Toward perceptually-consistent stereo: A scanline study

Jialiang Wang
Harvard University
jialiangwang@g.harvard.edu

Daniel Glasner
AiCure *
dglasner@gmail.com

Todd Zickler
Harvard University
zickler@seas.harvard.edu

Abstract

Two types of information exist in a stereo pair: correlation (matching) and decorrelation (half-occlusion). Vision science has shown that both types of information are used in the visual cortex, and that people can perceive depth even when correlation cues are absent or very weak, a capability that remains absent from most computational stereo systems. As a step toward stereo algorithms that are more consistent with these perceptual phenomena, we re-examine the topic of scanline stereo as energy minimization. We represent a disparity profile as a piecewise smooth function with explicit breakpoints between its smooth pieces, and we show this allows correlation and decorrelation to be integrated into an objective that requires only two types of local information: the correlation and its spatial gradient. Experimentally, we show the global optimum of this objective matches human perception on a broad collection of well-known perceptual stimuli, and that it also provides reasonable piecewise-smooth interpretations of depth in natural images, even without exploiting monocular boundary cues.

1. Introduction

There are two sources of shape information in a stereo pair. One is the correlation (matching) signal from smooth surface regions that are visible to both cameras, which provides direct information about depth and possibly higher-order shape. The other is the decorrelation (anti-matching) signal at regions that are visible to only one camera, which provides information about the locations and amplitudes of depth discontinuities by the half-occlusions they induce. These two sources of information are complimentary, and vision scientists have invented a variety of perceptual stimuli [23, 2, 5, 1, 34, 25, 6, 30, 31], some of which are included in Figure 1, that convincingly demonstrate how biological vision can exploit one or both of them.

In the computer vision community, there has been substantial recent progress in exploiting correlation information, both in terms of designing or learning effective local matching functions (*e.g.*, census [36, 16] and MC-CNN [39]), and in terms of developing message passing methods and other techniques for aggregating matching information across piecewise-smooth scenes. However, there has been less progress in using the decorrelation information. There were early attempts to build it into scanline algorithms [12, 5, 7, 41], but more recently, the trend has been to treat it secondarily. A typical approach is to recover an initial depth map using correlation and then "clean it up" with left-right consistency checks [35] and other post-processing schemes [9] that aim to identify and fix the regions of half-occlusion. This strategy can be very effective when tuned on particular stereo datasets [27, 28, 22], but as shown in Figure 1, it is inconsistent with the human ability to accurately perceive depth discontinuities when correlation and color/texture cues are absent or very weak.

As a step towards stereo systems that are more consistent with human perception, and hopefully more likely to generalize beyond specific datasets, this paper re-examines scanline stereo by energy minimization, and it introduces an objective that more effectively combines the two sources of stereo information. The key idea is to represent the disparity map (and thus the depth map) as a piecewise smooth function over the visual field, with "smoothness" specified by any basis of a low-dimensional function space. In a single scanline, the stereo problem is formalized as identifying the best piecewise smooth cut through disparity space, as specified by a finite number of boundary locations and the shape coefficients for the smooth pieces between them. By explicitly representing the depth discontinuities at these boundary points, our approach provides a natural way to merge half-occlusion cues with correlation cues.

Although simple scanline formulations like ours do not provide state-of-the-art results on standard benchmarks, they have the important property of allowing exact global optimization by dynamic programming. This allows us to study the properties of a stereo objective without the complications of approximate inference and local minima. Following this approach, we find that an objective using

*Most work was done when Daniel Glasner was at Harvard University

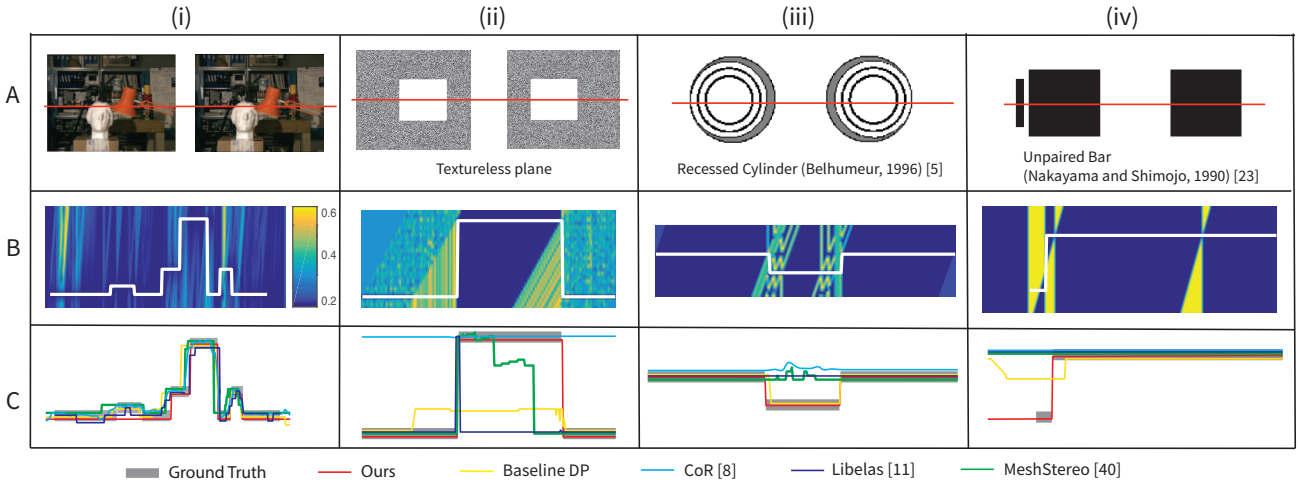

Figure 1. **Correlation and decorrelation cues in a stereo scanline.** Most stereo benchmarks contain images with rich matching cues (column (i)), but perceptual stimuli demonstrate that people can infer accurate depth even when matching information is absent. Columns (ii)-(iv) are examples, with row B showing pseudo-colored disparity space images and the perceived "ground-truth" disparity profiles in white. While most existing stereo methods do not match human perception on these stimuli, we introduce one that does.

only two types of local signals—correlation and the spatial derivative of correlation—is sufficient to match human perception on most perceptual stimuli, while also providing reasonable piecewise-smooth disparity profiles for images in the early Middlebury dataset [27]. These findings may inspire new ways to integrate half-occlusion into modern 2D stereo algorithms. In support of this endeavor, all the stimuli created for this paper have been made available at: http://vision.seas.harvard.edu/stereo.

2. Related work

An early perceptual study of depth from stereoscopic decorrelation is due to Gallium and Borsting [13], who used evidence from randomdot stereograms to reject the prevailing idea that unmatched regions are noise that disturb fusion [13, 15]. Other landmark studies were performed by Nakayama and colleagues [23, 2], who coined the phrase "da Vinci stereopsis" for the recovery of depth from half-occlusions (as opposed to depth from matching). They suggest that half-occlusions are processed early and in conjunction with matching, providing local information about both the location and amplitude of depth discontinuities. Anderson and Nakayama [2] also suggest a design for receptive fields that could serve as stereo boundary detectors, and the decorrelation gradient that we use in Section 3 is inspired by this design. There is evidence that some sort of detector like these exists in V2 [32], and there is speculation that such things could be built from local combinations of phase-based and displacement-based disparity tuned cells [29, 3].

In computer vision it is more common to handle half-occlusion in post-processing, after an initial depth map is constructed. Egnal and Wildes [9] provide a summary of these techniques. Most prevalent is the left-right consistency check, which duplicates the stereo effort, and compares two depth maps that are independently constructed from left and right viewpoints. There are noteworthy exceptions that isolate or label half-occluded pixels *during* matching [19, 24], but they do so without enforcing the geometry of the depth discontinuities that these pixels would imply.

One place where decorrelation has been more naturally incorporated into stereo processing is in scanline algorithms. Like ours, these are formulated as energy minimization and optimized by dynamic programming, with the output being a cut through disparity space [27]. Belhumeur [5] observed that rectified cameras allow half-occlusions to be represented as $45°$ "shadows" cast from occluding boundary points in disparity space, and he used this to build a Bayesian approach with priors for piecewise smoothness. The same half-occlusion geometry plays a useful role in our paper. The method of Bobick and Intille [7] is another notable precursor to ours, and part of our motivation is to generalize beyond their piecewise constant disparity profiles by allowing a more general notion of smoothness, and to reduce reliance on monocular cues and control points.

Like other scanline formulations, our approach uses a pre-computed correlation cost in disparity space. Since we focus on well-controlled input, we find it sufficient to use a simple correlation measure based on absolute intensity differences. In an applied stereo system, this pre-computed cost would come from a different correlation measure, such as one based on contrast polarities [16, 36] or learned from data [38, 21, 20, 37, 14]. Another notable aspect of our study is that it ignores monocular boundary cues from texture and color. An applied system would incorporate these

as well, because they are complimentary to the decorrelation cues that we study here, and are critical to obtain high performance on standard stereo benchmarks (*e.g.*, [22, 26]).

3. Correlation and decorrelation

Our input is a pair of stereo images that is rectified, meaning that the two effective cameras have parallel optical axes and equal focal lengths, and that all corresponding left and right pixels are contained in corresponding scanlines. As is common, we align the visual field with the left image, so that the output on a scanline is a disparity profile that is a scalar function defined on the left image plane. (This choice of visual field sacrifices some of the half-occlusion information that would be available in a cyclopean visual field, but it has the advantage of enabling global optimization by dynamic programming.) On a scanline, we assume a discrete domain for the visual field, indexed by $n \in \{1 \dots N\}$. For notational convenience, we choose a flipped coordinate system, with $n = 1$ the right-most pixel in the scanline and $n = N$ the left-most one.

We represent disparity profiles as piecewise smooth functions. The notion of smoothness is flexible, and is specified *a priori* by choosing a small set of C^1 functions $\{B_b(n)\}_{b=1\dots M}$ that are each defined over the entire visual field. Once these global basis functions are specified, any disparity profile that is smooth can be represented as a linear combination $d(n) = \sum_{b=1}^{M} \theta(b) B_b(n)$ with shape coefficients $\theta = \{\theta(b)\}_{b=1\dots M}$. Similarly, any piecewise-smooth disparity profile can be represented by a finite set of breakpoints $s_i \in \{1 \dots N\}$ along with sets of coefficients $\theta_i \in \mathbb{R}^M$ describing the shape within each smooth interval: $\forall n \in [s_{i-1}, s_i - 1], d(n) = \sum_{b=1}^{M} \theta_i(b) B_b(n)$. As will become clear, the complexity of our optimization scheme grows quickly with the dimension M of the shape space, so we mainly discuss the piecewise constant case ($B_1(n) = 1$) and the piecewise linear case ($B_1(n) = 1, B_2(n) = n$) in the sequel. Figure 2 shows an example of the former. We use notation $d_\theta(n)$ for the disparity profile, defined over the entire visual field, that is associated with a given set of shape parameters, *i.e.* $d_\theta(n) \triangleq \sum_{b=1}^{M} \theta(b) B_b(n)$.

Finding a good disparity profile in a scanline amounts to simultaneously finding a good partition of the visual field into intervals $\mathcal{I} = \{[s_{i-1}, s_i - 1]\}_{i=1}^{|\mathcal{I}|+1}$, $s_0 = 1, s_{|\mathcal{I}|+1} - 1 = N$ with accompanying shape coefficients $\Theta = \{\theta_i\}_{i=1}^{|\mathcal{I}|}$. We want to formalize this in terms of an objective $\mathcal{L}(\mathcal{I}, \Theta)$ that incorporates both correlation and decorrelation cues.

Correlation cues are handled in the usual manner. We assume there is a function that encodes the correlation between left-image receptive fields at pixel n and right-image receptive fields at pixel $n + d$. As is common, we represent this as a cost function over disparity space, $C(n, d) \in [0, 1]$ with $d \in [d_{\min}, d_{\max}] \subset \mathbb{R}$, such that a low cost $C(n, d)$ represents high correlation. For our experiments it is sufficient to use simple mean absolute differences in 3×3 windows, $C(n, d) = \frac{1}{9} \sum_{i=-1}^{1} \sum_{j=-1}^{1} |I_l(n+i, j) - I_r(n+i+d, j)|$, but in typical applications one would instead use a learned measure or one that is more robust. Regardless of how this function is defined, it immediately induces a related cost function over the M-dimensional model space via simple replication of values: $C(n, \theta) \triangleq C(n, d_\theta(n))$. In the piecewise-constant case ($M = 1$), the cost function can be visualized as in the top of Figure 2, where a "good" disparity profile consists of a piecewise horizontal cut with low integral cost.

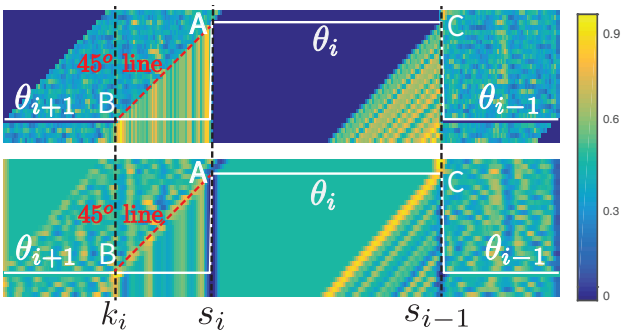

Figure 2. **Correlation and its gradient. Top:** correlation cost (absolute intensity difference) for the scanline of Fig. 1(ii). **Bottom:** proposed boundary measure G based on the correlation gradient. We see elevated gradient magnitudes (of particular polarities) at occluding boundaries A and C and at half-occlusion boundary B, which is obtained by casting a $45°$ "shadow" from A.

For decorrelation cues, we find it useful to consider the spatial derivative of the correlation cost, $\partial C / \partial n \approx f_n * C$. It is convenient to have it normalized to $[0, 1]$, say by:

$$G(n, \theta) = \left(1 + \exp\left(-\beta (f_n * C(n, \theta))\right)\right)^{-1}, \qquad (1)$$

with normalization parameter β. A visualization of this decorrelation signal G, using a horizontal nine-tap filter $\frac{1}{8}[1, 1, 1, 1, 0, -1, -1, -1, -1]$ and $\beta = 10$ is shown in the bottom of Figure 2.

When used with our piecewise smooth representation, the correlation gradient provides a convenient encoding of the local information about the location and amplitude of a depth discontinuity. To see this, consider the occlusion geometry at a discontinuity like the one at s_i in Figure 2, which separates two piecewise smooth regions with shape coefficients θ_i, θ_{i+1}. The breakpoint and shape coefficients imply two critical points in disparity space. The first is the boundary of the occluding surface, $A = (d_{\theta_i}(s_i), s_i)$. The second is the boundary of the half-occluded region on the occluded surface, $B = \left(d_{\theta_{i+1}}(k_i), k_i\right)$, whose spatial location k_i has a deterministic form $k_i = k(s_i, \theta_i, \theta_{i+1})$ thanks to the rectified camera geometry [5]: If one draws a $45°$ line in disparity space through the occluding boundary A, then

k_i is the (right-most) intersection of this line with $d_{\theta_{i+1}}(n)$.

Now, we expect an extreme value in the correlation gradient at point B because, whenever the texture of the occluded surface is distinct from the occluding one, there will be a sharp correlation change as we move along $d_{\theta_{i+1}}(n)$ from the half-occluded region to the binocular one. Similarly, we expect an elevated gradient signal at C, and at A with opposite sign. (One would expect a fourth point D similar to B if using the cyclopean visual field). Gradient signals at A and C are stronger if we use correlation measurements that encode polarity such as Census, but to a lesser degree if using simple absolute intensity differences. Nevertheless, this means that the correlation gradient can provide information about both the location and amplitude of a depth discontinuity, via the critical points A, B, C. This is quantified in the next section.

While our use of the correlation gradient at occluding points A and C may be interpreted as an embodiment of the correlation-decorrelation receptive fields proposed by Anderson and Nakayama [2], our additional use of the gradient at half-occlusion boundaries B seems new. One of its advantages is conceptual simplicity, because it implies a stereo system with only two types of local computational units: one set of disparity-tuned units to measure the binocular correlations $C(n, \theta)$ and another set to measure the finite spatial differences $G(n, \theta)$ between them. This is different from most alternative formulations of half-occlusion signals [9], such as left-right consistency checks, which would imply a greater variety of local units [31].

4. Objective and constraints

Our goal is to combine the correlation and decorrelation cues into an objective,

$$\mathcal{L}(\mathcal{I}, \Theta) = \sum_{i=1}^{|\mathcal{I}|} \rho\left(s_{i-1}, s_i, \theta_{i-1}, \theta_i\right), \tag{2}$$

that can be globally optimized by dynamic programming.

We begin by aligning the visual field with the left image plane, which ensures that half-occlusions always occur in the direction of increasing n (left in Figure 2). Breakpoints may or may not induce occlusions, depending on the sign of the discontinuity. A breakpoint s_i is an occlusion-inducing one if it satisfies $d_{\theta_{i+1}}(s_i) < d_{\theta_i}(s_i)$, in which case the location of the half-occlusion boundary is $k_i = k(s_i, \theta_i, \theta_{i+1})$ according to the 45° rule described previously. Otherwise the breakpoint satisfies $d_{\theta_{i+1}}(s_i) \geq d_{\theta_i}(s_i)$ and does not induce occlusions, and we indicate this by the convention $k_i = s_i$. Then, the objective can be written in three terms,

$$\begin{aligned}&\rho\left(s_{i-1}, s_i, \theta_{i-1}, \theta_i\right) = \\ &\quad \rho_c(s_{i-1}, s_i, \theta_{i-1}, \theta_i) + \lambda_1 \rho_g(s_{i-1}, \theta_{i-1}, \theta_i) + \lambda_2,\end{aligned} \tag{3}$$

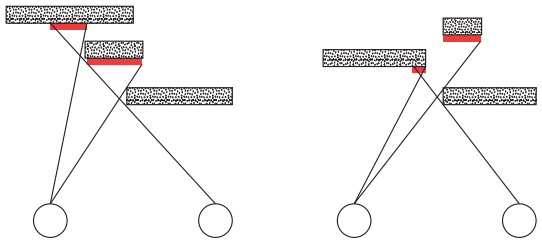

Figure 3. **Examples of disallowed events.** In addition to the ordering constraint, we prohibit the existence of depth discontinuities inside of half-occluded regions, shown here in red.

that respectively encode correlation cues ρ_c, decorrelation cues ρ_g via the correlation gradient, and a constant λ_2 that behaves like a geometric prior [17] to discourage solutions that have many small intervals.

The correlation term is simply the integral of the matching cost along the binocular portion of each interval,

$$\rho_c(s_{i-1}, s_i, \theta_{i-1}, \theta_i) = \sum_{n=k(s_{i-1}, \theta_{i-1}, \theta_i)}^{s_i - 1} C\left(n, \theta_i\right).$$

The decorrelation term measures the appropriately-signed magnitude of the correlation gradient at the occluding boundary and the half-occlusion boundary when they exist:

$$\rho_g(s_{i-1}, \theta_{i-1}, \theta_i) = \begin{cases} 1 - G(s_{i-1}, \theta_i), & \text{if } d_{\theta_i}(s_{i-1}) \geq d_{\theta_{i-1}}(s_{i-1}) \\ G(s_{i-1}, \theta_{i-1}) - G(k(s_{i-1}, \theta_{i-1}, \theta_i), \theta_i), & \text{otherwise.} \end{cases}$$

Like all dynamic programming scanline stereo algorithms, this formulation does not permit violations of the ordering constraint [4]. It furthermore assumes that each interval $[s_i, s_{i+1} - 1]$ is occluded, at most, by the interval $[s_{i-1}, s_i - 1]$ that immediately precedes it. Thus it does not allow situations like Figure 3 that have discontinuities inside of the half-occluded region. In fact, in testing on captured stereo images we have found it useful to go further than this by imposing a hard lower bound K on the size of the binocular portion of segments that has occlusion, $\forall i, s_i - k(s_{i-1}, \theta_{i-1}, \theta_i) \geq K$ if $d_{\theta_i}(s_{i-1}) < d_{\theta_{i-1}}(s_{i-1})$.

Finally, we note that our formulation requires that the slope of the disparity function $d_{\theta_i}(n)$ within each interval to be less than 45° so as not to cause a half-occluding limb. This is trivial to enforce for the one-dimensional and two-dimensional function spaces that we consider here, but for other high-order formulations, it would require greater care.

5. Optimization by dynamic programming

Our objective can be optimized by dynamic programming using an algorithm inspired by Jackson *et al.* [17]. Let

N be the number of pixels in a scanline, and let us partition the continuous space of shape coefficients, a subset of $\mathbb{R}^M$, into L bins indexed by $\ell \in \{1, \ldots, L\}$. This restricts the shape of each segment θ_i to one of L possibilities. Associated with this partition, we precompute and store a discrete $L \times N$ correlation cost table as

$$\tilde{C}(n, \ell) = \min_{\theta \in \mathcal{N}(\ell)} C(n, \theta), \tag{4}$$

with $\mathcal{N}(\ell)$ the set of shape coefficients in bin ℓ. Using the minimum operator here ensures that accuracy degrades gracefully for coarser discretizations (see Figure 6).

Let $opt(u, \ell_u)$ denote the scalar cost associated with the optimal disparity sub-profile over interval $[1, u]$ with the constraint that the final segment, the one containing u, has shape ℓ_u. For the first pixel we have $opt(1, \ell) = \tilde{C}(1, \ell) + \lambda_2$ for all $\ell \in \{1 \ldots L\}$, and from there, we can visit the remaining pixels $u \in \{2 \ldots N\}$ in sequence, recursively computing the L values of $opt(u, \cdot)$ at each pixel u. For each pair u, ℓ_u, we search for the optimal location of the *previous* breakpoint v. We set $v = 1$ if having only one segment is optimal for interval $[1, u]$. If $v \neq 1$, v will necessarily be a breakpoint between the shape of the final segment ℓ_u and a different shape, say ℓ_v, of the segment before it. Thus, we can write the recursion as

$$opt(u, \ell_u) = \min_{\Gamma} \left(\rho(v, u, \ell_v, \ell_u) + opt(v, \ell_v) \right), \tag{5}$$

where Γ is a subset of pairs (v, ℓ_v) within $[1, u-1] \times [1 \ldots L]$ that satisfy the constraints of the previous section.

To be able to build each valid subset Γ during recursion, we also maintain a record of the optimal beginnings (the minimizers of Equation 5):

$$arg(u, \ell_u) = \arg\min_{\Gamma} \left(\rho(v, u, \ell_v, \ell_u) + opt(v, \ell_v) \right). \tag{6}$$

This data structure has size $N \times L \times 2$. In our notation, $(v, \ell_v) = arg(u, \ell_u)$ means that among all possible sub-profiles defined on interval $[1, u]$ with final segment shape ℓ_u, the one with lowest cost (and cost equal to $opt(u, \ell_u)$) is smooth over interval $[v, u]$ and has a breakpoint at v marking a transition to shape ℓ_v. Using this data structure, the valid pairs (v, ℓ_v) comprising Γ can be specified as those satisfying both the half-occlusion constraint, $u - k(v, \ell_v, \ell_u) \geq K$ if $d_{\ell_v}(v) > d_{\ell_u}(v)$, and the ordering constraint[1]: $v - x > d_{\ell_v}(x) - d_{\ell_u}(v)$ with $(x, \ell_x) = arg(v, \ell_v)$, if $d_{\ell_v}(v) > d_{\ell_u}(v)$.

Once the recursion terminates at pixel $u = N$, we trace the optimal profile by using the arg data structure to accumulate the profile's breakpoints and shape-transitions, from the last breakpoint at pixel N to the first breakpoint at pixel

[1]Strictly speaking, this is stronger than the ordering constraint (sufficient but not necessary). See [33]

Algorithm 1 Find optimal disparity profile

for all $\ell \in L$ **do**
  $opt(1, \ell) \leftarrow \tilde{C}(1, \ell) + \lambda_2$
  $arg(1, \ell) \leftarrow (1, \ell)$
end for
for $u \leftarrow$ *2 to N* **do**
  for $\ell_u \leftarrow$ *1 to L* **do**
    $\Gamma \leftarrow$ BUILDVALIDSUBSET$(u, \ell_u, arg(u, \ell_u))$
    $opt(u, \ell_u) \leftarrow \min_{\Gamma} (\rho(v, u, \ell_v, \ell_u) + opt(v, \ell_v))$
    $arg(u, \ell_u) \leftarrow \arg\min_{\Gamma} (\rho(v, u, \ell_v, \ell_u) + opt(v, \ell_v))$
  end for
end for
$\mathcal{I}, u \leftarrow N$ ▷ initialize trace back
$\Theta, \ell_u \leftarrow \arg\min_{\ell} opt(N, \ell)$
while $u > 1$ **do**
  $(u, \ell_u) \leftarrow arg(u, \ell_u)$
  APPEND$(\mathcal{I}, u)$
  APPEND(Θ, ℓ_u)
end while

1. Algorithm 1 provides pseudo code. Further details about the dynamic programming algorithm and constraints can be found in a supplemental document [33].

6. Experiments

We test the algorithm on perceptual stimuli and some images from the Middlebury 2001 benchmark [27]. For the correlation gradient G we use the same nine-tap filter from Equation 4. Our parameter values are $\lambda_1 = 1, \lambda_2 = 1, \beta = 10, K = 10$, and $\lambda_1 = 0.1, \lambda_2 = 0.19, \beta = 40, K = 10$ for perceptual stimuli and natural images, respectively.

6.1. Perceptual Stimuli

We rendered twelve synthetic stereo pairs that have impoverished amounts of matching information and/or impoverished amounts of monocular boundary information. Many of these stimuli are modeled after previously published work [23, 5, 30, 2, 18] and others we created are variations of these. The collection can be used to benchmark a stereo system's ability to combine correlation and decorrelation cues in a perceptually-consistent manner. Each stimulus was rendered in Matlab from a custom-designed three-dimensional scene, so the "ground truth" disparity is precisely known. The ground truth also matches the perceptual outcomes that were reported in the respective publications.

We tested two flavors of our algorithm, one with a piecewise-constant model ($M = 1$) and another with a piecewise-linear one ($M = 2$). Our results with the piecewise-constant model are in Figures 1 and 5 and rows $1 - 5$ of Figure 4; and our results with the piecewise-linear model are in rows 6 and 7 of Figure 4. In all cases, the disparity space is defined as $[d_{gt_{min}} - 5, \cdots, d_{gt_{max}} + 5]$ with d_{gt} the ground truth disparity. Note that the aspect ratios of the visualized disparity spaces are not all equal, so that half-occlusion lines are often visualized with angles that differ

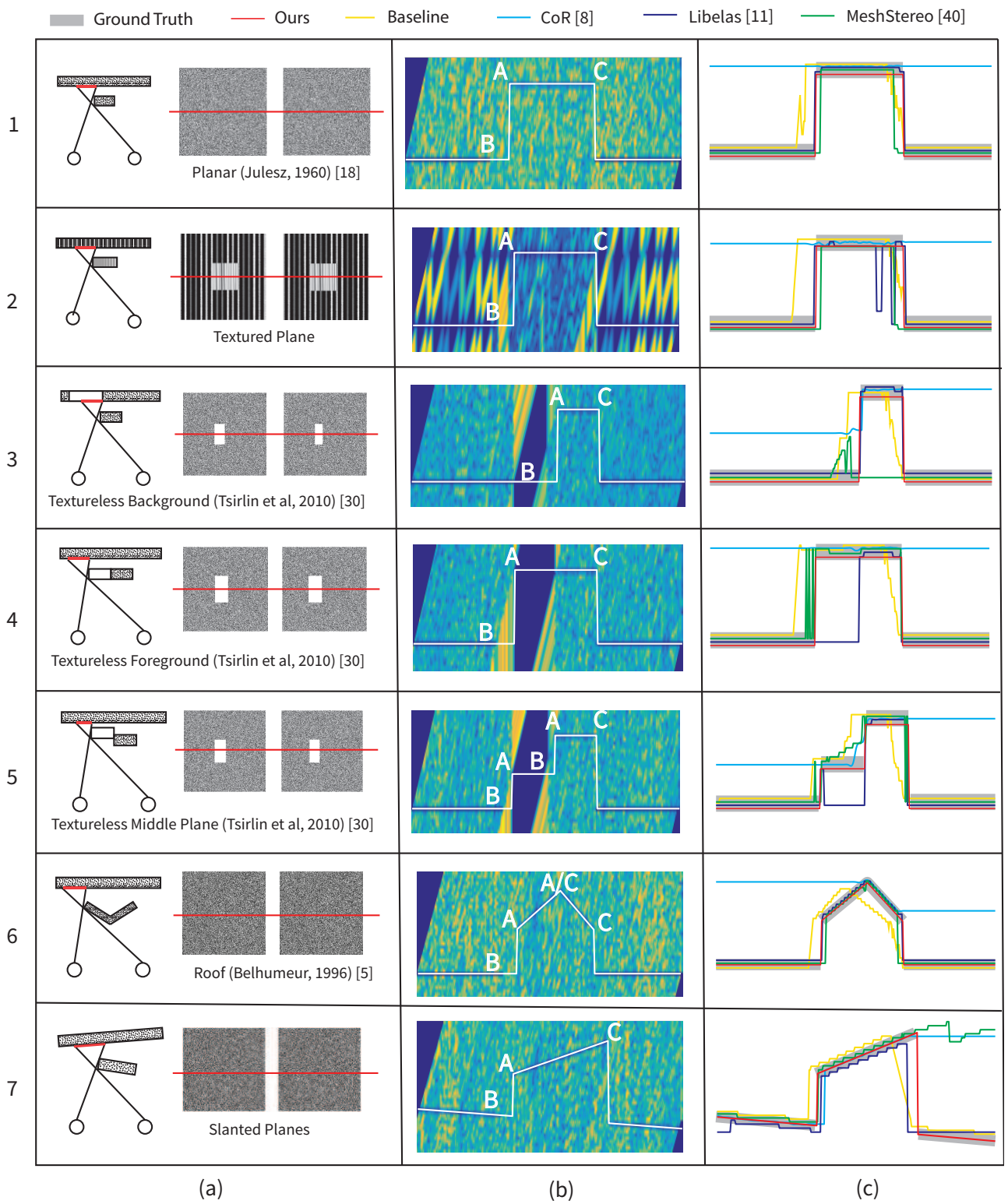

Figure 4. **Results on perceptual stimuli: (a)** Scene and rendered stereo pairs, with scanline of interest marked in red. **(b)** Ground truth disparity. A, B, C-type points have elevated correlation gradient magnitudes. **(c)** Results of some previous stereo methods and of our method with both piecewise constant (rows 1-5) and piecewise linear (rows 6 & 7) models. The latter avoids the "staircase" effect and provides high-precision depth information.

from 45°. In some cases, we mark boundaries with type A, B or C in accordance with the earlier discussion.

We find that our algorithm recovers the correct disparity for every stimulus in Figures 1 and 4. For a qualitative comparison, we applied three modern stereo systems: MeshStereo [40], CoR Stereo [8] and Libelas [11]. We also applied a basic scanline method (Baseline DP) that uses dynamic programming along a scanline to minimize a block-matching plus L2-smoothness energy from each of the left and right viewpoints, and then merges the two disparity profiles using the standard left-right consistency check [10]. For visualization purposes, when overlaid on the ground truth disparity profile, the recovered profiles are each given a unique, small additive disparity offset to separate them vertically in Figs. 1 and 4. We see that MeshStereo and Libelas seem to do well when there is sufficient correlation information (rows 1, 2 and 5 of Fig. 4), but they break down when correlation cues are absent or very weak (Fig. 1).

Figure 5 shows our results on the wallpaper stimuli of Anderson and Nakayama [2], which we interpret as a failure case. These are ambiguous stimuli, and human observers variously perceive the striped wallpaper as being either in front or behind the surrounding reference plane. Bias for one percept over the other can be induced by changing the luminance of the reference [2]: when it is very bright (row 1) the wallpaper is more often perceived as being in front; and when it is very dark (row 2) the wallpaper tends to be perceived as being behind. In contrast, we find that our algorithm produces the opposite result, because while there are strong A, B, C-type boundary signals in both configurations, these signals are stronger for the opposite interpretation. This effect could perhaps be corrected by incorporating monocular boundary information, which is certainly used by the humans but has been left out of our model.

Figure 6 studies the piecewise-linear version of our algorithm for different quantizations L of the 2D shape space. Very fine discretization ($L = 5000$) allows the dynamic programming algorithm to recover precise disparity given sufficient computation time, and due to the use of the minimum operator in Equation 4, we can trade between computation time and accuracy in a graceful manner. The reported execution times are from un-optimized Matlab implementations, and to put them in context, comparable implementations of the piecewise constant version of our algorithm and the Baseline DP method using the same number of pixels and 41 disparity levels run in 3.25s and 0.08s, respectively.

6.2. Natural images

As a sanity check, we also test the piecewise constant version of our algorithm on scanlines from the Middlebury 2001 benchmark. Some representative results are shown Figures 7 and 8. Overall, the algorithm produces reasonable approximations to the true disparity; and as expected,

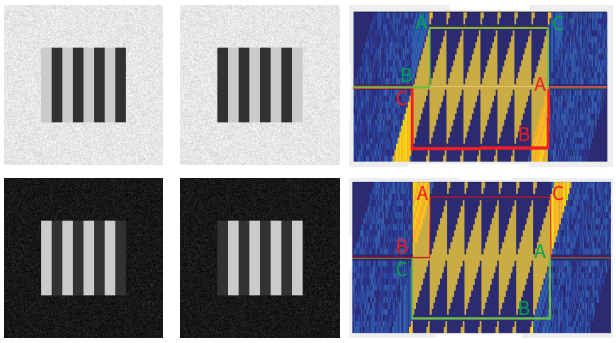

Figure 5. **Failure cases:** wallpaper stimuli from Anderson/Nakayama [2]. **Top row:** Light background. Human observers perceive as a plane in front of randomdot background. **Bottom row:** Dark background. Human observers perceive as a plane behind. Red curves are human observations and green curves are results from our algorithms. All A, B, C points in both explanations have correlation gradients G consistent with our objective. However, the incorrect explanation has stronger G.

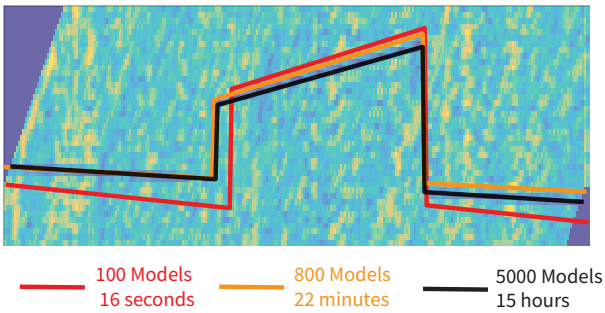

Figure 6. **Accuracy vs. runtime.** For piecewise-linear and higher-order smoothness models, the quantization L of the shape space can grow large and cause dynamic programming to run slowly. But when the quantized cost table $\tilde{C}$ is properly defined, there is a graceful trade-off between execution time and accuracy.

applying the piecewise constant model to a slanted scene produces a staircase effect (rows 3 and 6 of Figure 7). The approach tends to perform well in regions where the occluding and occluded surfaces have distinct textures (rows 1 and 2 of Figure 7) and even better when the half-occluded surface has some texture, since this produces a stronger correlation gradient at A-type points. Long segments that are textureless usually also perform well since there are high correlation gradient costs at these regions (row 3).

When the algorithm breaks down, it is for the reasons one expects. By design, it fails to recover small segments that have few mutually visible pixels, are fully half-occluded, or violate the ordering constraint (row 4). These are structures that would benefit most from 2D approaches that can exploit monocular texture and color cues for spatial grouping. Another cause of error is the simplicity of

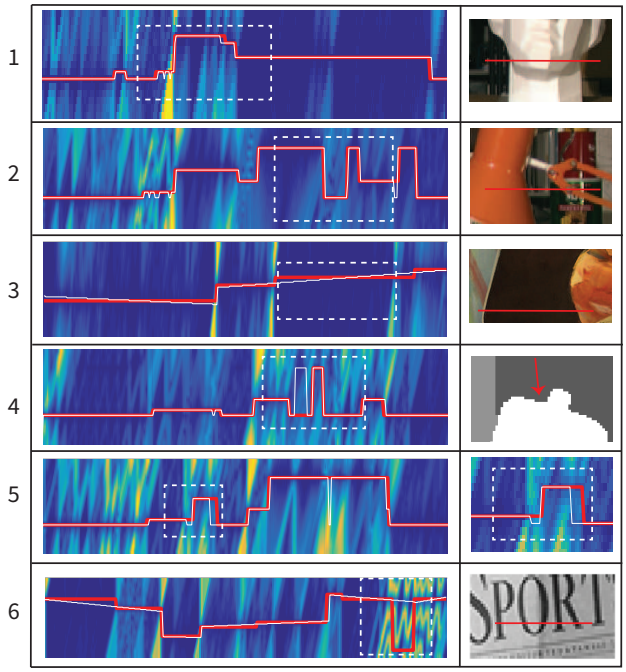

Figure 7. **Scanline results of natural images.** Left column shows scanline examples and right column shows context for highlighted region near white dotted box. **1-2:** examples of successful cases where the occluding and occluded region have distinct texture. **3:** segment changes at large textureless regions are avoided by the decorrelation measurment G. **4:** failure caused by limitations of our scanline algorithm. **5:** $\tilde{C}$ is higher at the correct model than an incorrect one. **6:** Repeated texture created by texture along a particular scanline gives strong G causing additional boundaries, which can be cleaned up using inter-scanline consistency check.

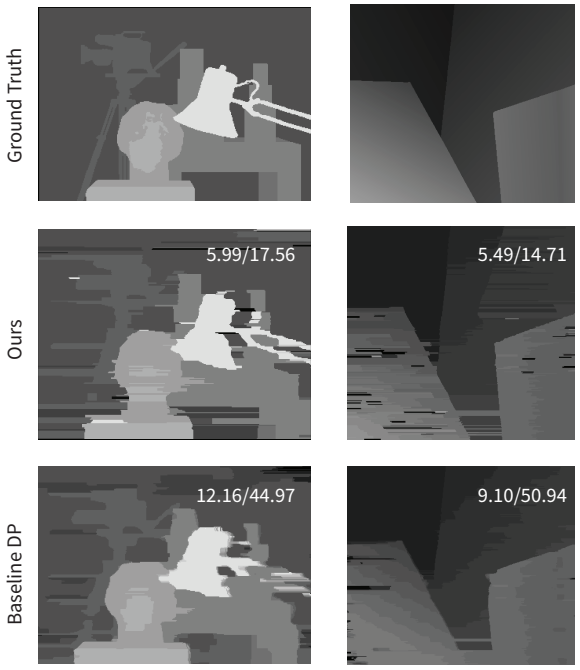

Figure 8. **Results on Venus and Tsukuba.** White error numbers are: (i) percentage of bad pixels (disparity error > 1.5) over the whole image; and (ii) percentage bad pixels over regions affected by occlusion, defined as the union of half-occluded pixels and the Middlebury-defined "boundary" pixels.

our correlation cost (absolute intensity differences), which can cause extended regions to have lower cost for the incorrect shape than for the correct shape (row 5). Also, in the presence of particular textures, the gradient of this simple correlation measure can be strong at spurious points in disparity space, which are then interpreted as additional depth discontinuities (row 6).

To provide a sense of the overall performance, Figure 8 shows the depth maps for two scenes in the dataset, Tsukuba and Venus (ignored boundary 18 pixels). These were created by running our algorithm and Baseline DP in each scanline individually, without any inter-scanline processing. Overall, our method provides a reasonable interpretation of the scene, with sharp localization of boundaries. This is in spite of the fact that it entirely ignores monocular grouping and boundary cues that are available from texture and color.

7. Conclusion

We propose a scanline stereo objective that combines matching and half-occlusion cues, and that mimics human perception of stimuli that have weak or absent correlation cues. The key to our objective is representing disparity (and thus depth) as a piecewise smooth function with an explicit set of breakpoints. This allows direct reasoning about half-occlusions, and also provides reasonable piecewise smooth interpretations of depth along scanlines in natural images, with substantial room for improvement by adding monocular boundary information (which we ignore).

Along a single scanline, the correlation gradient is a scalar quantity, but it becomes a vector quantity in two dimensions. Thus, extending its use to two dimensional stereo algorithms will require local detectors that measure the correlation gradient at multiple spatial orientations, analogous to monocular boundary detectors. The extension to two-dimensions will also require different optimization techniques that can infer piecewise smooth two-dimensional functions with explicit boundaries and per-segment shape parameters. Progress is this direction can be found in the work of Chakrabarti *et al*. [8], which shows how to infer piecewise smooth two-dimensional functions by passing sparse messages among image patches at multiple scales.

Acknowledgements. This work was funded by National Science Foundation awards IIS-1212928 and IIS-1618227.

References

[1] B. L. Anderson. The role of partial occlusion in stereopsis. *Nature*, 367(6461):365–368, 1994.

[2] B. L. Anderson and K. Nakayama. Toward a general theory of stereopsis: binocular matching, occluding contours, and fusion. *Psychological review*, 101(3):414, 1994.

[3] A. Assee and N. Qian. Solving da vinci stereopsis with depth-edge-selective v2 cells. *Vision research*, 47(20):2585–2602, 2007.

[4] H. H. Baker. Depth from edge and intensity based stereo. Technical report, DTIC Document, 1982.

[5] P. N. Belhumeur. A bayesian approach to binocular steropsis. *IJCV*, 19(3):237–260, 1996.

[6] P. N. Belhumeur and D. Mumford. A bayesian treatment of the stereo correspondence problem using half-occluded regions. In *CVPR*, 1992.

[7] A. F. Bobick and S. S. Intille. Large occlusion stereo. *IJCV*, 33(3):181–200, 1999.

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

[9] G. Egnal and R. P. Wildes. Detecting binocular half-occlusions: Empirical comparisons of five approaches. *T-PAMI*, 24(8):1127–1133, 2002.

[10] P. Fua. A parallel stereo algorithm that produces dense depth maps and preserves image features. *Machine vision and applications*, 6(1):35–49, 1993.

[11] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In *ACCV*, 2010.

[12] D. Geiger, B. Ladendorf, and A. Yuille. Occlusions and binocular stereo. *IJCV*, 14(3):211–226, 1995.

[13] B. Gillam and B. E. The role of monocular regions in stereoscopic displays. *Perception*, 17:603–608, 1988.

[14] R. Haeusler, R. Nair, and D. Kondermann. Ensemble learning for confidence measures in stereo vision. In *CVPR*, 2013.

[15] J. M. Harris and L. M. Wilcox. The role of monocularly visible regions in depth and surface perception. *Vision research*, 49(22):2666–2685, 2009.

[16] M. Humenberger, T. Engelke, and W. Kubinger. A census-based stereo vision algorithm using modified semi-global matching and plane fitting to improve matching quality. In *CVPR-W*, 2010.

[17] B. Jackson, J. D. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan, and T. T. Tsai. An algorithm for optimal partitioning of data on an interval. *Signal Processing Letters*, 12(2):105–108, 2005.

[18] B. Julesz. Binocular depth perception without familiarity cues. *Science*, 145(3630):356–362, 1964.

[19] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *ICCV*, volume 2, pages 508–515. IEEE, 2001.

[20] W. Luo, A. G. Schwing, and R. Urtasun. Efficient deep learning for stereo matching. In *CVPR*, 2016.

[21] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016.

[22] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015.

[23] K. Nakayama and S. Shimojo. Da vinci stereopsis: Depth and subjective occluding contours from unpaired image points. *Vision research*, 30(11):1811–1825, 1990.

[24] A. S. Ogale and Y. Aloimonos. Shape and the stereo correspondence problem. *IJCV*, 65(3):147–162, 2005.

[25] K. Prazdny. Detection of binocular disparities. *Biological cybernetics*, 52(2):93–99, 1985.

[26] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition*, pages 31–42. Springer, 2014.

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

[28] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *CVPR*, 2003.

[29] D. Y. Tsao, B. R. Conway, and M. S. Livingstone. Receptive fields of disparity-tuned simple cells in macaque v1. *Neuron*, 38(1):103–114, 2003.

[30] I. Tsirlin, L. M. Wilcox, and R. S. Allison. Monocular occlusions determine the perceived shape and depth of occluding surfaces. *Journal of Vision*, 10(6):11–11, 2010.

[31] I. Tsirlin, L. M. Wilcox, and R. S. Allison. A computational theory of da vinci stereopsis. *Journal of vision*, 14(7):5–5, 2014.

[32] R. von der Heydt, H. Zhou, and H. S. Friedman. Representation of stereoscopic edges in monkey visual cortex. *Vision Research*, 40(15):1955 – 1967, 2000.

[33] J. Wang, D. Glasner, and T. Zickler. A dynamic programming algorithm for perceptually-consistent stereo. *Harvard Computer Science Group Technical Report*, (TR-02-17), 2017.

[34] D. Weinshall. Perception of multiple transparent planes in stereo vision. *Nature*, 341(6244):737–739, 1989.

[35] J. Weng, N. Ahuja, and T. S. Huang. Two-view matching. In *ICCV*, 1988.

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

[37] S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. In *CVPR*, 2015.

[38] J. Zbontar and Y. LeCun. Computing the stereo matching cost with a convolutional neural network. In *CVPR*, 2015.

[39] J. Zbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(1-32):2, 2016.

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

[41] C. L. Zitnick and T. Kanade. A cooperative algorithm for stereo matching and occlusion detection. *T-PAMI*, 22(7):675–684, 2000.