragments. From these fragments, we can compromise a suitable description of the 3D scene, which avoids the risk of early-decision in picking the segmentation method and parameters.

In this paper, we propose a strategy to generate multiple disparity proposals from different segmentations as priors to regularize the disparities. Disparity proposals include absolute disparities as well as relative disparities between pixel pairs. The absolute disparities are the initial disparities refined by surface models, thus are more reliable. The relative disparities focus on the surface structure. Unlike the traditional local smoothness priors, these pairwise relations are structure-dependent, and the long range relations can encode the large structure directly.

To find out suitable descriptions from multiple proposals, we develop a coordinating scheme consisting of point-wise competition and pairwise collaboration. Point-wise competition picks the best absolute disparities exclusively, which can effectively suppress the outliers. Pairwise collaboration casts a vote on relative disparities, which can distinguish the artificial boundaries. The coordinating scheme is realized by a MRF model with non-local connections. During inference, we adapt the SGM algorithm by executing the 1D dynamic programming in various step sizes, so the long range connections can be better preserved without losing much computation efficiency.

The contributions of this paper are mainly three-fold:

- We use 3D surfaces to generate multiple disparity proposals, where the absolute disparity proposals are reliable point-wise estimations and relative disparity proposals reflect the 3D structures.
- We develop a coordinating scheme to mediate multiple

disparity proposals. The scheme includes point-wise competition and pairwise collaboration. Both of them are embedded in a unified MRF model.

- We propose to find an approximate optima of the MRF model with non-local connections by performing 1D dynamic programming in different directions with various spatial step sizes.

2. Related Work

Existing stereo methods are generally classified into local methods and global methods. Local methods [15, 36, 28, 31] aggregate matching cost on a support region. Global methods [33, 5, 32, 26] usually optimize a probability model that typically includes an observation term and a prior term, which our method belongs to. For a comprehensive taxonomy about stereo methods, we refer readers to the survey [23]. Here we only discuss the most related work.

Disparity priors of neighboring points play an important role in global methods. First-order smoothness priors [18, 5] force the adjacent two pixels have the same disparity value. Classical methods such as SGM [13] and graph cut based method [5] have adopted these priors, which prefer to recover the disparities of fronto-parallel planes. Second order smoothness priors penalize large second derivatives of disparity and can better model general plane structures [2, 30]. Woodford et al. [33] decompose the triple cliques which represent the second-order terms into equivalent pairwise representations. Olsson et al. [19] use pairwise interactions with tangent planes to represent second-order smoothness. Recently, some methods attempt to learn the priors from data, e.g., Wei et al. [32] retrieve semantic-similar patches in training set to regularize the estimate. Güney and Geiger [11] take object-category specific disparity proposals as priors for a certain object class. Our work exploits multiple regular 3D surfaces to obtain structure priors, which can not only act on adjacent pixels, but also propose relative disparity between the points at a longer distance.

Stereo estimation can also benefit from color image segmentation. Segmentation provides clues about depth discontinuities and potentially groups the points belonging to a same regular surface. Based on the segmentation methods [29, 3, 14, 16, 27, 4, 38, 34, 35] build a 3D surface model to estimate the disparity. In these methods, segmentation cues can be employed either as hard or soft constraints. Hard constraints indicate pixels within a superpixel strictly lie on a same 3D surface. Tao et al. [29] model each color segment as a single disparity plane, which fails if segments straddle depth boundaries. Soft constraints denote the solutions should be consistent with a given color segment but also allow for deviation. Sun et al. [27] bias the disparity map towards the fitted disparity over superpixels by adjusting the data term. Bleyer et al. [3] encourage the segmentation assumption to be fulfilled in subsegments. Our method extends the idea of the soft constraints. We make use of the segmentation to derive multiple disparity proposals, and prefer our final output to be consistent with both the absolute proposals and the relative proposals.

Multi-segmentation supplies a rich set of representation of the surface structures. In order to seek most reliable structures, Bleyer et al. [3] enforce the consistency between the estimations of the segments and the subsegments. Chakrabarti et al. propose a consensus framework [6] that can simultaneously infer inlier regions and enforce smoothness over neighboring inlier regions. We develop a coordinating scheme to mediate disparity proposals based on multi-segmentation. The scheme includes point-wise competition and pairwise collaboration. Point-wise competition picks the best absolute disparities exclusively, while pairwise collaboration casts a vote on relative disparities. We embed both mechanisms into a MRF model to exert their complementary abilities.

3. Our Approach

Given a rectified left image I_L and a right image I_R, we aim at estimating the disparity map D of the reference image (e.g., the left image I_L). The flow chart of our method is shown in Fig. 1. First we calculate an initial disparity by some off-the-shelf method. Then we produce multiple segmentations of the reference image, and fit the initial disparity within each fragment by a 3D disparity surface. These 3D surfaces are used to generate proposals for the absolute disparities of individual pixels and the relative disparity between pairs of pixels. The proposals are used as priors and coordinated by point-wise competition and pairwise collaboration within a MRF model. The output disparity map is the one that fits the global model best.

3.1. Surface Fragments from Multi-Segmentation

We generate 3D surface fragments based on multiple segmentations. Reference image I_L are segmented with different methods and parameters, producing M segmentation maps. Putting the 2D segments together with an initial disparity map D^0, we will obtain a set of 3D surface fragments $\mathcal{S}$. Each surface fragment $s \in \mathcal{S}$ is further fitted into a regular shape. For simplicity we use the 3D plane, which can be parameterized by $\Theta_s = [a_s, b_s, c_s]^T$. More complex surfaces, such as quadratic surfaces, are also applicable here.

To obtain 3D surface fragments that are adequate for representing the scene structure, we first adopt left-right consistency check (LRC) to select reliable initial disparities, then apply RANSAC[9] with least squares to derive 3D planes. Unreliable initial disparities caused by occlusion or ambiguous matches can be filtered out by LRC.

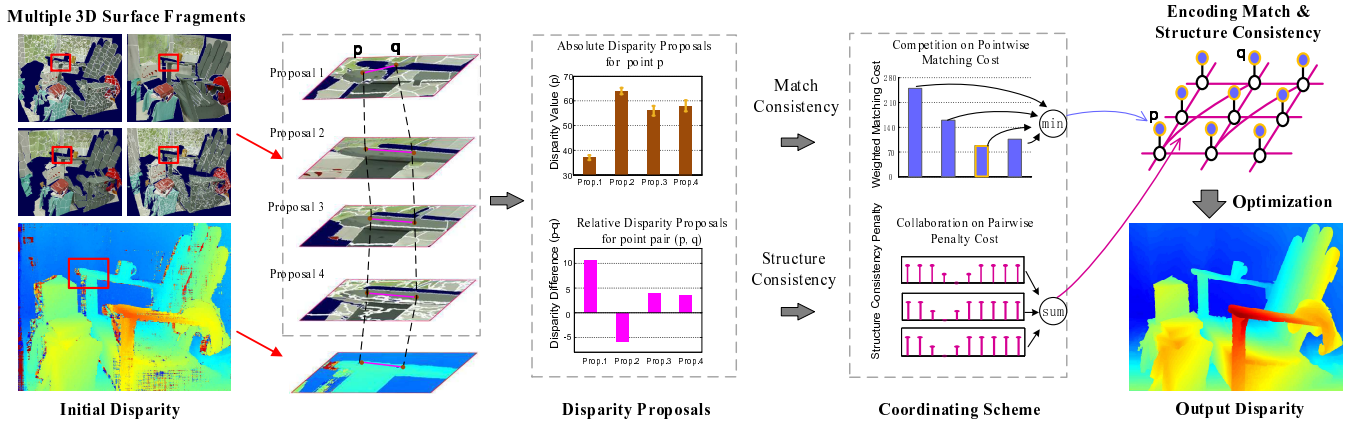


Figure 1: **Flow Chart of Our Approach.** Given the initial disparity together with 2D segments, we can obtain a set of 3D surface fragments, as visualized using the toolkit cvkit-1.6.6[12]. Surface fragments generate absolute disparities with risks(illustrated by yellow error bars)and relative proposals. Competition scheme selects the disparity with highest matching consistency as pointwise estimation. Collaboration scheme sums the penalty costs of relative disparity proposals to enforce structure consistency.The relative proposal 1 is rejected since $\mathbf{p}$ and $\mathbf{q}$ locate in different segments. Finally a MRF model encodes the matching and structure consistency, in which long range interactions are also included. The output disparity is a global approximate solution.

3.2. Disparity Proposals

The fitted planes can be used to predict the absolute disparities of individual pixels. The predicted disparities are "flattened" version of the initial disparities, where the outliers and the high-frequency noises will be removed. The fitted planes can also be used to calculate the relative disparities between pixel pairs, which capture the geometric structure of the scene. The details of calculating the disparities are given below.

Absolute Disparity. For a pixel $\mathbf{p} = [p_x, p_y]^\top$, the absolute disparity predicted by the plane of surface fragment s is:

$$A_s(\mathbf{p}) = a_s p_x + b_s p_y + c_s. \tag{1}$$

The surface fragments, either from the same or different segmentation map, may have quite different orientations or offsets, so we will obtain a diverse set of disparity proposals. We find that there are always some good proposals inside this set. Through point-wise competition, we can find a good disparity as described in Section 3.3.

Relative Disparity. For a pixel pair $\mathbf{p}$ and $\mathbf{q}$, their relative disparity on the plane of surface fragment s is given by:

$$\begin{aligned} R_s(\mathbf{p}, \mathbf{q}) &= (a_s p_x + b_s p_y + c_s) - (a_s q_x + b_s q_y + c_s) \\ &= a_s(p_x - q_x) + b_s(p_y - q_y). \end{aligned} \tag{2}$$

Here the offset of the plane is canceled out, so the relative disparity will focus on the structure of the surface. Note that, the pixel pairs here are not restricted to adjacent pixels. The long range relations are essential for capturing the large structures of the scene.

3.3. Coordinating Model

Different surface fragments may generate quite different disparity proposals. We coordinate them through point-wise competition and pairwise collaboration.

Point-wise Competition. The absolute disparities predicted by different surface fragments may contain a lot of outliers, since many surface fragments, especially some large-scale fragments, cannot be fitted precisely into planes. Therefore we keep only the optimal proposal in a winner-take-all fashion, as follows:

$$\begin{aligned} \bar{D}_{\mathbf{p}} &= A_{s^*}(\mathbf{p}) \\ s^* &= \arg\min_{s \in \mathcal{S}_{\mathbf{p}}} W_s \cdot C(\mathbf{p}, A_s(\mathbf{p})), \end{aligned} \tag{3}$$

where $\mathcal{S}_{\mathbf{p}} = \{s | \mathbf{p} \in s\}$ is the set of surface fragments containing pixel $\mathbf{p}$. The optimal proposal is supposed to be the one with the minimum matching cost C between the left and right images, weighted by the risk W_s of surface fitting.

The risk of surface fitting is defined as $W_s = e^{-\gamma_s}$, which is negatively correlated to the confidence of surface fitting γ_s. Since only reliable initial disparities are used to fit the planes, we leverage the fraction of robust matches, $\gamma_s = |\hat{s}|/|s|$, to measure the confidence of fitting. Here $|\hat{s}|$ is the number of the valid matches within the surface fragment s, and $|s|$ is the total number of pixels in s. Weighted by W_s, the influence of occluded pixels can be eliminated in some degree.

The matching cost is a combination of the sum of absolute gradients difference and the Hamming distance of two Census transformed pixels [37], as follows:

$$\begin{aligned} C(\mathbf{p}, d) = \sum_{\mathbf{q} \in N(\mathbf{p})} \{ &| \nabla I_L([p_x, p_y]^\top) - \nabla I_R([q_x - d, q_y]^\top) | \\ &+ \lambda \mathcal{H}(T_L([p_x, p_y]^\top), T_R(\rho([q_x - d, q_y]^\top))) \}, \end{aligned} \tag{4}$$

where $\nabla I(\cdot)$ is the gradient image, $T(\cdot)$ is the Census transform. $\mathcal{H}(\cdot)$ denotes the Hamming distance between two bit

strings and λ is a constant parameter. The neighborhood N is specified to be 5×5 squares in practice.

Pairwise Collaboration. From Eq. 2 we can obtain a set of proposals for the relative disparities, each from a regularized surface fragment. These proposals can impose constraints on the structure of the output disparity map. Theoretically, a surface fragment can generate a proposal for the relative disparity between any two pixels. We assume that the majority of multiple proposals is a good estimate for the disparity, while the error of the proposal is uncontrollable if the pixels fall outside the surface fragment. Therefore we remove these unreliable proposals, and collaborate the rest of them into the following loss function:

$$E_{\mathbf{p},\mathbf{q}}(D_\mathbf{p}, D_\mathbf{q}) = \sum_{s \in \mathcal{S}_\mathbf{p} \bigcap \mathcal{S}_\mathbf{q}} \phi(|D_\mathbf{p} - D_\mathbf{q} - R_s(\mathbf{p},\mathbf{q})|), \quad (5)$$

where ϕ maps the fitting errors $\varepsilon_{\mathbf{p},\mathbf{q}} = D_\mathbf{p} - D_\mathbf{q} - R_s(\mathbf{p},\mathbf{q})$, measured in pixels, into three discrete penalties:

$$\phi(|\varepsilon_{\mathbf{p},\mathbf{q}}|) = \begin{cases} 0 & \text{if} \quad \varepsilon_{\mathbf{p},\mathbf{q}} = 0 \\ \beta_1 & \text{if} \quad \varepsilon_{\mathbf{p},\mathbf{q}} = \pm 1 \\ \beta_2 & \text{otherwise} \end{cases}. \quad (6)$$

We ensure that $\beta_2 \geq \beta_1$. Through minimizing the loss in Eq.5, the structure of the output disparity map will be consistent to the prior surface structure encoded by $R_s(\mathbf{p},\mathbf{q})$.

As multiple proposals jointly determine the final value, our pairwise term enjoys two advantages: (1) If two pixels $\mathbf{p}$ and $\mathbf{q}$ always locate in two different fragments, i.e., $\mathcal{S}_\mathbf{p} \bigcap \mathcal{S}_\mathbf{q} = \emptyset$, a scene boundary may possibly exist between this two pixels. Cutting off all the interactions between $\mathbf{p}$ and $\mathbf{q}$ keeps the boundary from being over-smoothed. (2) If $\mathbf{p}$ and $\mathbf{q}$ are separated to different fragments in some of the segmentations, i.e., $0 < |\mathcal{S}_\mathbf{p} \bigcap \mathcal{S}_\mathbf{q}| < M$, parts of the artificial edges raised by over-segmentation can be eliminated via voting over multiple proposals.

3.4. Objective Function

To integrate pointwise and pairwise proposals into a global model, we formulate the inference of the optimal disparity map by a MRF model, as follows:

$$E(D) = \sum_\mathbf{p} E_\mathbf{p}(D_\mathbf{p}) + \sum_\mathbf{p} \sum_{\mathbf{q} \in \mathcal{N}(\mathbf{p})} E_{\mathbf{p},\mathbf{q}}(D_\mathbf{p}, D_\mathbf{q}). \quad (7)$$

The unary term is defined as:

$$E_\mathbf{p}(D_\mathbf{p}) = |D_\mathbf{p} - \bar{D}_\mathbf{p}|, \quad (8)$$

which measures how well the disparity map D agrees with the estimated disparity map $\bar{D}$ (Eq. 3).

The binary terms are inherited from the loss function in Eq. 5. Note that only the proposals generated by the surface fragments that contain both two pixels are taken into account. This arrangement derives a new type of neighborhood that two pixels are connected as neighbors if and only if they occur in at least one surface fragment, i.e., $\mathcal{N}(\mathbf{p}) = \{\mathbf{q} | \mathcal{S}_\mathbf{p} \bigcap \mathcal{S}_\mathbf{q} \neq \emptyset\}$. The extent of this neighborhood is consistent with the shape of fragment, which is more natural than the traditional 4-connected or 8-connected neighborhood. Moreover, the long range relations supply critical constraints to deal with large textureless regions and to predict the disparities in occluded regions.

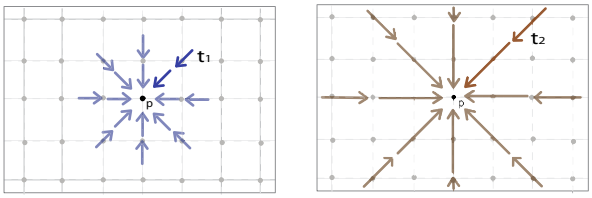


Figure 2: **Aggregation from different directions, each direction is decomposed into Markov chains of various spatial step sizes.**

4. Optimization

The optimal disparity map can be obtained by minimizing the energy in Eq. 7. This is a NP-hard problem. Widely used semi-global matching (SGM) [13] approximates the optima by aggregating matching costs in 1D from all directions. We follow the SGM pipeline to divide the 2D problem into a lot of 1D problems defined on different directions. However, non-local connections are encoded in our MRF model, that is, the 1D problems have pairwise constraints over point pairs at different distances. We further approximate the high-order 1D graph by many Markov chains as shown in Fig. 2. Each Markov chain $\mathbf{t}$ is of selected, fixed step size, and 1D dynamic programming (DP) can find the optimal solution and minimum costs for it. The cost $\mathbf{E}^\mathbf{t}(\mathbf{p}, D_\mathbf{p})$ along chain $\mathbf{t}$ of a pixel $\mathbf{p}$ at disparity $D_\mathbf{p}$ is calculated recursively by DP as

$$\mathbf{E}^\mathbf{t}(\mathbf{p}, D_\mathbf{p}) = E_\mathbf{p}(D_\mathbf{p}) + \min_d \{E_{\mathbf{p}-\mathbf{t},\mathbf{p}}(d, D_\mathbf{p}) + \mathbf{E}^\mathbf{t}(\mathbf{p}-\mathbf{t}, d)\} \quad (9)$$

where $\mathbf{p} - \mathbf{t}$ is the pixel before $\mathbf{p}$ specified by Markov chain $\mathbf{t}$, and d is the scalar disparities of $\mathbf{p} - \mathbf{t}$.

The output disparity for a pixel $\mathbf{p}$ is the one that minimizes the total aggregate cost $\mathbf{E}^A(\mathbf{p}, D_\mathbf{p})$, which is calculated by summing over the costs $\mathbf{E}^\mathbf{t}(\mathbf{p}, D_\mathbf{p})$ along all the chains of different directions and step sizes:

$$\mathbf{E}^A(\mathbf{p}, D_\mathbf{p}) = \sum_{\mathbf{t} \in \mathcal{T}} \mathbf{E}^\mathbf{t}(\mathbf{p}, D_\mathbf{p}), \quad (10)$$

where $\mathcal{T}$ is a set of Markov chains along all directions with various step sizes. However, considering all chains is computationally infeasible, we therefore pre-define some specific Markov chains to reduce the connectivity of the original graph. We had tested several configurations of chains,

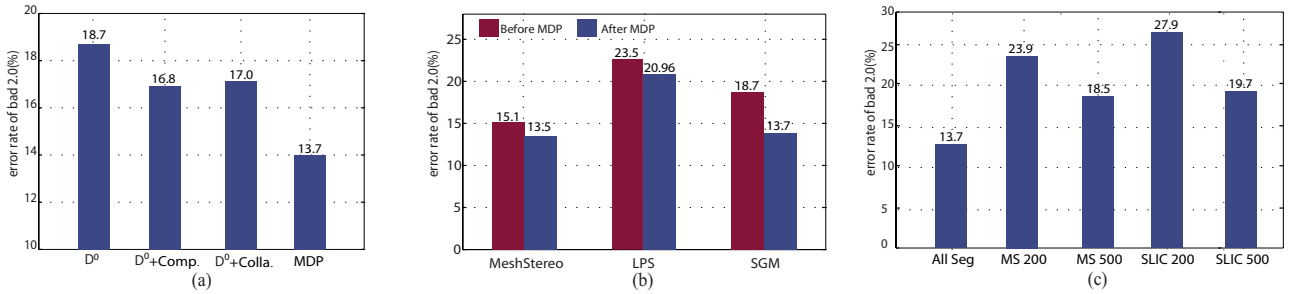


Figure 3: **Error rates for empirical analysis.** (a) The error rates for evaluating the effect of coordinating scheme. (b) The error rates for investigating the influence of initial disparities. (c) The error rates for comparing MDP with individual disparity proposals.

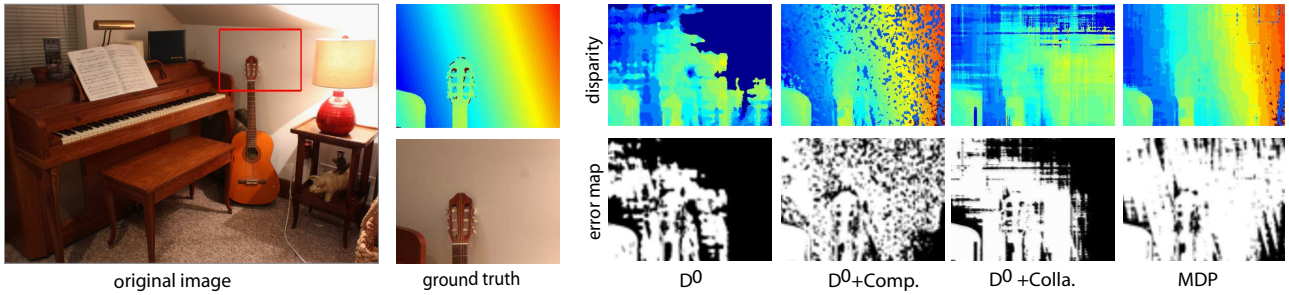


Figure 4: **Effect of coordinating Scheme.**

and finally chose 4 directions with 2 step sizes as a tradeoff between effectiveness and efficiency.

The modified SGM inherits the advantages of computational efficiency, accuracy and simplicity from the original SGM. Specifically, when finding the minima in Eq. 9, d can be restricted to a narrow searching scope. For given $D_{\mathbf{p}}$ and a relative proposal $R_s(\mathbf{p}-\mathbf{t},\mathbf{p})$, only 4 states of d need to be taken into consideration according to Eq. 6. Thus the recursive procedure of Eq. 9 requires $O(D_{max})$ steps at each pixel, where D_{max} is the total number of disparity levels. The total complexity is $O(WHD_{max})$ with W as image width and H as image height.

5. Experiments

Our approach that coordinates **M**ultiple **D**isparity **P**roposals is denoted by MDP. In this section, we introduce the experimental setup, conduct a set of breakdown analysis to investigate different aspects of our method, and compare our method with the state-of-the-art approaches.

5.1. Experimental Setup

Implementation details. To generate the 3D surfaces, we first apply Meanshift [8] and SLIC [1] for color image segmentation. Each method generates two segmentation maps with different scales: one has about 200 super-pixels and the other has about 500 super-pixels. We therefore obtain $M = 4$ segmentation maps. The initial disparity D^0 is obtained by the results of SGM, which is an implementation from Yamaguchi et al. [35] and we accelerate the code using AVX512 instruction set. We find that such initialization is slightly better than the direct matching cost without semi-global smoothing.

Parameters are set empirically: λ in Eq. 4 is set to $1/16$. Constant penalties β_1 and β_2 in Eq. 6 are set to 2.5 and 30 respectively. When using RANSAC, the outlier threshold is set to 1 pixel. In optimization, the dynamic programming is performed along 4 directions, each with 2 spatial step sizes: 10 pixels and 1 pixels to capture long range relationships and short range relationships. The estimated scalar disparity is accurate to 1 pixel.

Dataset and evaluation protocol. We analyze our method on 2014 Middlebury stereo dataset [23]. This dataset contains 30 image pairs, where 15 image pairs with available disparity ground-truth are used for training while other 15 image pairs are used for online evaluation. Compared with older version Middlebury dataset, it provides more images with higher quality. Most image pairs have large amount of textureless regions and are imperfect rectified, making the dataset very challenging.

Following the standard evaluation protocol, we use error rate (%) for error threshold of 2.0 pixels at non-occluded regions as the evaluation metric. We present a selection of important results, readers can refer to [24] for more results.

5.2. Empirical Analysis

We systematically analyze each component over the training set, which provides the ground truth disparities that enable us to evaluate our approach step by step.

The effect of coordinating scheme. The coordinating

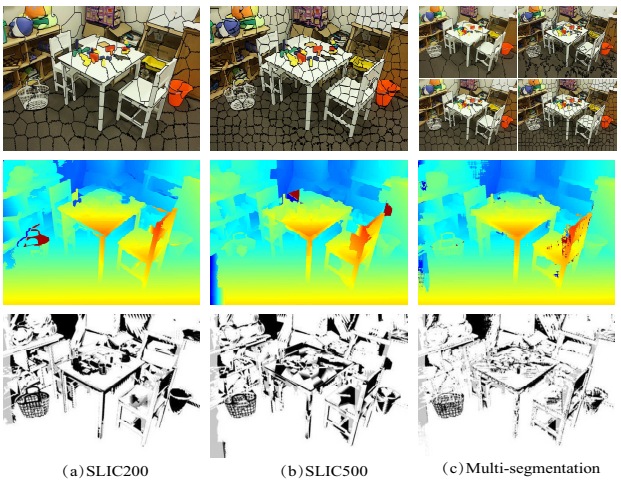


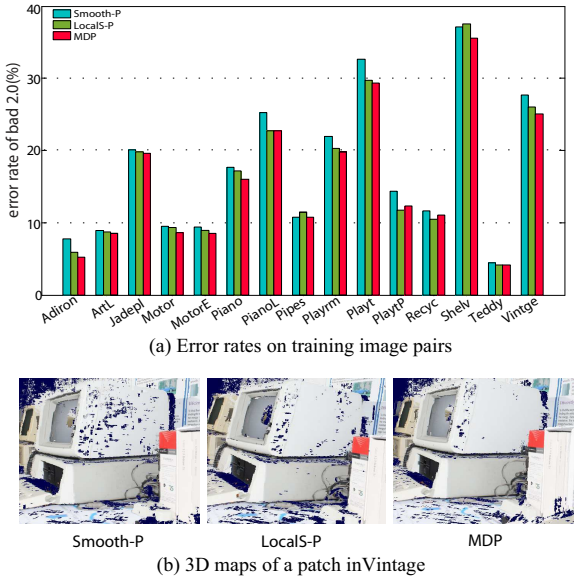
Figure 5: **Qualitative evaluation on multiple segmentation proposals.**

scheme consists of two parts: point-wise competition and pairwise collaboration. To investigate the effect of each part, we demonstrate four stepwise results: the initialization disparity D^0, the result of only performing point-wise competition (D^0+Comp.), the result of only performing pairwise competition (D^0+Colla.) and the result of performing both point-wise competition and pairwise collaboration (MDP).

As illustrated in Fig. 4, the initial disparity D^0 has great errors on the extracted textureless region. We therefore utilize multiple proposals. By only performing point-wise competition scheme, the disparities in D^0+Comp. can select the reliable disparity estimations, but there remains a lot of wrongly estimated points because all the M proposals can be wrong. By only performing pairwise competition, the disparities in D^0+Colla. reflect the structure of scene, such as the wall at the right of guitar, but the estimated absolute disparities may not be accurate without regarding the point-wise matching consistency. MDP combines the point-wise competition and pairwise collaboration, therefore yields relative satisfactory results.

The average error rates over 15 image pairs in the training set have been reported in Fig. 3 a. It can be seen that performing point-wise competition and pairwise collaboration independently can reduce the error rate by 1.9% and 1.7% respectively, and performing them simultaneously can reduce the error rate by 5%. The results suggest the complementary ability of the competition and collaboration.

The effect of multiple segmentation proposals. To verify the necessity of coordinating multiple segmentation proposals, we compared MDP with 4 original disparity proposals that are induced by 4 segmentation methods. The disparity proposals corresponding to Meanshift with 200 superpixels, Meanshifit with 500 superpixels, SLIC with 200 superpixels and SLIC with 500 superpixels are denoted by MS200, MS500, SLIC200, SLIC500, respectively. As shown in Fig. 3 c, MDP has the lowest average error rates over other 4 single disparity proposals. The results indicate that our coordinating strategy can effectively exploit reliable information from different proposals.



(a) Error rates on training image pairs

(b) 3D maps of a patch inVintage

Figure 6: **Effect of surface structure prior.**

A concrete example is displayed in Fig. 5. According to the error maps in the last row, employing larger scale segmentation (SLIC200) can help recover the textureless regions, i.e., the table surface, but fails at the objects on the table with fine structures. Employing smaller scale segmentation (SLIC500) fails at large textureless regions but excels at recovering smaller structures. Our method integrates the advantages of multiple segmentation proposals, as shown in the last error map, both textureless regions and tiny structures can be recovered.

Influence of Initial Disparity. To investigate the influence of different disparities, we apply different types of initializations for our method. In addition to SGM, we also utilize the results of MeshStereo [39] and LPS [26]. The average error rates of different initial disparities and the disparities after MDP are demonstrated in Fig. 3 b. We find that our method is not just effective on a specified initial method but can generally reduce the error of different disparity methods. The results indicate that our method is complementary to previous stereo computation methods, and it can potentially incorporate more types of initialization to further improve our method.

Surface Structure Prior and Long Range Constraints. The pairwise term in Eq. 7 is distinguished from those of previous methods in two aspects. (1) We adopt a structure preserving prior instead of first-order smoothness prior. (2) We can impose the long range constraints while most methods can only consider the constraints between adjacent points. To show the improvement due to these differences, we construct two variants Smooth-P and LocalS-P. Smooth-

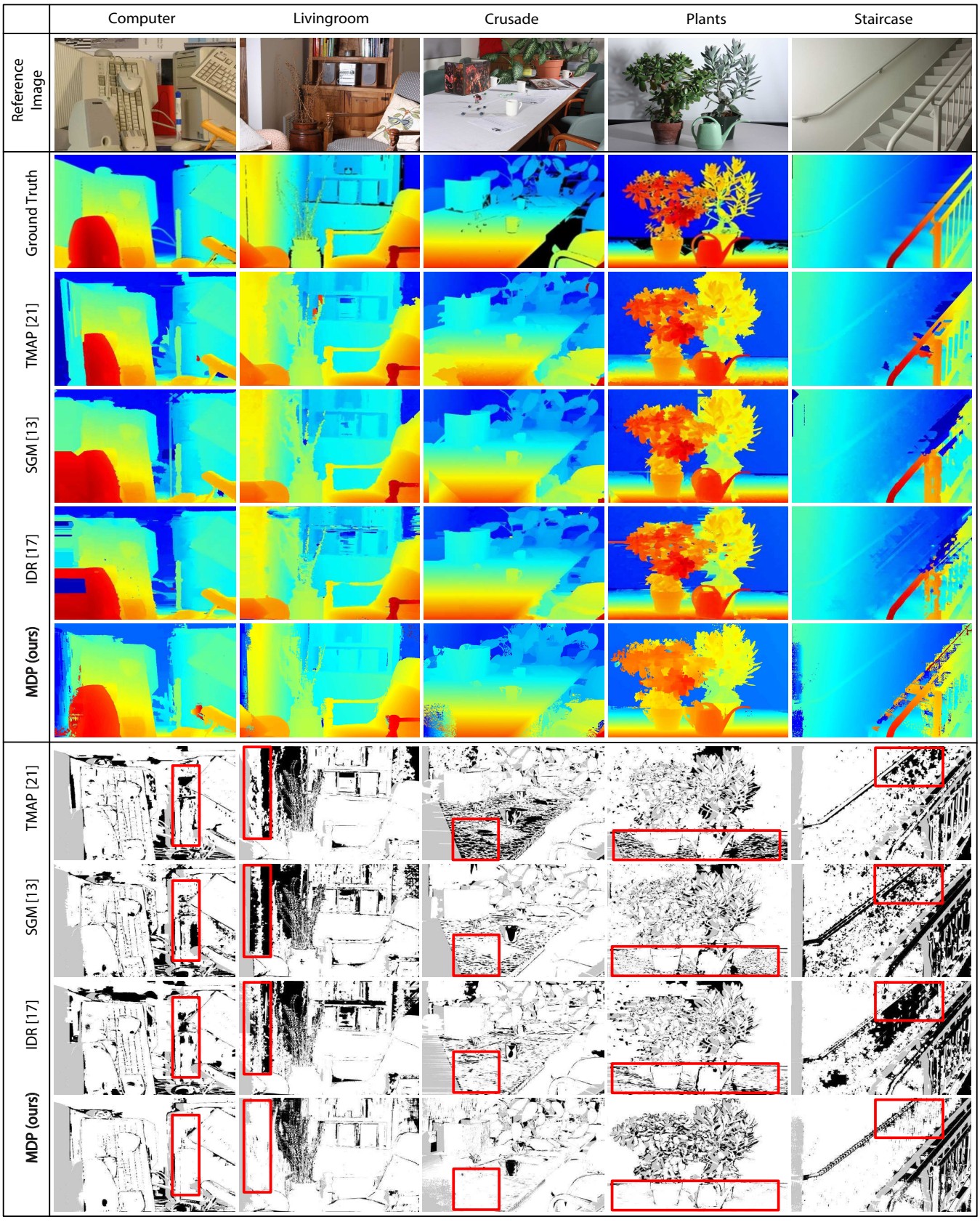


Figure 7: **Qualitative Comparison on Datasets from the Middlebury Stereo Benchmark.**

Name	Res	Avg	Austr	AustrP	Bicyc2	Class	ClassE	Compu	Crusa	CrusaP	Djemb	DjembL	Hoops	Livgrm	Nkuba	Plants	Stairs
MC-CNN-acrt	H	$\mathbf{8.29_1}$	$\mathbf{5.59_1}$	4.55_4	$\mathbf{5.96_1}$	$\mathbf{2.83_1}$	11.4_2	8.44_2	$\mathbf{8.32_1}$	8.89_2	$\mathbf{2.71_1}$	$\mathbf{16.3_1}$	$\mathbf{14.1_1}$	13.2_2	$\mathbf{13.0_1}$	$\mathbf{6.4_1}$	$\mathbf{11.1_1}$
MDP(Ours)	H	12.6_2	14.4_5	4.99_7	10.6_9	10.7_4	27.2_3	$\mathbf{8.11_1}$	12.5_5	$\mathbf{8.07_1}$	4.27_2	30.4_6	20.5_2	$\mathbf{12.6_1}$	17.8_2	13.4_3	17.3_2
MeshStereo[39]	H	13.4_3	5.90_2	4.88_5	10.8_{10}	12.9_7	$\mathbf{10.6_1}$	13.6_4	12.2_4	9.01_3	5.39_5	27.4_3	23.5_4	17.7_3	21.0_7	15.4_7	20.9_4
LCU	Q	17.0_4	24.7_7	7.59_{11}	11.6_{12}	11.9_5	27.9_4	14.0_5	19.3_6	15.8_8	8.10_{15}	36.1_{10}	29.1_8	21.3_6	18.4_3	14.1_4	23.8_6
TMAP [21]	H	17.1_5	20.2_6	4.94_6	8.13_5	12.8_6	30.0_5	14.1_7	27.9_{11}	20.4_{12}	5.09_3	31.5_8	23.1_3	20.9_5	19.0_4	18.8_{11}	18.0_3
IDR [17]	H	18.4_6	37.5_{14}	$\mathbf{4.08_1}$	7.49_3	23.3_{14}	40.6_8	15.7_{14}	24.5_7	11.3_7	5.46_7	33.1_9	26.0_5	21.5_7	21.7_8	15.3_6	21.2_5
SGM[13]	H	18.7_7	40.3_{15}	4.54_3	8.03_4	22.9_{13}	40.5_7	14.6_10	24.7_8	10.1_5	5.40_6	29.6_6	28.5_7	23.9_8	20.0_5	14.2_5	30.9_10
LPS[26]	H	19.4_8	6.14_3	5.34_8	9.24_6	7.53_2	96.0_{25}	15.0_{12}	9.61_2	9.40_3	5.18_4	92.4_{25}	27.4_6	24.3_{11}	23.0_{10}	10.0_2	25.6_8
LPS[26]	F	20.3_9	6.72_4	6.06_9	9.72_7	9.87_3	94.3_{24}	14.1_6	11.2_3	11.2_6	5.88_9	89.3_{24}	36_{12}	20.5_4	23.8_{12}	16.0_8	25.4_7

Table 1: **Quantitative Evaluation on the Middleburry Stereo Benchmark at 2 Error Threshold.** Our method is ranked at 2nd out of 25 competing methods at time of submission. Evaluation is only performed on non-occlusion regions. In each cell the number denotes bad pixel rate, Res column denotes the image resolution the method works on, and subscript denotes ranking. We highlight in bold when our approach outperforms the state-of-the-art.

P changes the pairwise term of MDP to traditional first order smoothness prior, and LocalS-P only imposes the short range structure prior for the pairwise term.

We visualize the 3D colored point clouds generated from disparities in Fig. 6 b. It can be seen that there are a lot of holes on the 3D reconstruction model when using the first-order smoothness prior. That is, models with first-order smoothness prior always fail to recovery slanted surfaces. LocalS-P reduce the holes, which indicates even short range structure prior can help to recover the slanted surface with weak texture. The fact that MDP further reduces the holes reveals the effectiveness of long range constraints. More quantitative comparison are demonstrated in Fig. 6 a. For most of the sequences, imposing the long range structure prior can achieve best results. We adopt the plane model for 3D proposals, which makes our method excel at recovering slanted surfaces even with weak texture, but it may fail to predict the curved surface. In the future, we will adopt the quadratic model to propose the disparity.

Runtime We require 3.1 sec/megapixels (s/mp) for initialization, 14.3s/mp for segmentation and plane-fitting in 4 scales, 1.9s/mp for graph construction and 5.2s/mp for optimization, thus 24.5s/mp in total on a 3.4GHz CPU.

5.3. Comparison to State-of-the-Art

Results on 2014 Middlebury. The quantitative comparison to the current state-of-the-art on the Middlebury stereo benchmark are summarized in Table 1 and the full version can be checked on the Middlebury evaluation website[1]. Our method is currently (October 2015) ranked at the second place amongst 25 competitors for error threshold 2.0. There are 3 out of 15 images where we are ranked at the first as highlighted in Table 1. All of the three images have large textureless slant planes. It is evident that our method outperforms others in such regions.

Fig. 7 shows some qualitative results of our method. As evidenced by the error map, it is difficult for other methods,such as SGM which imposes traditional first-order smoothness constraints to recover slanted surfaces. However, our method excels at estimating slanted surfaces even with weak texture (marked by red rectangles) due to the surface structure constraints and long range prior.

Results on KITTI 2015. To verify the performance across dataset, we additionally evaluate overall results on KITTI 2015 dataset. This dataset contains 200 training and 200 test image pairs of outdoor scenes. Compared to 2014 Middlebury dataset, it is relatively more challenging for us since accurate segmentation is harder.

The error rates of the non-occluded and total areas on the training set are 4.79% and 5.13%, compared to 7.33% and 8.94% of the initial disparity D^0, our method can consistently improve the baseline even on another type of dataset. We achieve an error rate over total areas of 5.36% on the test set and rank 8th on the KITTI 2015 leaderboard[2].

6. Conclusion

We have presented a novel stereo method which coordinates multiple disparity proposals. The disparity proposals are generated from 3D surface fragments based upon multi-segmentation, and then coordinated by point-wise competition and pairwise collaboration into a MRF model with long-range connections. In the experiments, we have shown that our method can integrate advantages of multiple segmentation proposals. Then we verified the complementary ability of the two components of our coordinating scheme. Further more, with the long range constraints, our method has the ability to deal with the low texture regions. Rankings on 2014 Middlebury and KITTI 2015 stereo datasets indicate that our method has achieved comparable results to state-of-the-art.

Acknowledgement

This work was supported by National Basic Research Program of China (No.2015CB351703), National Natural Science Foundation of China (No.61573280, No.61231018), and 111 Project (No.B13043).

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art

[1] http://vision.middlebury.edu/stereo/eval3/

[2] http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark\=stereo

superpixel methods. *IEEE Trans. PAMI.*, 34(11):2274–2282, 2012.

[2] A. Blake and A. Zisserman. *Visual Reconstruction*. MIT Press, 1987.

[3] M. Bleyer, C. Rother, and P. Kohli. Surface stereo with soft segmentation. In *Proc. CVPR*, 2010.

[4] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, and S. N. Sinha. Object stereo - joint stereo matching and object segmentation. In *Proc. CVPR*, 2011.

[5] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. PAMI*, 23(11):1222–1239, 2001.

[6] A. Chakrabarti, Y. Xiong, S. J. Gortler, and T. Zickler. Low-level vision by consensus in a spatial hierarchy of regions. In *Proc. CVPR*, 2015.

[7] G. K. M. Cheung, S. Baker, and T. Kanade. Visual hull alignment and refinement across time: A 3d reconstruction algorithm combining shape-from-silhouette with stereo. In *Proc. CVPR*, 2003.

[8] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. PAMI*, 24(5):603–619, 2002.

[9] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *ACM*, 24(6):381–395, 1981.

[10] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proc. CVPR*, 2012.

[11] F. Güney and A. Geiger. Displets: Resolving stereo ambiguities using object knowledge. In *Proc. CVPR*, 2015.

[12] H. Hirschmüller. Computer vision toolkit (cvkit). `http://vision.middlebury.edu/stereo/code/`.

[13] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. PAMI.*, 30(2), 2008.

[14] L. Hong and G. Chen. Segment-based stereo matching using graph cuts. In *Proc. CVPR*, 2004.

[15] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Trans. PAMI.*, 35(2):504–511, 2013.

[16] A. Klaus, M. Sormann, and K. F. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *Proc. ICPR*, 2006.

[17] J. Kowalczuk, E. Psota, and L. C. Pérez. Real-time stereo matching on CUDA using an iterative refinement method for adaptive support-weight correspondences. *IEEE Trans. CSVT*, 23(1):94–104, 2013.

[18] M. G. Mozerov and J. van de Weijer. Accurate stereo matching by two-step energy minimization. *IEEE TIP*, 24(3), 2015.

[19] C. Olsson, J. Ulén, and Y. Boykov. In defense of 3d-label stereo. In *Proc. CVPR*, 2013.

[20] J. H. Park and H. Park. Fast view interpolation of stereo images using image gradient and disparity triangulation. In *Proc. ICIP*, 2003.

[21] E. Psota, J. Kowalczuk, M. Mittek, and L. Perez. Map disparity estimation using hidden markov trees. In *Proc. ICCV*, 2015.

[22] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. In *Proc. CVPR*, 2011.

[23] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3):7–42, 2002.

[24] D. Scharstein, R. Szeliski, and H. Hirschmüller. Middlebury stereo evaluation - version 3. `http://vision.middlebury.edu/stereo/eval3/`.

[25] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. CVPR*, 2006.

[26] S. N. Sinha, D. Scharstein, and R. Szeliski. Efficient high-resolution stereo matching using local plane sweeps. In *Proc.CVPR*, 2014.

[27] J. Sun, Y. Li, and S. B. Kang. Symmetric stereo matching for occlusion handling. In *Proc. CVPR*, 2005.

[28] X. Tan, C. Sun, D. Wang, Y. Guo, and T. D. Pham. Soft cost aggregation with multi-resolution fusion. In *Proc.ECCV*, 2014.

[29] H. Tao, H. S. Sawhney, and R. Kumar. A global matching framework for stereo computation. In *Proc. ICCV*, 2001.

[30] D. Terzopoulos. Multilevel computational processes for visual surface reconstruction. *Computer Vision, Graphics, and Image Processing*, 24(1):52–96, 1983.

[31] M. Veldandi, S. Ukil, and K. G. Rao. Robust segment-based stereo using cost aggregation. In *Proc. BMVC*, 2014.

[32] D. Wei, C. Liu, and W. T. Freeman. A data-driven regularization model for stereo and flow. In *Proc. 3DV*, 2014.

[33] O. J. Woodford, P. H. S. Torr, I. D. Reid, and A. W. Fitzgibbon. Global stereo reconstruction under second order smoothness priors. In *Proc. CVPR*, 2008.

[34] K. Yamaguchi, T. Hazan, D. A. McAllester, and R. Urtasun. Continuous markov random fields for robust stereo estimation. In *Proc.ECCV*, 2012.

[35] K. Yamaguchi, D. A. McAllester, and R. Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *Proc.ECCV*, 2014.

[36] Q. Yang. A non-local cost aggregation method for stereo matching. In *Proc.CVPR*, 2012.

[37] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In *Proc. ECCV*, 1994.

[38] C. Zhang, Z. Li, R. Cai, H. Chao, and Y. Rui. As-rigid-as-possible stereo under second order smoothness priors. In *Proc. ECCV*, 2014.

[39] C. Zhang, Z. Li, Y. Cheng, R. Cai, and Y. Rui. Meshstereo: A global stereo model with mesh alignment regularization for view interpolation. In *Proc. ICCV*, 2015.