Displets: Resolving Stereo Ambiguities using Object Knowledge

Fatma Güney
MPI Tübingen
fatma.guney@tue.mpg.de

Andreas Geiger
MPI Tübingen
andreas.geiger@tue.mpg.de

Abstract

Stereo techniques have witnessed tremendous progress over the last decades, yet some aspects of the problem still remain challenging today. Striking examples are reflecting and textureless surfaces which cannot easily be recovered using traditional local regularizers. In this paper, we therefore propose to regularize over larger distances using object-category specific disparity proposals (displets) which we sample using inverse graphics techniques based on a sparse disparity estimate and a semantic segmentation of the image. The proposed displets encode the fact that objects of certain categories are not arbitrarily shaped but typically exhibit regular structures. We integrate them as non-local regularizer for the challenging object class 'car' into a superpixel based CRF framework and demonstrate its benefits on the KITTI stereo evaluation. At time of submission, our approach ranks first across all KITTI stereo leaderboards.

1. Introduction

"In many cases it is sufficient to know or conjecture that an object exhibits certain regularities in order to correctly interpret its perspective image as a 3D shape."
Hermann von Helmholtz, 1867

Since David Marr's influential investigations into the human representation and processing of visual information [40], stereo matching has been considered primarily a low-level process that builds upon bottom-up representations like Marr's primal sketch. And indeed, the vast majority of binocular stereo matching algorithms that have been proposed in the literature rely on low-level features in combination with relatively simple first or second order smoothness assumptions about the world. Little is known about the importance of recognition for this problem and the leading entries in current benchmarks such as Middlebury [50] or KITTI [13] even completely ignore semantic information. This is in stark contrast to current trends in single image reconstruction [8, 19, 32, 35, 49] where recognition plays a central role. Thus how important are Helmholtz's early observations [59] for binocular stereo matching after all?

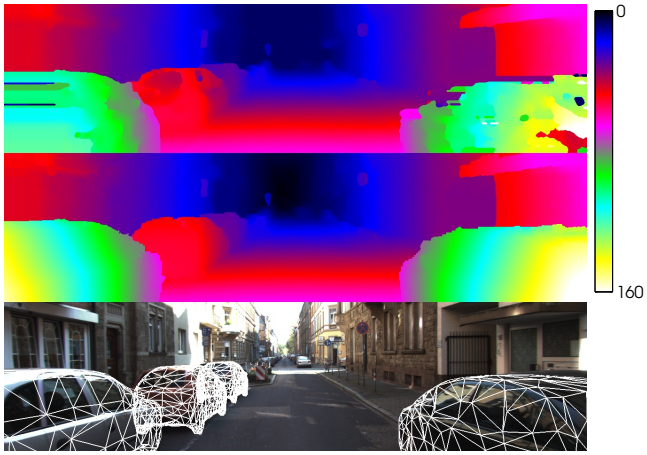


Figure 1: **Resolving Stereo Matching Ambiguities:** Current state-of-the-art stereo methods often fail at reflecting, textureless or semi-transparent surfaces (top, [68]). By using object knowledge, we encourage disparities to agree with plausible surfaces (center). This improves results both quantitatively and qualitatively while simultaneously recovering the 3D geometry of the objects in the scene (bottom).

In this paper we investigate the utility of mid-level processes such as object recognition and semantic segmentation for stereo matching. In particular, we focus our attention on the reconstruction of well-defined objects for which the data term is weak and current methods perform poorly, such as cars. Due to their textureless, reflective and semi-transparent nature, those object categories represent a major challenge for current state-of-the-art algorithms, as illustrated in Fig. 1 (top). While the reconstruction of purely specular surfaces has been successfully demonstrated using multiple frames [21,62], such techniques are hard to employ in our setting due to the superposition of several deteriorating visual effects in real-world scenes.

In contrast, as humans we are able to effortlessly extract information about the geometry of cars even from a single

image thanks to our object knowledge and shape representation. Inspired by this fact, we introduce object knowledge for well-constrained object categories into a slanted-plane MRF and estimate a dense disparity map. Towards this goal, we leverage semantic information and inverse graphics to sample a set of plausible object disparity maps given an initial semi-dense disparity estimate. We encourage the presence of these 2.5D shape samples (or "displets") in our MRF formulation depending on how much their geometry and semantic class agrees with the image observations. Our experiments indicate that the proposed framework is able to resolve stereo ambiguities on challenging stereo pairs from the KITTI benchmark as illustrated in Fig. 1 (center). At the same time our method is able to extract 3D object representations which are consistent with the estimated disparity map and may serve as input to higher-level reasoning, see Fig. 1 (bottom) for an illustration. In combination with recently proposed deep features [68], our model ranks first in all KITTI stereo evaluation tables. We make our code and object annotations available on our project website[1].

2. Related Work

As one of the oldest and most fundamental problems in computer vision [22], computational stereo has witnessed great progress making it nowadays an attractive alternative to costly laser scanners for outdoor environment perception, e.g., in autonomous driving [11, 12]. The recent success of stereo methods can be attributed to the development of benchmarks such as Middlebury [50]. More recently, the KITTI benchmark [13] has pushed limits even further by providing a larger and more realistic stereo dataset with meaningful object classes and ground truth annotations.

Local stereo methods [14, 20, 23, 27, 37, 46, 50, 67] often fail in challenging scenarios as they suffer from matching ambiguities in weakly textured, saturated or reflective regions. Thus, recent efforts mainly focus on global methods [2, 16, 41, 42, 54, 60] which are able to overcome some of these problems by imposing smoothness constraints between adjacent pixels or superpixels. On the *pixel-level*, approaches based on semi-global matching (SGM) [17, 18, 52, 53] enjoy great popularity due to their computational efficiency, accuracy and simplicity. As first-order approaches such as SGM are not able to express planarity priors, second-order models have recently been investigated either by introducing triple cliques into the formulation [63], by increasing the parameter space [34,43] or by following a total generalized variational (TGV) approach [30,45]. In contrast, *superpixel-based* methods [51,64–66] model each entity as a slanted plane, thus enforcing planarity implicitely and allowing for larger ranges of interaction, depending on the size of the superpixels. In this paper, we follow this second line of work, but go beyond pairwise interactions by modeling constraints connecting up to several hundred superpixels.

Planes, B-splines and quadratic surfaces have been proposed in [3–5, 24, 69] to better constrain the underlying problems and simplify inference. Given a set of geometric proposals, the stereo matching problem can be cast as a discrete labeling problem where each pixel is assigned to a proposal. This allows the application of standard tools for discrete optimization such as belief propagation or graph cuts. While promising, such regularizes ignore the semantic context which heavily constrains the shape of the geometric primitives. In contrast, the method proposed in this paper explicitely conditions on the semantic class, thus allowing for richer geometries and interactions spanning a larger region in the image.

While the mutual benefits of recognition and reconstruction have been shown using simple priors and multiple views [15,29], little work has addressed the binocular stereo problem in this context with notable exceptions [33,48,61]. Saxena et al. [48] proposed to directly integrate depth-from-appearance constraints into the data term. In contrast, Ladicky et al. [33] model stereo estimation and semantic segmentation jointly by learning the dependency between height over ground and the semantic class. Wei et al. [61] follow a data-driven approach which directly transfers disparity information from regions with similar appearance in the training data using SIFT flow [36]. Unfortunately, the nature of interaction in these models is very local and thus cannot constrain large ambiguous regions well enough. We thus propose to leverage object knowledge, akin to [1, 7] where geometric priors are used to improve multi-view reconstruction for single objects. In contrast to [1,7], the number of objects is unknown in our case. We jointly infer the disparity map, the number of objects and their geometry.

The proposed method also borrows ideas from binary segmentation approaches leveraging pattern-based potentials [26,39,47] to encourage plausible label configurations. While our displet masks can be interpreted as pattern potentials, we differ in that we optimize a continuous label space and model interactions far beyond the typical 10×10 pixel patches used in segmentation approaches.

3. Stereo Matching using Displets

In this paper, we tackle the classical task of binocular stereo matching. That is, given the left and right images of a synchronized, calibrated and rectified stereo camera, we are interested in estimating the disparity at each pixel of the reference image (e.g., left image). We assume that the image can be decomposed into a set of planar superpixels which we obtain using the StereoSLIC algorithm [65]. In addition to unary and pairwise constraints, we also introduce long-range interactions into our model using *dis-*

[1] http://www.cvlibs.net/projects/displets/

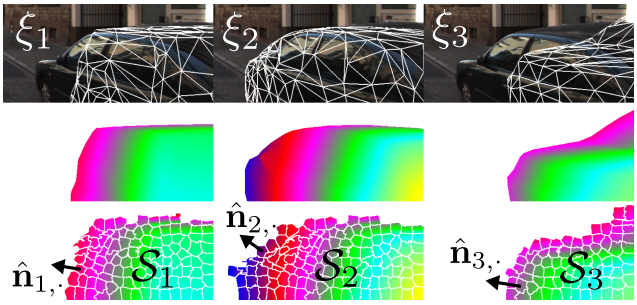


Figure 2: **Illustration of Displets:** We sample 3D CAD model configurations ξ_k (top+center) and extract the plane parameters $\hat{\mathcal{N}}_k = (\hat{\mathbf{n}}_{k,1}, \ldots, \hat{\mathbf{n}}_{k,|\mathcal{S}_k|})^T$ of the corresponding displet $k \in \mathcal{D}$ by fitting planes to the rendered disparity map for all involved superpixels $\mathcal{S}_k$ (bottom).

plets, a set of physically plausible disparity maps of a certain semantic class, associated with an image mask and a score (see Section 4 for further details). Intuitively, displets can be thought of as a representative finite subset of the infinitely large set of disparity maps for that class conditioned on the image. For example, car displets should cover the most likely 3D car shapes given the two input images. In this section, we show how displets can be incorporated as soft constraint into a CRF model. This allows us to jointly optimize for the displets and the disparity map in a principled manner.

More formally, let $\mathcal{S}$ and $\mathcal{D}$ denote the set of superpixels and displets in the reference image. Each superpixel $i \in \mathcal{S}$ is associated with a region $\mathcal{R}_i$ in the image and a random variable $\mathbf{n}_i \in \mathbb{R}^3$ describing a plane in 3D ($\mathbf{n}_i^T\mathbf{x} = 1$ for $\mathbf{x} \in \mathbb{R}^3$ on the plane). Each displet $k \in \mathcal{D}$ is associated with its class label $c_k \in \mathcal{L}\backslash\{\text{background}\}$, a fitness value $\kappa_k \in \mathbb{R}$, a set of superpixels $\mathcal{S}_k \subseteq \mathcal{S}$ on which it is defined and the corresponding plane parameters $\hat{\mathcal{N}}_k = (\hat{\mathbf{n}}_{k,1}, \ldots, \hat{\mathbf{n}}_{k,|\mathcal{S}_k|})^T$. The plane parameters are obtained by local plane fitting to the rendered disparity map of the corresponding CAD model (Fig. 2). An additional random variable $d_k \in \{0, 1\}$, which can be interpreted as auxiliary variable in a high-order CRF, denotes the presence ($d_k = 1$) or absence ($d_k = 0$) of the displet in the scene. Furthermore, we assume that we have access to a rough semantic segmentation of the image $\mathbf{S} \in \mathcal{L}^{W \times H}$ with $|\mathcal{L}|$ the number of semantic labels[2] and $W \times H$ the image dimensions. We obtain this segmentation using ALE [31] and refer the reader to Section 5 for further details.

Given the left and right image, our goal is to infer all superpixel plane parameters $\mathbf{n}_i$ as well as the presence or absence of all displets d_k in the scene. We specify our CRF in terms of the following energy function

$$E(\mathbf{n}, \mathbf{d}) = \sum_{i \in \mathcal{S}} \varphi_i^{\mathcal{S}}(\mathbf{n}_i) + \sum_{i \sim j} \psi_{ij}^{\mathcal{S}}(\mathbf{n}_i, \mathbf{n}_j) + \sum_{k \in \mathcal{D}} \varphi_k^{\mathcal{D}}(d_k) + \sum_{k \in \mathcal{D}} \sum_{i \in \mathcal{S}_k} \psi_{ki}^{\mathcal{D}}(d_k, \mathbf{n}_i) \quad (1)$$

where $\mathbf{n} = \{\mathbf{n}_i | i \in \mathcal{S}\}$ and $\mathbf{d} = \{d_k | k \in \mathcal{D}\}$ and $i \sim j$ denotes the set of adjacent superpixels in $\mathcal{S}$.

3.1. Data Term

The data term models the fact that corresponding points in the left and right image should be similar in appearance. While many options are possible, we simply penalize deviations from an initial sparse disparity map $\hat{\Omega}$, calculated with a semi-dense feature matching algorithm:

$$\varphi_i^{\mathcal{S}}(\mathbf{n}_i) = \sum_{\mathbf{p} \in \mathcal{R}_i \cap \hat{\Omega}_+} \rho_{\tau_1}(\omega(\mathbf{n}_i, \mathbf{p}) - \hat{\omega}(\mathbf{p})) \quad (2)$$

Here, $\hat{\Omega}_+$ denotes the set of valid pixels in $\hat{\Omega}$, $\omega(\mathbf{n}_i, \mathbf{p})$ is the disparity of plane $\mathbf{n}_i$ at pixel $\mathbf{p}$, and $\hat{\omega}(\mathbf{p})$ represents the value of the reference disparity map $\hat{\Omega}$ at pixel $\mathbf{p}$. Note that given the calibration parameters, the function $\omega(\mathbf{n}_i, \mathbf{p})$ is straightforward to derive and we specify all necessary details in the supplementary material. To account for outliers, we chose $\rho_\tau(\cdot)$ as the robust l_1 penalty function $\rho_\tau(x) = \min(x, \tau)$. In Section 5 we further evaluate and compare two state-of-the-art feature matching algorithms which yield the initial sparse disparity map $\hat{\Omega}$.

3.2. Local Smoothness

We encourage local smoothness in our formulation by penalizing discontinuities at superpixel boundaries as well as by encouraging similar orientations of adjacent superpixels. In particular, our smoothness term decomposes as

$$\psi_{ij}^{\mathcal{S}}(\mathbf{n}_i, \mathbf{n}_j) = \theta_1 \sum_{\mathbf{p} \in \mathcal{B}_{ij}} \rho_{\tau_2}(\omega(\mathbf{n}_i, \mathbf{p}) - \omega(\mathbf{n}_j, \mathbf{p})) + \theta_2 \; \rho_{\tau_3}\left(1 - |\mathbf{n}_i^T \mathbf{n}_j| / (\|\mathbf{n}_i\| \|\mathbf{n}_j\|)\right) \quad (3)$$

where $\mathcal{B}_{ij}$ denotes the set of shared boundary pixels between superpixel i and superpixel j and the other functions are defined as above. The weights θ_1, θ_2 control the importance of each term with respect to the other terms in Eq. 1. Inspired by contrast-sensitive smoothness priors, we downweight θ_1 and θ_2 if neighboring superpixels i and j are likely to be separated by an occlusion boundary. This likelihood is computed by simply detecting large changes in the gradient of the input disparity map $\hat{\Omega}$.

3.3. Displet Potentials

In order to encode long-range interactions, we introduce displet potentials which encourage plausible geometries in regions corresponding to a certain semantic class.

[2] While keeping our exposition general, we only consider "car" vs. "background" in our experiments as cars are the most challenging object category in KITTI while still sufficiently restricted in terms of geometry.

The unary potential for displet d_k is defined as

$$\varphi_k^{\mathcal{D}}(d_k) = -\theta_3 \left[d_k = 1\right] \cdot \left(\left|\left[\mathbf{S} = c_k\right] \cap \mathbf{M}_k\right| + \kappa_k\right) \quad (4)$$

where $[\cdot]$ denotes the (element-wise) Iverson bracket, $\mathbf{M}_k$ represents the pixel mask corresponding to the set of superpixels $\mathcal{S}_k$, and κ_k is a fitness score assigned to displet k (see Section 4 for further details). Intuitively, this potential tries to explain as many regions in the image which have been assigned the semantic class label c_k using displets whose shape corresponds to typical objects in class c_k.

Furthermore, we define a potential between each displet and all superpixels it comprises as follows

$$\psi_{ki}^{\mathcal{D}}(d_k, \mathbf{n}_i) = \lambda_{ki} \left[d_k = 1\right] \cdot \left(1 - \delta(\mathbf{n}_i, \hat{\mathbf{n}}_{k,z_i})\right) \quad (5)$$

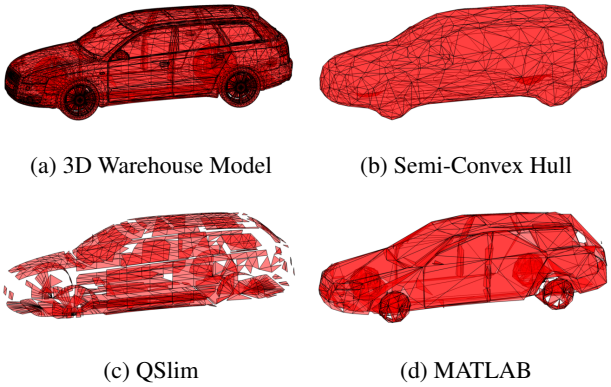
where z_i denotes the index of the plane corresponding to superpixel i in $\mathcal{N}_k$ and $\delta(\cdot,\cdot) = 1$ if both arguments are equivalent and 0 otherwise. Further, λ_{ki} is a penalty value which takes $\lambda_{ki} = \infty$ inside the object (hard constraint) and $\lambda_{ki} < \infty$ at the object boundaries (soft constraint) to better account for inaccuracies in the displet mask $\mathbf{M}_k$. While many choices are possible, we define λ_{ki} as a sigmoid function of the distance transform of $\mathbf{M}_k$ and estimate its parameters from training data. It is important to note that the hard constraint avoids evidence undercounting, i.e., it ensures that displets don't overlap and explain the same region in the image.

3.4. Inference

Minimizing Eq. 1 is a non-convex mixed continuous-discrete optimization problem which is NP-hard to solve. We leverage greedy max-product particle belief propagation (MP-PBP) [44,55] with sequential tree-reweighted message passing (TRW-S) [25] using 30 particles and 50 iterations to find an approximate solution. At every iteration, plane particles are sampled from a normal distribution around the previous MAP solution and using the plane parameters of spatially neighboring superpixels. Both strategies complement each other and we found their combination important for efficiently exploring the search space. To ensure that displets are selected with non-zero probability, we augment the proposal set for a superpixel by the plane parameters of all overlapping displets. We initialize all superpixel planes using the StereoSLIC algorithm [65].

4. Rapid Inverse Graphics

This section describes how we subsample the infinitely large space of disparity maps using inverse graphics, yielding the set of displets $\mathcal{D}$ used in the previous section. We make use of MCMC to draw a set of representative samples corresponding to a certain object category (e.g., cars). In contrast to [38], our generative process produces disparity maps from CAD models using the camera intrinsics. Our likelihood model compares the rendered disparity map with the input disparity map $\hat{\Omega}$ and returns a score depending on the level of agreement. This makes our algorithm invariant to the actual image intensities which are hard to model in a generative way, in particular in the presence of reflecting or translucent surfaces.



(a) 3D Warehouse Model (b) Semi-Convex Hull

(c) QSlim (d) MATLAB

Figure 3: **Mesh Simplification.** For efficient rendering, we simplify 3D CAD Models with ∼100k faces (a) by a smooth semi-convex approximation using 1k faces only (b). The application of generic mesh simplification algorithms using the same number of faces produces undesirable holes and self-intersections in the mesh as illustrated in (c+d).

4.1. Semi-Convex Hull

We start with a set of representative CAD models from Google 3D Warehouse[3] which capture most of the 3D shape variability of the object category. Unfortunately, CAD models downloaded from Google Warehouse are not directly applicable as they are often designed with love of detail resulting in hundreds of thousands of vertices and faces, slowing down the rendering process significantly. We found that tools like MATLAB's `reducepatch` function[4] or QSlim [10] are not able to simplify these models to an affordable level of detail without introducing holes or artifacts as illustrated in Fig. 3. In this section, we thus propose a simple method which reduces a CAD model of geometrically simple classes such as cars to around 1000 faces while preserving the hull of the object and removing all interior elements which are not affecting the rendered depth map.

We initialize a mesh using the convex hull of the object and gradually relax it to a smooth approximation, subject to the constraint that the volume of the model comprises all surface points. We call this representation the "semi-convex hull" of an object. In particular, we minimize the squared point-to-point distances between all vertices of the mesh and densely sampled points on the original 3D model.

[3] https://3dwarehouse.sketchup.com/

[4] http://www.mathworks.de/help/matlab/ref/reducepatch.html

Algorithm 1: Mesh Simplification

Input: 3D CAD model
Output: Semi-convex hull $(\mathcal{M}, \mathbf{x})$

$\mathcal{P} \leftarrow$ draw samples from 3D CAD model
$(\mathcal{M}, \mathbf{x}) \leftarrow$ convex hull of 3D CAD model
$(\mathcal{M}, \mathbf{x}) \leftarrow$ remeshing of $(\mathcal{M}, \mathbf{x})$ using [9]
while *not converged* **do**
  $\mathbf{x} \leftarrow \mathbf{x} - \gamma \nabla E(\mathbf{x})$
  if $\mathcal{P} \nsubseteq \mathit{Vol}(\mathcal{M}, \mathbf{x})$ **then**
    $\alpha \leftarrow \min(\{\alpha > 0 \,|\, \mathcal{P} \subseteq \mathrm{Vol}(\mathcal{M}, \alpha\,\mathbf{x})\})$
    $\mathbf{x} \leftarrow (\alpha + \epsilon)\,\mathbf{x}$
    $(\mathcal{M}, \mathbf{x}) \leftarrow$ remeshing of $(\mathcal{M}, \mathbf{x})$ using [9]
$(\mathcal{M}, \mathbf{x}) \leftarrow$ simplification of $(\mathcal{M}, \mathbf{x})$ using [10]

More formally, let $\mathcal{P}$ denote the set of 3D points obtained by uniformly sampling a large number of 3D points from the union of all surfaces of the object (full resolution CAD model). Let further $\mathcal{M} = \{\mathcal{V}, \mathcal{F}\}$ denote a mesh with vertices $\mathcal{V}$, faces $\mathcal{F}$ and edges $\mathcal{E}(\mathcal{F})$. Each vertex $i \in \mathcal{V}$ is associated with a variable $\mathbf{x}_i \in \mathbb{R}^3$ specifying the location of the vertex. We initialize $\mathcal{M}$ and $\mathbf{x} = \{\mathbf{x}_i | i \in \mathcal{V}\}$ by uniformly remeshing the object's convex hull using isotropic surface remeshing [9] and formulate our objective as minimizing

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \|\mathbf{x}_i - nn(\mathbf{x}_i)\|^2 + \sum_{(i,j) \in \mathcal{E}} \left(\|\mathbf{x}_i - \mathbf{x}_j\|^2 - \bar{l}_0 \right)^2$$
$$\text{s.t.} \quad \mathcal{P} \subseteq \mathrm{Vol}(\mathcal{M}, \mathbf{x}) \tag{6}$$

where $nn(\mathbf{x}_i)$ denotes the nearest neighbor of $\mathbf{x}_i$ in $\mathcal{P}$, $\bar{l}_0$ is the average edge length of the initial mesh, and $\mathrm{Vol}(\mathcal{M}, \mathbf{x})$ denotes the set of 3D points inside the mesh. We solve Eq. 6 using gradient descent and enforce the closure constraint $\mathcal{P} \subseteq \mathrm{Vol}(\mathcal{M}, \mathbf{x})$ by mesh rescaling and uniform remeshing on violation. After convergence, we obtain a smooth semi-convex hull of the object which we simplify to 1k faces using QSlim [10]. See Algorithm 1 for further details.

4.2. Sampling the Space of Displets

For a given object category c, we are interested in subsampling the space of plausible displets given a semi-dense disparity image $\hat{\Omega}$, a semantic segmentation $\mathbf{S}$, and the semi-convex hull of all CAD models of this object category. We approach this inverse graphics problem using MCMC, i.e., we sample pose parameters $\xi \in SE(3)$ directly from the observation model $p(\xi|\hat{\Omega}) \propto \exp(-E_{\hat{\Omega}}(\xi))$ with

$$E_{\hat{\Omega}}(\xi) = \sum_{\mathbf{p} \in \hat{\Omega}_+ \cap \mathcal{O}} \frac{\min\left(|\bar{\omega}(\mathbf{p}, \xi) - \hat{\omega}(\mathbf{p})|, \tau_1\right)}{|\hat{\Omega}_+ \cap \mathcal{O}|} + \beta \sum_{\mathbf{p} \in \hat{\Omega}_+} [\bar{\omega}(\mathbf{p}, \xi) > \hat{\omega}(\mathbf{p}) + \tau_2] \tag{7}$$

Here, $\mathcal{O}$ denotes a 2D object instance in the image, $\bar{\omega}(\mathbf{p}, \xi)$ is the disparity of the CAD model in pose ξ rendered at pixel $\mathbf{p}$, and $\beta, \tau_1, \tau_2 > 0$ are parameters of the model. Intuitively, $E_{\hat{\Omega}}(\xi)$ encourages displets to explain all pixels within object region $\mathcal{O}$ in terms of disparity (first term) while avoiding the occlusion of other objects (second term). In principle, the use of object proposals $\mathcal{O}$ could be avoided by directly sampling according to the semantic labeling $\mathbf{S}$, but we found instance level information to improve the diversity of the displet set. While a large number of generic object proposal algorithms [6, 28, 56, 70] can be applied to obtain the set of object proposal regions $\{\mathcal{O}\}$, we found a much more simple strategy to be sufficient for our goals: First, we project all valid pixels of class c ($\mathbf{p} \in \hat{\Omega}_+ \cap [\mathbf{S} = c]$) into 3D. Next, we apply kernel density estimation (KDE) along the principal coordinate axes x and z of the camera. As object boundaries frequently coincide with minima of the KDE, we propose one object region $\mathcal{O}$ for each pair of adjacent minima by projecting all 3D points in this range back into the image. It is important to note that we do not assume precise object boundaries for the proposals due to the robust term in Eq. 7.

We run one Markov chain for each combination of CAD models and object proposals using Metropolis-within-Gibbs (MWG) sampling (5.000 iterations) with randomly chosen blocks. For each combination, we select the 8 most dominant modes after burn-in and combine all results to yield the final set of displets. An illustration is given in the supplementary material. As our semi-convex mesh has a low number of vertices and faces, we are able to draw a large number of samples on commodity graphics hardware. In practice, we achieve $\sim$ 8200 fps on a single NVIDIA Quadro 4000 GPU using 12 threads.

5. Experimental Results

This section provides a thorough quantitative and qualitative analysis of the proposed displet model. As the number of images and objects per category in Middlebury [50] is too small to allow for a meaningful evaluation, we chose the more recent and challenging KITTI stereo dataset [13] as testbed for our experiments. Following the KITTI evaluation protocol, we perform all ablation studies on the training set from which we randomly select 50 images for training the parameters in our model. The remaining images are used for validation. In addition we submit our best performing configuration to the KITTI server for evaluation on the test set. We perform block coordinate descent on the 50 training images to obtain the model parameters $\{\theta\}$ and $\{\tau\}$ which we fix throughout all our experiments. For further details we refer the reader to the supplementary material.

Image Features: We manually annotated the training and test sets with pixel-wise car versus background labels and trained the associative hierarchical random fields model [31] for semantic segmentation. In order to obtain sparse

but high-quality input disparity maps we process all stereo pairs using the semi-global matching framework [18] followed by a simple left-right consistency check to remove outliers. For calculating the matching costs, we use a combination of Census and Sobel features ("SGM", [64]), as well as more recently proposed features based on convolutional neural networks ("CNN", [68]).

Ablation Study: Our first set of experiments conducted on the KITTI training set aims at assessing the contribution of each individual term in our energy. As we expect most notable improvements at reflective surfaces, we evaluate the error in all image regions (Table 1b) as well as the error only in reflective regions (Table 1a) using the KITTI ground truth. Unless otherwise stated, we report the percentage of outliers using the default outlier threshold of 3 pixels.

The first row in 1 shows the results of the input disparity maps, interpolated using the KITTI development kit. The following rows show our results when using only the unary term or combining it with one of the pairwise smoothness terms for superpixel boundaries ("Pair (Boundary)"), orientations ("Pair (Orientation)") and both ("Pair"), respectively. In combination, both smoothness terms are able to reduce the error by 16.6% for reflective and by 5.7% for all image regions considering the better performing CNN features. Adding the occlusion sensitive weight further reduces the error in all image regions but makes results slightly worse at reflective surfaces. This can be attributed to the fact that the input disparity maps contain holes and errors at reflective regions which are sometimes erroneously identified as occlusion boundaries hence lowering the smoothing effect for these regions. Finally, adding the proposed displets to the model dramatically improves the results, reducing the number of outliers by additional 53.3% in reflective regions and by 8.6% in all regions.

Next, we evaluate the influence of the number of object proposals as well as the variety of CAD models used for generating the displets. For these experiments, we focus our attention on the reflective regions as those are most affected by limiting the number of displets. As we obtain a different number of displet proposals in each image, we randomly draw subsets and plot the error with respect to the acceptance probability of a displet in Fig. 4 (left). Here, $P = 0$ corresponds to the case without displets and $P = 1$ corresponds to the case when making use of all available proposals for inference. The performance with respect to the number of CAD models used is shown in Fig. 4 (right). For this experiment, we randomly select a subset of models for each image (ranging from 0 to all 8 models) and discard all the proposals from all other models. Both plots illustrate that reasonable results can be obtained with a small number of displet proposals and 3 models only. However, in both cases performance keeps increasing when adding more displets.

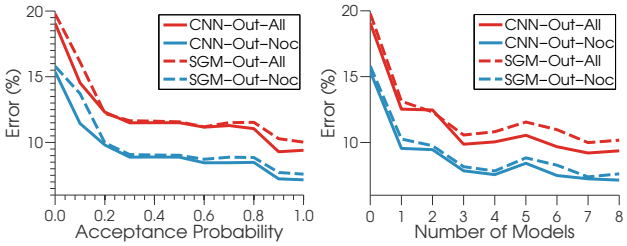


Figure 4: **Number of Proposals and Models.** This figure shows the performance in reflective regions when limiting the number of object proposals (left) and models (right).

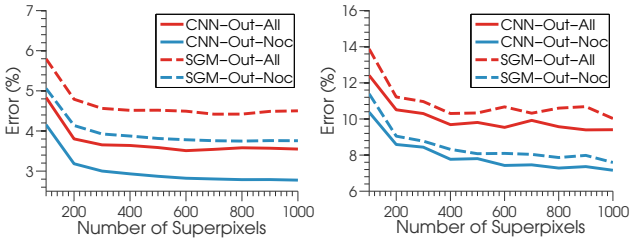


Figure 5: **Number of Superpixels.** These figures show the impact when varying the number of superpixels for all regions (left) and for reflective regions (right).

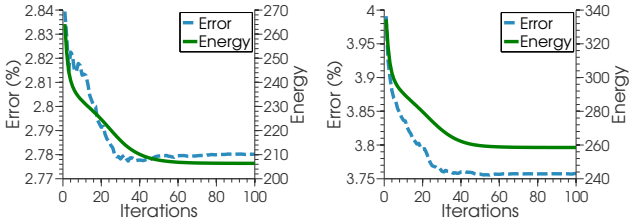


Figure 6: **Convergence of Error and Energy.** This figure shows the decrease in error and energy on all image regions vs. MP-PBP iterations using CNN (left) and SGM (right).

In Fig. 5, we investigate the impact of the number of superpixels on the performance of our model. Similarly to [64], we observe diminishing returns beyond 500 superpixels and chose 1000 superpixels as a reasonable trade-off between accuracy and performance for all other experiments.

As our inference procedure is an iterative process with runtime linear in the number of iterations we plot the energy and errors through the iterations to determine the point of convergence. Fig. 6 shows the average energy and error over the images in validation set using CNN (left) and SGM (right) as input disparity maps. As both the error and the energy stabilize after about 40 iterations, we set the maximum number of iterations to this value throughout all our experiments.

Results on the KITTI Stereo Benchmark: This section compares our performance with respect to the current state-

	CNN		SGM	
	Out-Noc	Out-All	Out-Noc	Out-All
Input $\hat{\Omega}$ (Interpolated)	19.84 %	22.98 %	22.60 %	25.52 %
Unary Only	17.72 %	21.72 %	19.38 %	23.36 %
Unary + Pair (Boundary)	15.96 %	19.67 %	16.18 %	19.94 %
Unary + Pair (Normal)	17.06 %	20.80 %	18.24 %	21.86 %
Unary + Pair	14.78 %	18.77 %	14.91 %	18.85 %
Unary + Pair + Occ	15.32 %	19.32 %	15.79 %	19.60 %
Unary + Pair + Disp	7.08 %	9.30 %	7.45 %	9.98 %
Unary + Pair + Occ + Disp	7.16 %	9.41 %	7.59 %	10.02 %

(a) **Reflective Regions**

	CNN		SGM	
	Out-Noc	Out-All	Out-Noc	Out-All
Input $\hat{\Omega}$ (Interpolated)	3.35 %	4.28 %	5.13 %	6.08 %
Unary Only	3.31 %	4.60 %	4.71 %	5.96 %
Unary + Pair (Boundary)	3.21 %	4.15 %	4.28 %	5.23 %
Unary + Pair (Normal)	3.28 %	4.31 %	4.52 %	5.60 %
Unary + Pair	3.12 %	3.95 %	4.13 %	4.93 %
Unary + Pair + Occ	3.04 %	3.88 %	4.07 %	4.80 %
Unary + Pair + Disp	2.87 %	3.64 %	3.87 %	4.57 %
Unary + Pair + Occ + Disp	2.78 %	3.55 %	3.76 %	4.50 %

(b) **All Regions**

Table 1: **Importance of Different Terms in the Model.** This table shows the performance of various model configurations on the validation set of KITTI for reflective regions (a) and for all regions (b) using the default error threshold of 3 pixels.

Rank	Method	Out-Noc	Out-All	Avg-Noc	Avg-All
1	Our Method	**8.40 %**	**9.89 %**	**1.9 px**	**2.3 px**
2	VC-SF * [57]	11.58 %	12.29 %	2.7 px	2.8 px
3	PCBP-SS [65]	14.26 %	18.33 %	2.4 px	3.9 px
4	SPS-StFl * [66]	14.74 %	18.00 %	2.9 px	3.6 px
5	CoP	15.30 %	19.15 %	2.7 px	4.1 px
6	SPS-St [66]	16.05 %	19.34 %	3.1 px	3.6 px
7	DDS-SS [61]	16.23 %	19.39 %	2.5 px	3.0 px
8	PCBP [64]	16.28 %	20.22 %	2.8 px	4.4 px
9	PR-Sf+E * [58]	17.85 %	20.82 %	3.3 px	4.0 px
10	StereoSLIC [65]	18.22 %	21.60 %	2.8 px	3.6 px
11	MC-CNN [68]	18.45 %	21.96 %	3.5 px	4.3 px
12	PR-Sceneflow * [58]	19.22 %	22.07 %	3.3 px	4.0 px
⋮	⋮	⋮	⋮	⋮	
62	ALE-Stereo [33]	83.80 %	84.37 %	24.6 px	25.4 px

(a) **Reflective Regions**

Rank	Method	Out-Noc	Out-All	Avg-Noc	Avg-All
1	Our Method	**2.47 %**	**3.27 %**	**0.7 px**	**0.9 px**
2	MC-CNN [68]	2.61 %	3.84 %	0.8 px	1.0 px
3	SPS-StFl * [66]	2.83 %	3.64 %	0.8 px	**0.9 px**
4	VC-SF * [57]	3.05 %	3.31 %	0.8 px	0.8 px
5	SPS-St [66]	3.39 %	4.41 %	0.9 px	1.0 px
6	PCBP-SS [65]	3.40 %	4.72 %	0.8 px	1.0 px
7	CoP	3.78 %	4.63 %	0.9 px	1.1 px
8	DDS-SS [61]	3.83 %	4.59 %	0.9 px	1.0 px
9	StereoSLIC [65]	3.92 %	5.11 %	0.9 px	1.0 px
10	PR-Sf+E * [58]	4.02 %	4.87 %	0.9 px	1.0 px
11	PCBP [64]	4.04 %	5.37 %	0.9 px	1.1 px
12	PR-Sceneflow * [58]	4.36 %	5.22 %	0.9 px	1.1 px
⋮	⋮	⋮	⋮	⋮	⋮
62	ALE-Stereo [33]	50.48 %	51.19 %	13.0 px	13.5 px

(b) **All Regions**

Table 2: **Quantitative Evaluation on the KITTI Stereo Benchmark.** This table shows the KITTI stereo leaderboards at time of submission using the default error threshold of 3 pixels. Evaluation is performed separately for reflective regions (a) and for all regions (b) of the KITTI test set. The numbers represent outliers (in %) and average disparity error (in pixels). Methods marked with an asterisk are scene flow methods which use two or more stereo image pairs as input.

of-the-art. We submitted our results for the KITTI test set using the best performing configuration according to Table 1b (CNN+Full model) to the KITTI evaluation server. As shown in Table 2, our method ranks first amongst more than 60 competitors in all evaluation categories. As expected, our improvements are particularly pronounced in reflective regions, but also improve overall performance even with respect to scene flow methods which take two or more stereo pairs of the sequence as input. The relatively weak performance of ALE-Stereo [33] can be attributed to the simple semantic interaction model as well as the suboptimal graph-cuts based inference procedure.

Qualitative Results: Fig. 7 shows some qualitative results of our method without displets (left column in each subfigure) as well as our full model including displets (right column in each subfigure). As evidenced by the error images in the last row of each subfigure, the proposed displet significantly reduces errors for the category car in a large number of images and even in extremely challenging scenarios such as the subfigure in row 2, column 2. Two failure cases are highlighted at the bottom: In the left scene, errors on the caravan increase slightly as our collection of 3D CAD models doesn't contain an instance of this rather rare vehicle type. In the right scene the trunc of the car is extrapolated towards the building due to failures in the semantic segmentation (indicated in green) while the overall reconstruction of the trunk has improved.

Runtime: Our non-optimized MATLAB implementation with C++ wrappers requires on average 60 seconds for sampling the displets (parallelized using 12 cores), 5 seconds for initialization and 2.5 seconds for each of the 40 MP-PBP iterations, including graph construction (2.2 seconds)

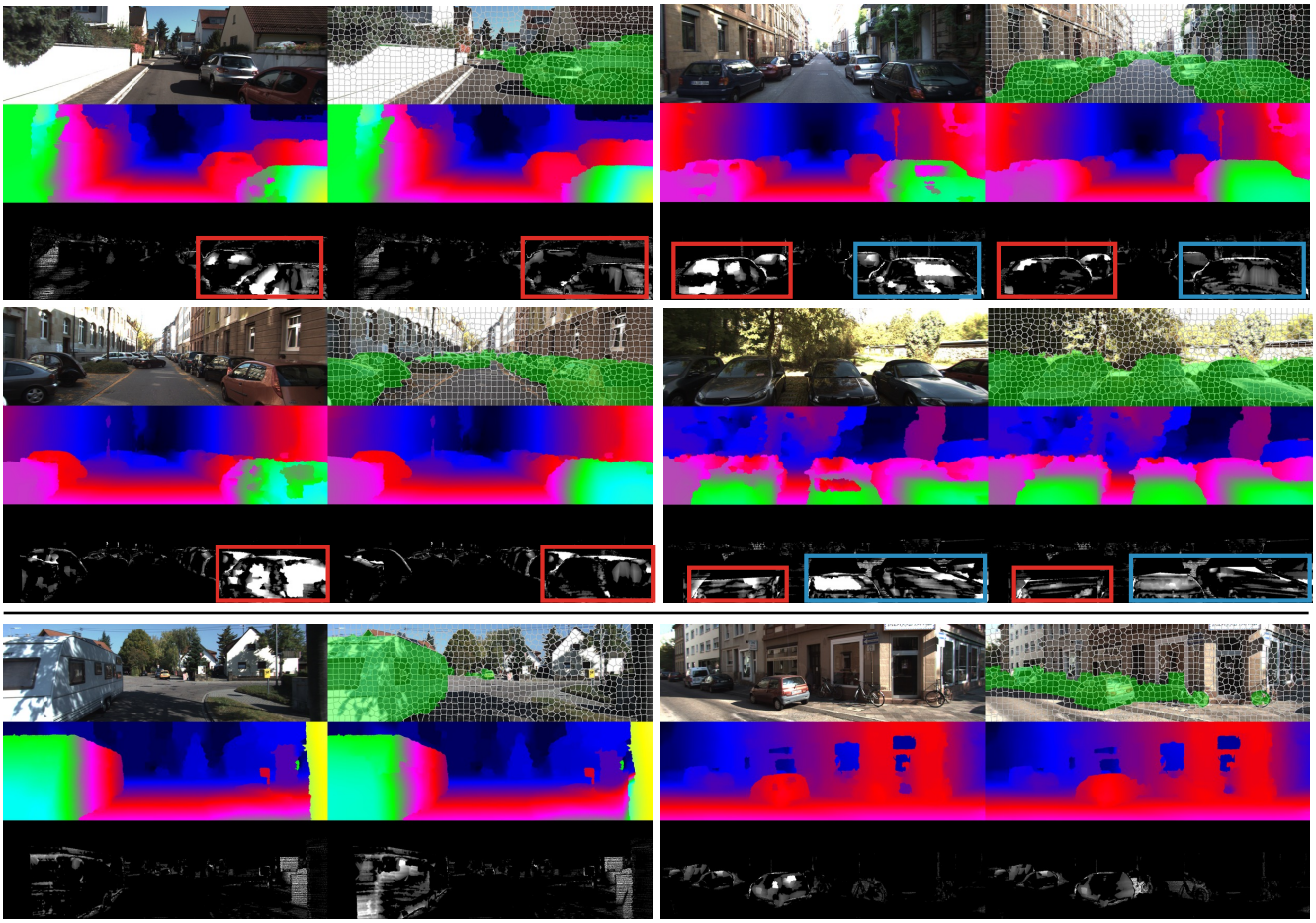


Figure 7: **Qualitiative Results.** Each subfigure shows from top-to-bottom: The input image, the superpixels and the semantic segments which our method takes as input (first row), our inference results without and with displets (second row), and the corresponding error maps from $\leq$ 1 pixel error in black to $\geq$ 5 pixels error in white (third row). We mark regions with big improvements using red and blue rectangles. Two failure cases where the displets do not fit the data in terms of depth (bottom-left) and shape (bottom-right) are illustrated below the horizontal line.

and TRW-S inference (0.3 seconds). In addition, ALE [31] requires 3.5 second per binary image segmentation and we obtain our initial disparity maps in 5 seconds for SGM or 100 seconds for CNN. We thus require 265 seconds in total for processing a single image using the full model in combination with CNN based matching costs.

Supplementary Material: We encourage the reader to look at our project website as well as the supplementary material[5] where we provide an analysis of performance with respect to variation of parameters, number of particles as well as additional qualitative results on the KITTI stereo validation set. The project page also provides two videos comparing our 3D reconstruction performance to that of Zbontar et al. [68] and demonstrating the optimization of the semi-convex hull.

[5] http://www.cvlibs.net/projects/displets/

6. Conclusion and Future Work

We propose displets as expressive non-local prior for resolving ambiguities in stereo matching. By conditioning on the semantic class of an object, displets provide valuable category specific shape information which allows for regularizing the solution over very large distances in the image. Our displet model for cars attains state-of-the-art performance on the challenging KITTI stereo benchmark, lowering errors in reflective and textureless regions by 50%. While in this paper we deliberately focused on this particular object category, we plan to investigate the applicability of displets to other geometrically well constrained object classes in the future. Buildings, for instance, often lack texture but their shape can be well described by a set of planes or 3D box primitives. Another interesting future direction is the extension of displets to flowlets for serving as non-local category specific prior in optical flow and scene flow.

References

[1] S. Bao, M. Chandraker, Y. Lin, and S. Savarese. Dense object reconstruction with semantic priors. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[2] F. Besse, C. Rother, A. Fitzgibbon, and J. Kautz. PMBP: PatchMatch Belief Propagation for correspondence field estimation. *International Journal of Computer Vision (IJCV)*, 110(1):2–13, 2014. 2

[3] M. Bleyer, C. Rhemann, and C. Rother. Extracting 3D scene-consistent object proposals and depth from stereo images. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2012. 2

[4] M. Bleyer, C. Rother, and P. Kohli. Surface stereo with soft segmentation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010. 2

[5] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, and S. Sinha. Object stereo - joint stereo matching and object segmentation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011. 2

[6] J. Carreira and C. Sminchisescu. CPMC: automatic object segmentation using constrained parametric min-cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 34(7):1312–1328, 2012. 5

[7] A. Dame, V. Prisacariu, C. Ren, and I. Reid. Dense reconstruction using 3D object shape priors. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[8] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. 1

[9] S. Fuhrmann, J. Ackermann, T. Kalbe, and M. Goesele. Direct resampling for isotropic surface remeshing. In *Vision, Modeling and Visualization (VMV)*, 2010. 5

[10] M. Garland and P. S. Heckbert. Surface simplification using quadric error metrics. In *ACM Trans. on Graphics (SIGGRAPH)*, pages 209–216, 1997. 4, 5

[11] S. K. Gehrig, F. Eberli, and T. Meyer. A real-time low-power stereo vision engine using semi-global matching. In *Proc. of the International Conf. on Computer Vision Systems (ICVS)*, 2009. 2

[12] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun. 3D traffic scene understanding from movable platforms. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 36(5):1012–1025, 2014. 2

[13] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1, 2, 5

[14] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In *Proc. of the Asian Conf. on Computer Vision (ACCV)*, 2010. 2

[15] C. Haene, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3D scene reconstruction and class segmentation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[16] P. Heise, S. Klose, B. Jensen, and A. Knoll. PM-Huber: patchmatch with huber regularization for stereo matching. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, December 2013. 2

[17] S. Hermann and R. Klette. Iterative semi-global matching for robust driver assistance systems. In *Proc. of the Asian Conf. on Computer Vision (ACCV)*, 2013. 2

[18] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 30(2):328–341, 2008. 2, 6

[19] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *International Journal of Computer Vision (IJCV)*, 75(1):151–172, October 2007. 1

[20] A. Hosni, M. Bleyer, M. Gelautz, and C. Rhemann. Local stereo matching using geodesic support weights. In *Proc. IEEE International Conf. on Image Processing (ICIP)*, 2009. 2

[21] B. Jacquet, C. Hane, K. Koser, and M. Pollefeys. Real-world normal map capture for nearly flat reflective surfaces. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, December 2013. 1

[22] B. Julesz. Binocular Depth Perception of Computer-Generated Patterns. *Bell System Technical Journal*, 39(5):1125–1161, 1960. 2

[23] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. In *Proc. IEEE International Conf. on Robotics and Automation (ICRA)*, 1994. 2

[24] A. Klaus, M. Sormann, and K. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *Proc. of the International Conf. on Pattern Recognition (ICPR)*, 2006. 2

[25] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 28(10):1568–1583, 2006. 4

[26] N. Komodakis and N. Paragios. Beyond pairwise energies: Efficient optimization for higher-order MRFs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[27] K. Konolige. Small vision system. hardware and implementation. In *Proc. of the International Symposium on Robotics Research (ISRR)*, 1997. 2

[28] P. Krähenbühl and V. Koltun. Geodesic object proposals. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 5

[29] A. Kundu, Y. Li, F. Dellaert, F. Li, and J. Rehg. Joint semantic segmentation and 3d reconstruction from monocular video. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 2

[30] G. Kuschk and D. Cremers. Fast and accurate large-scale stereo reconstruction using variational methods. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV) Workshops*, 2013. 2

[31] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical random fields. *IEEE Trans. on Pat-*

tern Analysis and Machine Intelligence (PAMI), 36(6):1056–1077, 2014. 3, 5, 8

[32] L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1

[33] L. Ladicky, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. Torr. Joint optimisation for object class segmentation and dense stereo reconstruction. In *Proc. of the British Machine Vision Conf. (BMVC)*, 2010. 2, 7

[34] G. Li and S. W. Zucker. Differential geometric inference in surface stereo. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 32(1):72–86, 2010. 2

[35] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010. 1

[36] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman. SIFT flow: Dense correspondence across different scenes. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2008. 2

[37] Z. Ma, K. He, Y. Wei, J. Sun, and E. Wu. Constant time weighted median filtering for stereo matching and beyond. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, December 2013. 2

[38] V. Mansinghka, T. Kulkarni, Y. Perov, and J. Tenenbaum. Approximate bayesian image interpretation using generative probabilistic graphics programs. *NIPS 2013*, 2013. 4

[39] P. Marquez-Neila, P. Kohli, C. Rother, and L. Baumela. Non-parametric higher-order random fields for image segmentation. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 2

[40] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Company, 1983. 1

[41] X. Mei, X. Sun, W. Dong, H. Wang, and X. Zhang. Segment-tree based cost aggregation for stereo matching. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[42] X. Mei, X. Sun, M. Zhou, S. Jiao, H. Wang, and X. Zhang. On building an accurate stereo matching system on graphics hardware. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV) Workshops*, 2011. 2

[43] C. Olsson, J. Ulen, and Y. Boykov. In defense of 3D-label stereo. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[44] J. Pacheco, S. Zuffi, M. J. Black, and E. Sudderth. Preserving modes and messages via diverse particle selection. In *Proc. of the International Conf. on Machine learning (ICML)*, 2014. 4

[45] R. Ranftl, T. Pock, and H. Bischof. Minimizing TGV-based variational models with non-convex data terms. In *Proc. of the International Conference on Scale Space and Variational Methods in Computer Vision (SSVM)*, 2013. 2

[46] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011. 2

[47] C. Rother, P. Kohli, W. Feng, and J. Jia. Minimizing sparse higher order energy functions of discrete variables. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[48] A. Saxena, J. Schulte, and A. Y. Ng. Depth estimation using monocular and stereo cues. In *Proc. of the International Joint Conf. on Artificial Intelligence (IJCAI)*, 2007. 2

[49] A. Saxena, M. Sun, and A. Y. Ng. Make3D: learning 3D scene structure from a single still image. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 31:824–840, 2009. 1

[50] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision (IJCV)*, 47:7–42, 2002. 1, 2, 5

[51] M. Schönbein and A. Geiger. Omnidirectional 3d reconstruction in augmented manhattan worlds. In *Proc. IEEE International Conf. on Intelligent Robots and Systems (IROS)*, 2014. 2

[52] S. N. Sinha, D. Scharstein, and R. Szeliski. Efficient high-resolution stereo matching using local plane sweeps. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[53] R. Spangenberg, T. Langner, and R. Rojas. Weighted semi-global matching and center-symmetric census transform for robust driver assistance. In *Proc. of the International Conf. on Computer Analysis of Images and Patterns (CAIP)*, 2013. 2

[54] T. Taniai, Y. Matsushita, and T. Naemura. Graph cut based continuous stereo matching using locally shared labels. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[55] H. Trinh and D. McAllester. Unsupervised learning of stereo vision with monocular cues. In *Proc. of the British Machine Vision Conf. (BMVC)*, 2009. 4

[56] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 104(2):154–171, 2013. 5

[57] C. Vogel, S. Roth, and K. Schindler. View-consistent 3D scene flow estimation over multiple frames. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 7

[58] C. Vogel, K. Schindler, and S. Roth. Piecewise rigid scene flow. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2013. 7

[59] H. von Helmholtz. *Handbuch der physiologischen Optik*. L. Voss, 1867. 1

[60] L. Wang, H. Jin, and R. Yang. Search space reduction for MRF stereo. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2008. 2

[61] D. Wei, C. Liu, and W. Freeman. A data-driven regularization model for stereo and flow. In *Proc. of the International Conference on 3D Vision (3DV)*, 2014. 2, 7

[62] M. Weinmann, A. Osep, R. Ruiters, and R. Klein. Multi-view normal field integration for 3D reconstruction of mirroring objects. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2013. 1

[63] O. Woodford, P. Torr, I. Reid, and A. Fitzgibbon. Global stereo reconstruction under second-order smoothness priors. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 31:2115–2128, 2009. 2

[64] K. Yamaguchi, T. Hazan, D. McAllester, and R. Urtasun. Continuous markov random fields for robust stereo estimation. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2012. 2, 6, 7

[65] K. Yamaguchi, D. McAllester, and R. Urtasun. Robust monocular epipolar flow estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2, 4, 7

[66] K. Yamaguchi, D. McAllester, and R. Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 2, 7

[67] K.-j. Yoon, S. Member, and I. S. Kweon. Adaptive support-weight approach for correspondence search. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 28:650–656, 2006. 2

[68] J. Zbontar and Y. LeCun. Computing the stereo matching cost with a convolutional neural network. *arXiv.org*, 1409.4326, 2014. 1, 2, 6, 7, 8

[69] C. Zhang, Z. Li, R. Cai, H. Chao, and Y. Rui. As-rigid-as-possible stereo under second order smoothness priors. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 2

[70] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 5