# Stereo reconstruction using top-down cues

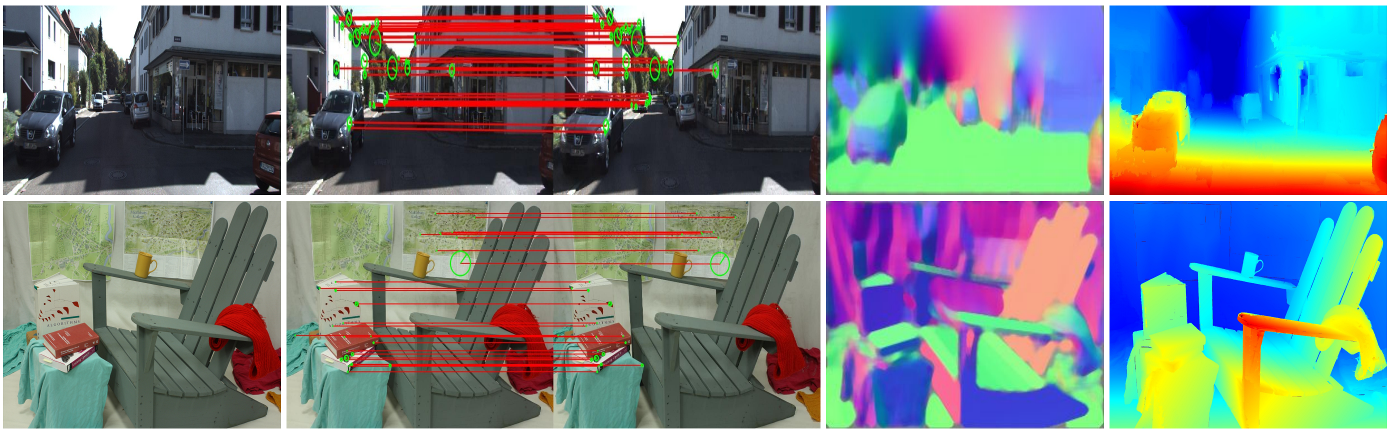
Simon Hadfield[a,*], Karel Lebeda[a], Richard Bowden[a]

[a] *CVSSP, University of Surrey, GU2 7XH Guildford, United Kingdom*

## Abstract

We present a framework which allows standard stereo reconstruction to be unified with a wide range of classic top-down cues from urban scene understanding. The resulting algorithm is analogous to the human visual system where conflicting interpretations of the scene due to ambiguous data can be resolved based on a higher level understanding of urban environments. The cues which are reformulated within the framework include: recognising common arrangements of surface normals and semantic edges (*e.g.* concave, convex and occlusion boundaries), recognising connected or coplanar structures such as walls, and recognising collinear edges (which are common on repetitive structures such as windows). Recognition of these common configurations has only recently become feasible, thanks to the emergence of large-scale reconstruction datasets. To demonstrate the importance and generality of scene understanding during stereo-reconstruction, the proposed approach is integrated with 3 different state-of-the-art techniques for bottom-up stereo reconstruction. The use of high-level cues is shown to improve performance by up to 15 % on the Middlebury 2014 and KITTI datasets. We further evaluate the technique using the recently proposed HCI stereo metrics, finding significant improvements in the quality of depth discontinuities, planar surfaces and thin structures.

*Keywords:* Stereo reconstruction, Scene understanding, biologically inspired, high level cues, bottom up, top down



(a) Input data (b) Appearance matching (c) Scene reasoning (d) Output reconstruction

Figure 1: An illustration of the different components which are unified within the proposed framework. Specifically including both bottom-up matching (b) and a top-down understanding (c) of an outdoor (top) and indoor (bottom) urban scene.

## 1. Introduction

One of the classic challenges of computer vision is estimating the 3D structure of an environment, using only visual information. It is one of the most commonly exploited techniques, with uses in mapping, robotics, surveying and many others. However, it is also one of the most challenging problems in computer vision to solve for general scenes as visual information is inherently ambiguous. The same observations may be produced by infinitely many combinations of structure, texture and illumination. However, not all these different configurations are equally probable in the real world. Particularly when the environment is highly structured, such as in an urban setting, there are strong cues which can be used to find a feasible interpretation of the scene. In this paper we demonstrate a general framework which enables top-down scene reasoning to be integrated into state-of-the-art stereo reconstruction pipelines, alongside traditional bottom-up appearance matching.

The formulation is inspired by the human visual system. At the lowest level the human vision system perceives

*Corresponding author. Telephone: +44 1483 682260.
*Email addresses:* `S.Hadfield@surrey.ac.uk` (Simon Hadfield), `K.Lebeda@surrey.ac.uk` (Karel Lebeda), `R.Bowden@surrey.ac.uk` (Richard Bowden)

*Preprint submitted to CVIU: Special Issue on 3D Modeling of Urban Indoor or Outdoor Scenes* *August 1, 2016*

depth by recognising matched elements of the observations from the eyes. These matches are then triangulated to estimate depth. This is analogous to traditional stereo reconstruction techniques from computer vision (with some additional assumptions about smoothness) dating back as far as the 1960s [1, 2]. In this paper, we refer to this approach as bottom-up reconstruction, because the reconstruction emerges from the matching of small scene sub-units. It is also interesting to note that before this bottom-up matching can be undertaken in most computer vision systems, an additional calibration task must first be solved: estimating the epipolar geometry. This is equivalent to the innate knowledge of the eyes' characteristics in the human visual system.

However, this is not the only approach employed by humans. The human visual system also exploits a wide range of very strong high-level cues to help understand the structure of the environment, particularly within more structured domains such urban environments. It is easy to recognise this, when one considers that a person has no difficulty understanding the layout of objects within the source environment for a photo or video, even though the information obtained via triangulation correctly states that the objects are all on a planar surface. The cues that the human vision system uses to achieve this have been well studied in computer vision, especially in this particular scenario when only a single image is available. This is often called *single-image scene understanding*.

As with bottom-up reconstruction, the human vision system has a significant advantage over computer vision, which is gained through experience and understanding of the world. Studies have found that these types of top-down cues take around 7 months of continuous online training to emerge in humans [3]. It is only recently, with the growth of large-scale 3D datasets that more accurate data-driven approaches to scene understanding have become feasible within computer vision [4, 5]. In previous work, some of the most commonly exploited cues involve assumptions about gravity and the viewing orientation, assumptions about the types of surfaces found in urban environments (for example the Manhattan world assumption) and assumptions about common configurations of objects and primitives. In this work we refer to these collectively as top-down reconstruction techniques, because the structure for each element of the scene is determined based on rules relating to the overall configuration of multiple sub-units.

One interesting advantage of the proposed unified framework is that the use of top-down cues reduces issues relating to the baseline which are common in traditional bottom-up reconstruction. The accuracy of these matching and triangulation based systems is strongly limited by the *baseline* (*i.e.* the separation of the two cameras). When the range of the scene is significantly larger than this baseline, even small errors in the initial matching and triangulation are manifested as much larger errors in the estimated depth. In contrast, our unified system is capable of smoothly transitioning from top-down reconstruction when the bottom-up cues become less reliable. In this way, the technique again mirrors the behaviour of the human vision system, as researchers have found that the different types of depth cue each have different operating ranges [6]. This results in 3 general "perceptual spaces" for both the human vision system and the proposed framework: the *near space* (where the triangulation cue dominates), the *ambient space* (where a combination of bottom-up and top-down reconstruction is used), and the *vista space* where only top-down information is useful, as light entering the eye/sensor is near parallel and observations are approximately identical between the viewpoints. Our approach smoothly interpolates between these 3 situations, due to penalisation of the inverse depth error.

In our previous conference publication [7] we undertook an initial investigation of stereo representations which enable the unification of bottom-up and top-down reconstruction, as illustrated in Figure 1. In this paper we expand on this in all respects. In particular, the methodology has been greatly expanded, and now fully formalizes all linearised matching cost functions in Section 6. The technique has also been extended to allow integration with existing bottom-up stereo-reconstruction algorithms. Following this, the evaluation has been expanded and now investigates the effect of top-down reconstruction cues on a range of traditional state-of-the-art algorithms in Section 7.1. The behaviour of the various sub-components within the framework is also investigated in Section 7.2, which makes it possible to obtain a much deeper understanding of the behaviour of the approach. Finally in Section 8 we examine the approach in terms of the newly released HCI stereo metrics benchmark [8]. This explores geometry-aware aspects of the stereo reconstruction, which are particularly important for applications such as robotics and augmented reality.

## 2. Related Work

### *2.1. Bottom-up reconstruction*

The traditional approach to stereo reconstruction is based on matching and triangulation between multiple views of the same environment. *Local* bottom-up reconstruction techniques are based on independently detecting and matching small numbers of distinctive feature points. The most prevalent of these approaches is the matching of SIFT or similar features [9, 10]. More recently a number of additional matching criteria have been proposed, including the census transform [11], generative models [12] and edge preserving filters [13].

In the most general case, the matching may be run across the entire scene, to allow an initial estimate of the calibration to be obtained [14, 15]. However, when the calibration is already known, these local matching approaches can be limited to search only along the epipolar lines defined by the camera configuration.

A recent extension to this epipolar search was proposed as part of the semi-global matching approach, originally proposed by Hirschmüller [16]. This has gained significant popularity, due to it's robustness to minor calibration inaccuracies. Recent extensions of this idea include the addition of weighting terms [17], an iterative variant [18] and the augmented Manhattan world assumption [19, 20].

However, whichever matching technique is chosen, it is still impossible to fully reconstruct general scenes using only local matching information. This limitation is imposed by the well known *aperture problem* [21] which states that regions without strong texture (or more explicitly regions without significant gradient information perpendicular to the epipolar line) cannot be reliably matched between images, regardless of the descriptor which is used for matching. This is because, by definition, the matching point may move along the epipolar line, without causing a change in the observation in the other image. Due to this limitation, most work on bottom-up reconstruction (including the previously mentioned SGM techniques) has focused on global techniques, wherein various formulations of spatial smoothness are optimised alongside the local appearance matching. Some examples of these spatial-smoothness constraints include Total Variation (using both L1 and L2 norms [22, 23]), Monte-carlo inspired PatchMatch approaches [24, 25] and the Total Generalized Variation [26, 27] which was designed to deal with staircasing artifacts which are often introduced by total variation techniques.

An alternative way to encourage local smoothness, which has received significantly less attention in recent years, is to reconstruct the environment as a collection of scene primitives. This obviates the need to explicitly match every pixel, instead the matches are inherent in the configuration of primitives. This also helps to deal with common artifacts such as staircasing and oversmoothing, however the primitives which are chosen impose their own restrictions on the scene. For example smoothly curving surfaces can never be perfectly modelled using simple oriented planar primitives as used in many previous techniques [28, 29, 30], while the curved surfaces used by Zhang *et al*. [31] cannot model rough surfaces such as jagged stone. Despite this, reasonably chosen primitives can prove extremely effective for reconstruction within specialised domains such as urban environments. The most complex examples of primitive based reconstruction include earlier work by Wu and Levine [32] using geometric subunits or "geons" (*i.e.* full 3D shapes as primitives). More recently this idea has been extended even further using whole-object primitives [33, 34]. In this work, we make use of oriented planar primitives, which are particularly well suited to urban reconstruction.

One of the primary issues with reconstruction based on primitives, is determining the number and arrangement of the primitives. One of the most common approaches is to use superpixel segmentations to hypothesise consistent regions which can be well modelled by a single primitive. Mičušík and Košecká made use of this representation to segment the scene into planar primitives similar to the proposed technique [35]. However, unlike the proposed technique they used a discrete set of possible orientated planes, and focussed purely on bottom up matching and smoothness costs. More recent work by Bódis-Szomorú *et al*. [36, 37] has investigated highly efficient piecewise planar reconstruction, based on sparse input control points (from a separate SfM system).

### 2.2. Top-down reconstruction

The most common types of top-down reconstruction employ similar primitive sub-units, and then propose relational constraints between multiple primitives. Examples of these kinds of relationship at the local level (*e.g.* dealing with small collections of primitives), include data driven techniques to recognise common configurations of oriented planes [38], or concave/convex edges [5], and recent deep learning approaches which exploit prelearned representations [4]. These data-driven techniques are designed to exploit the recent prevalence of large scale reconstruction datasets, to learn which configurations are the most common and recognisable. In addition to detecting the existence of common primitive configurations, there has been work on categorizing types of inter-primitive relationships such as "occluding", "supporting" or "on-top-of" [39].

As is the case for bottom-up reconstruction, these local top-down techniques are commonly embedded within a global reconstruction framework. The global top-down interpretations tend to be particularly focussed on urban environments, perhaps the most well known global constraint in top-down reconstruction is the Manhattan-world assumption [40, 41, 42]; that scenes are composed entirely of planes, each having one of 3 orthogonal orientations [43]. For indoor environments, this idea has been successfully extended to coarsely modelling rooms as the interior of a cuboid, with between 2 and 5 visible faces [44, 45], perhaps containing box-shaped furnishings (the so called "box world" approach).

### 2.3. Joint approaches

There has previously been little work looking at techniques to unify both bottom-up and top-down knowledge during 3D reconstruction of urban environments. Additionally, although there have been a number of data-driven approaches to top-down scene understanding, there has been only limited exploitation of these large-scale reconstruction datasets in the traditional bottom-up reconstruction literature. The work that does exist generally falls within the domain of "Reconstruction meets Recognition", using an initial detection stage to locate pre-determined classes, to condition the following reconstruction stage. This has proven effective in a number of specific problem domains such as the outdoor urban reconstruction of Hane *et al*. [46] which initially splits the scene into buildings, sky, ground, vegetation and clutter. The 3D reconstruction that follows can then exploit collections of pre-learned weightings which favour particular types of reconstruction for each category (*e.g.* planar for buildings or horizontal for ground regions). There has also been work which obviates the need for pre-learned weightings by instead transferring generic object models into the 3D reconstruction when particular objects are recognised. So far this type of approach has been mostly limited to cars in urban environments. Cornelis *et al*. [47] used computer graphics techniques to blend the generic car models into the environment. However they required the background reconstruction to already be completed in order to constrain the placement of the car models. More recently Guney and Geiger [48] proposed a technique where the 3D car models were inserted before background reconstruction took place (i.e. the recognition constrained the reconstruction).

This type of approach is perhaps the closest to that proposed in this paper. However the proposed technique is more general as the cues are model-free and data-driven, rather than relying on a pre-selected set of object classes. As far as we are aware, the only work which attempts this is Saxena *et al*. [49], which used data-driven low-level monocular cues to improve stereo depth estimation. These monocular cues are similar to those used by Thomas *et al*. [50], except that they are purely data driven and do not depend on the detection of pre-selected object classes. However, this work still did not include any top-down reasoning about the entire scene. By combining both types of cues, the proposed technique has significant advantages as it inherently balances between both types of information, based on the type and reliability of the information available.

In the following sections, we present our novel framework for 3D scene reconstruction, which unifies the use of top-down and bottom-up cues in a manner inspired by the human vision system (Section 3). Two building blocks of bottom-up scene reconstruction are formalised for use within this unified representation in Sections 4.1 and 4.2. Section 5 then introduces cues from the other side of the spectrum: several aspects of top-down scene knowledge are presented and integrated into the framework. An efficient linear joint optimisation scheme is then introduced in Section 6. Finally, the proposed technique is evaluated on the recent Middlebury 2014 [51] and KITTI [52] benchmarks in Section 7. This evaluation includes examples of utilising top-down reconstruction within 3 existing state-of-the-art bottom-up reconstruction algorithms (Section 7.1) and an exploration of the characteristics of various subsystems within the framework (Section 7.2). We extend this evaluation in Section 8 using the recently proposed HCI stereo metrics [8] to examine various "geometrically inspired" properties of the reconstructions.

## 3. Unified bottom-up and top-down reconstruction

In both bottom-up and top-down approaches from the literature, the scene is defined as consisting of *primitives*. In the proposed method we localise our primitives using superpixel segmentation to obtain small contiguous scene regions. Superpixels are, however, not transferable between different views of a scene, as the segmentation employed for their extraction is viewpoint variant (this is especially the case for wide baseline reconstruction). In short, this means that it is not feasible to match superpixels from one viewpoint against superpixels from the second viewpoint. Instead the superpixels are used as an intermediate step for matching between the pixels of the (*reference* and *target*) images directly.

To achieve this, we use oriented planes to represent superpixels within the scene; each superpixel $s_i$ is parameterised by a single vector $\boldsymbol{\alpha}_i \in \mathbb{R}^3$. The direction of this vector represents the normal direction of the respective oriented plane, while the vector's length is equal to the inverse of the shortest (perpendicular) distance from the plane to the reference camera centre, which is taken to be the origin. With this parameterisation, two important relationships are observed. Firstly, every 3D point $\mathbf{p} \in \mathbb{R}^3$ lying on plane $\boldsymbol{\alpha}$ satisfies

$$\boldsymbol{\alpha}^\top \mathbf{p} = 1 \,. \tag{1}$$

Secondly, every ray $\mathbf{r}$ cast from the origin intersects the plane $\boldsymbol{\alpha}$ at distance

$$d = \frac{1}{\mathbf{r}^\top \boldsymbol{\alpha}} \,. \tag{2}$$

Every 3D plane induces a unique planar homography between two views. Without loss of generality, we assume these views are taken by a pair of non-parallel cameras. The geometric transformation between these views is described by the rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{t}$. The homography which the plane $\boldsymbol{\alpha}_i$ induces between images from these two views is then expressed by

$$\mathbf{H}_i = \mathbf{R} + \mathbf{t}\boldsymbol{\alpha}_i^\top \,. \tag{3}$$

In the reference image $\mathbf{I}^r$, we define a homogeneous point $\mathbf{x}^r$ (pixel coordinates) which lies within superpixel $s_i$. Given the plane-induced homography, the corresponding point $\mathbf{x}^t$ in the target image $\mathbf{I}^t$ can be computed as:

$$\mathbf{x}^t = \mathbf{K}_t \mathbf{H}_i \mathbf{K}_r^{-1} \mathbf{x}^r \,, \tag{4}$$

where $\mathbf{K}_r$ and $\mathbf{K}_t$ are the intrinsic calibration matrices of the reference and target cameras, respectively. Throughout this publication, we will refer to this projection more concisely as a function $\mathbf{x}^t = \mathrm{H}\left(\mathbf{x}^r | \boldsymbol{\alpha}_i\right)$, conditioned on the plane parameters. Also note that the ray $\mathbf{r}$ corresponding to a particular reference pixel $\mathbf{x}^r$ is simply defined as

$$\mathbf{r} = \frac{\mathbf{K}_r^{-1}\mathbf{x}^r}{\left\|\mathbf{K}_r^{-1}\mathbf{x}^r\right\|} \,. \tag{5}$$

Because this is a trivial mapping, the remainder of the paper omits this conversion in order to improve conciseness; for example both pixels and rays are interchangeably extracted from superpixels.

## 4. Bottom-up reconstruction

Using this formalisation, a number of standard bottom-up cost functions can now be formulated in terms of oriented plane primitives. For simplicity, we do not distinguish between homogeneous and non-homogeneous coordinates throughout this paper and use the same notation $(\mathbf{x})$ for both pixel locations (to index 2D images as in Equation (6)) and uncalibrated 3D rays (for 3D geometric computations as in Equation (5)).

### *4.1. Appearance matching*

The *Brightness Constancy* assumption is one of the most basic matching relationships, expressing the direct difference between two images in terms of their intensities. The associated energy is defined as

$$E_{bc}(s_i) = \sum_{\mathbf{x}^r \in s_i} \psi\left(\mathbf{I}^r(\mathbf{x}^r) - \mathbf{I}^t(\mathrm{H}(\mathbf{x}^r|\boldsymbol{\alpha}_i))\right) . \tag{6}$$

where $\psi$ is a robust cost function, preventing outlier over-penalisation. Similar relationships are the *Gradient Constancy* assumption with energy

$$E_{gc}(s_i) = \sum_{\mathbf{x}^r \in s_i} \psi\left(\mathbf{I}^r_{\Delta}(\mathbf{x}^r) - \mathbf{I}^t_{\Delta}(\mathrm{H}(\mathbf{x}^r_i|\boldsymbol{\alpha}_i))\right) \tag{7}$$

($\mathbf{I}^r_{\Delta}$ is the gradient image of $\mathbf{I}^r$ and $\mathbf{I}^t_{\Delta}$ of $\mathbf{I}^t$) and the *Modified Census Transform*

$$E_{ce}(s_i) = \sum_{\mathbf{x}^r \in s_i} \psi\left(\mathbf{I}^r_C(\mathbf{x}^r) \oplus \mathbf{I}^t_C(\mathrm{H}(\mathbf{x}^r|\boldsymbol{\alpha}_i))\right) , \tag{8}$$

where $\mathbf{I}_C$ are the census transform images [53]. The $\oplus$ symbol denotes the "exclusive or" binary operation, which is required as census transform similarity is defined via the Hamming distance.

### *4.2. Triangulation*

In addition to these dense pixelwise assumptions, local (feature-based) bottom-up cues are available. These originate from matching and triangulation, similar to a standard Structure from Motion (SfM) reconstruction pipeline. These triangulated matches are inherently sparse, however they tend to be more reliable and thus have a higher confidence. This can help to avoid local minima during optimisation, by ensuring that the initialisation is reasonable for the scene. In this work, the primary matching is based on CNN descriptors $\omega \in \mathbb{R}^{128}$, however the technique is not dependent on this and can make use of any feature correspondences. The CNN descriptors are computed with the deep network of [54], which consists of 6 convolutional layers, each followed by max-pooling, subsampling and rectification layers. The network was pre-trained on the Middlebury06 dataset [55] and is used as provided by the authors of [54].

The set of correspondences $\mathcal{C}$ between the two images defined by this matching is taken as the subset of matches with a low cosine error

$$\mathcal{C} = \left\{ \left(\mathbf{x}^r, \mathbf{x}^t\right) \middle| \frac{\omega^r \cdot \omega^t}{\parallel \omega^r \parallel \parallel \omega^t \parallel} > \lambda \right\} , \tag{9}$$

where

$$\omega^r = \mathrm{CNN}\left(\mathbf{I}^r(\mathbf{x}^r)\right) \quad \text{and} \quad \omega^t = \mathrm{CNN}\left(\mathbf{I}^t(\mathbf{x}^t)\right) . \tag{10}$$

In order to obtain 3D points whose projections correspond to these matches, the point depths are estimated by triangulation of the correspondences. To provide robustness to errors in the camera calibration and point localisation, the maximum-likelihood depth value is computed for a given correspondence [56]

$$\hat{d} = \mathbf{r}^\top \operatorname*{arg\,min}_{\mathbf{p}} \sum_{\mathbf{x} \in \{\mathbf{x}^r, \mathbf{x}^t\}} \|\Pi(\mathbf{p}) - \mathbf{x}\| , \tag{11}$$

where $\{\mathbf{x}^r, \mathbf{x}^t\}$ are matches from $\mathcal{C}$ and $\Pi$ denotes a camera projection function (for $\mathbf{I}^r$ or $\mathbf{I}^t$, according to $\mathbf{x}$). In other words $\hat{d}$ is the projection of the maximum likelihood 3D point $(\mathbf{p})$ for the correspondence, onto the ray $\mathbf{r}$ in question. Note that we omit the conversion from homogeneous to non-homogeneous coordinates for clarity, the distances are computed in image coordinates. The confidence $\hat{\nu}$ of the depth estimation (*i.e.* triangulation quality score) is taken as the reciprocal of the residual for this triangulation. Thus a large residual leads to a low confidence and small residuals lead to high confidences.

This information can be integrated into the proposed framework by formulating a sparse feature-based cost function, in terms of the planar primitives. However, because larger depth values are less reliable (uncertainty grows with the distance from the cameras), this cost function is formulated such that inconsistencies in the inverse depth are penalised. Similarly to the appearance constancy costs (Equations (6)–(8)), this provides a confidence based on theoretical limits on the information that can be extracted from the images [57] and therefore ensures a smooth transition between our different information sources. A relative error measure is used to penalise these inverse-depth inconsistencies (which are also known as the fractional depth error): $(\hat{d}-d)/d$. Here $\hat{d}$ is the (fixed) estimated sparse depth of the correspondence and $d$ is the refined depth on the oriented plane (defined by superpixel $s_i$). By making use of the relationship defined in Equation (2), this error measure can be expressed in terms of the plane parameters $\boldsymbol{\alpha}_i$:

$$\frac{\hat{d} - d}{d} = \frac{1}{d}\hat{d} - 1 = \mathbf{r}^\top \boldsymbol{\alpha}_i \hat{d} - 1 . \tag{12}$$

Finally, an energy (which is linear in the plane parameters) can be formulated by aggregating the inverse depth errors over all the triangulated correspondences, while accounting for their triangulation confidences

$$E_{tr}(s_i) = \sum_{\mathbf{x}^r \in s_i \cap \mathcal{C}} \hat{\nu} \psi\left(\mathbf{r}^\top \boldsymbol{\alpha}_i \hat{d} - 1\right) . \tag{13}$$

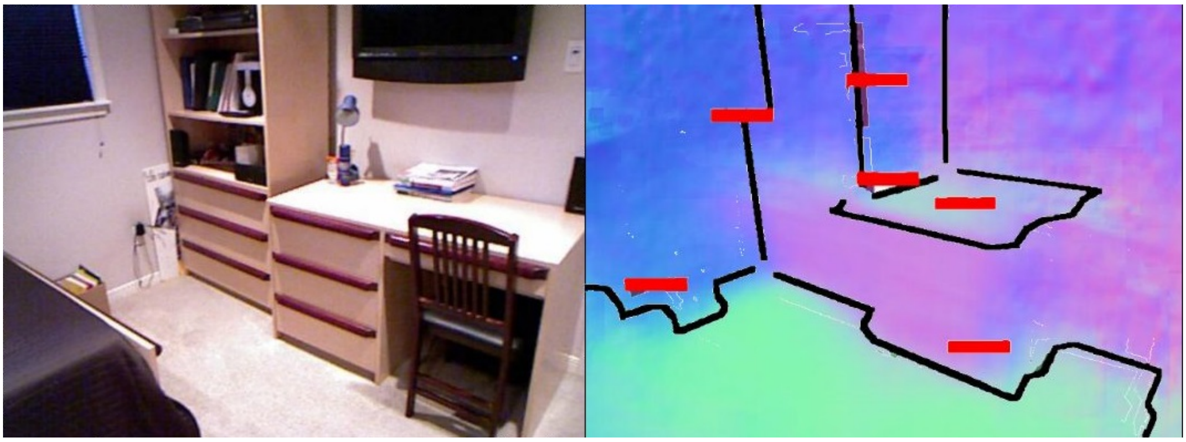


Figure 2: An example of an indoor scene interpretation in the "Origami world" [5, 38]. Left: input image; right: 3D scene interpretation with colour-coded normal directions (at every point the R-G-B colour channels represent the $z$-$x$-$y$ components of the vector normal to the surface.) We additionally show a set of discovered concave edges (denoted by the "minus" sign). No convex edges were detected in this example).

This triangulation energy function could also be used to integrate one or even several existing bottom-up reconstruction techniques into the proposed framework. This can be achieved simply by adding to the deep correspondences $\mathcal{C}$, and including any confidences $\hat{\nu}$ if available. In the experimental section 7.1 we exploit this to examine the value of top-down scene cues within current state-of-the-art stereo reconstruction systems.

## 5. Top-down reconstruction

All the bottom-up cues introduced in the previous section are formalised in terms of oriented planar primitives ($\boldsymbol{\alpha}$). This allows us to use them in a unified manner, alongside the top-down cues in the proposed joint framework. Top-down cues encapsulate knowledge of real human-made urban environments and can thus disambiguate between different solutions, greatly improving reconstruction accuracy.

One of the most commonly used high-level top-down cues is information about surface directions and types of edges present in the scene. We use a data-driven approach (following Fouhey *et al*. [5]) to probabilistically estimate surface normals and edge classes to create an "Origami world", as shown in Figure 2. This data-driven approach is able to exploit the recent availability of large-scale datasets to learn what scene structures are realistic. All of the top-down cues presented in this section are based on surface normals and edge classification provided by this approach, which is trained using the NYU dataset [58].

The edge categories recognised by the system are: *Concave*, *Convex*, *Occlusion boundary* and *No edge*. For every pixel in the image, the probability of belonging to each of these classes is estimated. In [5] this probability map is post-processed to extract consistent edges. However in this work we accumulate the class probabilities from each pixel along a particular superpixel boundary in order to estimate the class of that boundary.

### 5.1. *Surface normal consistency*

The obtained scene interpretation is used to produce a number of constraining energy functions. The first penalises inconsistencies between the *surface normal* images $\mathbf{I}_n^r$ and $\mathbf{I}_n^t$ estimated by [38]:

$$E_{sn}\left(s_i\right) = \sum_{\mathbf{x}^r \in s_i} \psi\left(\mathbf{R}\mathbf{I}_n^r\left(\mathbf{x}^r\right) - \mathbf{I}_n^t\left(\mathrm{H}\left(\mathbf{x}^r | \boldsymbol{\alpha}_i\right)\right)\right) . \quad (14)$$

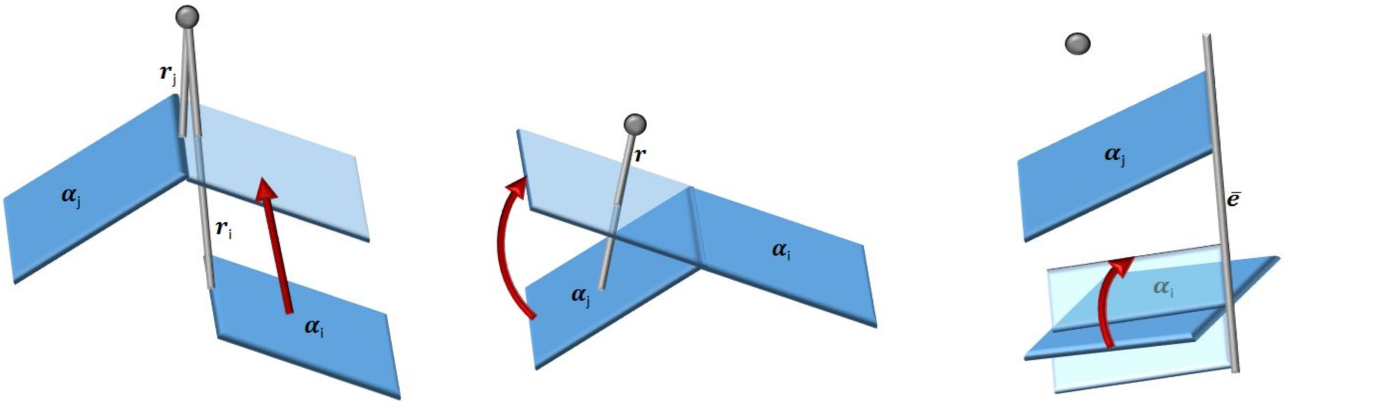
Rotation by  ($\mathbf{R}$) ensures that the normal directions are compared in a consistent co-ordinate frame (specifically the target camera's). Since this energy uses both images and penalises inconsistencies between them, it can be seen as a hybrid cue. Similar to the bottom-up cues, it is based on low level matching of elements of the top-down scene interpretation. Unlike the other top-down cues in this section this directly provides depth information for planes, not only pairwise constraints between planes.

### 5.2. *Connected structures*

In addition to enforcing a new type of consistency, we can use the interpretation of the scene to produce top-down pairwise constraints on the relationships between collections of oriented planes. One example of this is if the boundary between two neighbouring superpixels $s_i$ and $s_j$ is not detected as an occlusion boundary, we can favour reconstructions with a concave or convex (rather than disjoint) *connection* between the surfaces. If we define $\mathcal{N}_{i,j}$ as the set of pixels in $s_i$ which border $s_j$ then the fractional depth error across the boundary corresponds to $\frac{d_j - d_i}{\sqrt{d_i d_j}}$. This can be re-arranged in terms of $\boldsymbol{\alpha}$ using Equation (2) as:

$$E_{cn}\left(s_i, s_j\right) = \sum_{\mathbf{x}_i^r \in \mathcal{N}_{i,j}} \sum_{\mathbf{x}_j^r \in \mathcal{N}_{j,i}} \psi\left(\sqrt{d_i d_j}\left(\mathbf{r}_j^\top \boldsymbol{\alpha}_j - \mathbf{r}_i^\top \boldsymbol{\alpha}_i\right)\right) . \quad (15)$$

Figure 3a illustrates this idea. It should be noted that the neighbourhood pixel sets $\mathcal{N}$ are empty if the boundary between $s_i$ and $s_j$ is classified as an occlusion boundary.




(a) The cost function $E_{cn}$ encourages neighbouring planes to be connected in 3D (*i.e.* having aligned boundaries), if their configuration is not recognised as being an occlusion.

(b) The cost function $E_{cp}$ encourages neighbouring planes to be coplanar in 3D, assuming that both are subsections of a larger plane. This only occurs when the configuration is not recognised as being convex, concave or occlusion.

(c) The cost function $E_{cl}$ encourages a pair of non-neighbouring planes which have collinear edges in the image, to also have corresponding collinear edges in the 3D reconstruction.

Figure 3: Intuitive illustrations of the first 3 pairwise top-down configuration costs. The red arrow indicates the direction the plane should move to reduce each cost.

The inner brackets of this equation $(\mathbf{r}_j^\top \boldsymbol{\alpha}_j - \mathbf{r}_i^\top \boldsymbol{\alpha}_i)$ quantifies the difference in inverse depths between two ray and plane intersections. To simplify the remaining pairwise costs, we define a function

$$D(\mathbf{r}, \boldsymbol{\alpha}_i, \boldsymbol{\alpha}_j) = \mathbf{r}^\top \boldsymbol{\alpha}_i - \mathbf{r}^\top \boldsymbol{\alpha}_j \, , \tag{16}$$

to compute this difference of intersections (note that unlike Equation (15) this encodes the difference in intersections for the same ray with both planes).

### 5.3. Coplanarity

Similarly to this, the absence of any strong edge (either convex, concave or occlusion) indicates that neighbouring superpixels are part of a larger plane in 3D. Human-made urban scenes often contain large planar surfaces. This assumption is similar to, but less restrictive than, the full Manhattan world assumption. We can encode this as another top-down cue to aid the reconstruction, illustrated in Figure 3b. The resulting energy is based on the fractional depth change, arising from misalignment of the pair of neighbouring superpixels. This change is summed over the whole area of the two superpixels:

$$\begin{aligned} E_{cp}(s_i, s_j) = \gamma_{cp} \sum_{\mathbf{x}^r \in s_i} \psi\left(\sqrt{d_i d_j} D(\mathbf{r}, \boldsymbol{\alpha}_j, \boldsymbol{\alpha}_i)\right) + \\ \gamma_{cp} \sum_{\mathbf{x}^r \in s_j} \psi\left(\sqrt{d_i d_j} D(\mathbf{r}, \boldsymbol{\alpha}_j, \boldsymbol{\alpha}_i)\right) \, , \end{aligned} \tag{17}$$

where $\gamma_{cp}$ is an indicator function for *coplanar* superpixels, detected from the scene interpretation as described above.

### 5.4. Collinearity

Another constraint provided by the weak Manhattan-world assumption is the *collinearity constraint*. This says that if a pair of lines are collinear in the image, then they are likely to be parts of a repeating structure (such as a row of windows or bricks in urban environments) and thus they should also be collinear in the 3D reconstruction. This idea is shown in Figure 3c. The figure illustrates an important property of the collinearity cue: the two superpixels do not need to be adjacent to each other. Unlike the other pairwise cues, collinearity can encode long range relationships.

It should be noted that for this assumption to hold, we must also make a second assumption: that a straight line segment in the image relates to a straight line segment in the 3D world. This does not necessarily hold (for example if the edge's curvature is restricted to a planar subspace, and the camera centre also lies within that subspace). However, these cases are negligible in practice, and so it is reasonable to assume *a priori* that a straight line in 2D corresponds to a straight line in 3D.

A similar argument can be made about the validity of the collinearity cue. It is technically possible for lines which are collinear in 2D to be only parallel (not collinear) in 3D. However, this can only happen if the lines are offset from each other directly away from the camera (i.e. the two lines span a planar subspace which includes the camera centre). Again this situation is negligible in practice.

In order to formalize the collinearity constraint we note that the superpixel segmentation used to extract planar primitives already detects areas of significant gradient (*i.e.* line segments). Thus 2D lines $\bar{\mathbf{e}}$ are extracted from the boundaries of all superpixels in the image (note these are not line segments, but infinitely long lines). The set of pixels along each 2D line which also lie on the boundary of superpixel $s_i$, are referred to as $\mathcal{N}_{i,\bar{\mathbf{e}}}$ (no intersection means the set $\mathcal{N}_{i,\bar{\mathbf{e}}}$ is empty and the cost simplifies to zero). Given these points, which are collinear in 2D, we can then

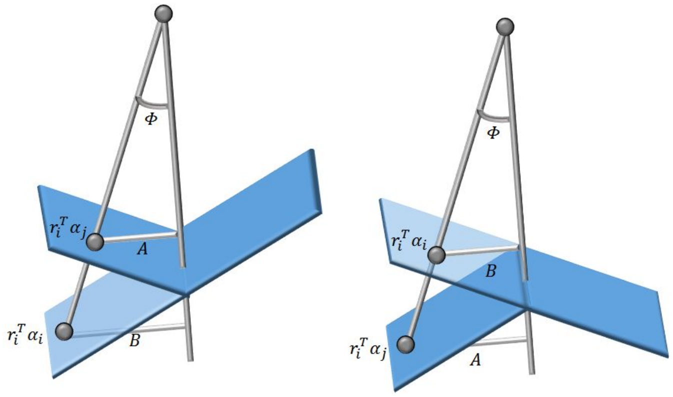



Figure 4: Intuitive illustration of the $E_{ed}$ cost function as defined in Equation (19), which encourages the reconstruction to obey the recognised convexity and concavity configurations. The left subfigure illustrates 2 planes ($i$ and $j$) with a concave relationship. The right subfigure shows a corresponding convex scenario. Rays which intersect plane $i$ can also be intersected with an extrapolated (shown as transparent) plane $j$. The relative magnitudes of $A$ and $B$ depend on the level of concavity.

quantify their degree of collinearity in 3D as:

$$E_{cl}\left(s_i, s_j\right) = \sum_{\mathbf{x}^r \in \mathcal{N}_{i,\bar{\mathbf{e}}}} \psi\left(\sqrt{d_i d_j} D\left(\mathbf{r}, \boldsymbol{\alpha}_j, \boldsymbol{\alpha}_i\right)\right) + \sum_{\mathbf{x}^r \in \mathcal{N}_{j,\bar{\mathbf{e}}}} \psi\left(\sqrt{d_i d_j} D\left(\mathbf{r}, \boldsymbol{\alpha}_j, \boldsymbol{\alpha}_i\right)\right) . \tag{18}$$

### 5.5. Edge categories

One final constraint arises directly from the scene interpretation, specifically from the extracted convex/concave *edge classification*. If such an edge lies between two neighbouring superpixels $s_i$ and $s_j$, we force the angle between these superpixels to agree with the edge class. This principle is illustrated in Figure 4. The concavity/convexity of the boundary is indicated by $\sin(\phi)(\hat{d}_j - d_i)$ where $\phi$ is the angle between the ray $\mathbf{r}$ and the ray through the superpixel boundary. The sign indicates the class (convex or concave), while the magnitude is proportional to the angle between the two oriented planes. The resulting energy is defined as:

$$E_{ed}\left(s_i, s_j\right) = \sum_{\mathbf{x}^r \in s_i} \psi_{ed}\left(\sin(\phi_i) D\left(\mathbf{r}, \boldsymbol{\alpha}_j, \boldsymbol{\alpha}_i\right)\right) + \sum_{\mathbf{x}^r \in s_j} \psi_{ed}\left(\sin(\phi_j) D\left(\mathbf{r}, \boldsymbol{\alpha}_j, \boldsymbol{\alpha}_i\right)\right) . \tag{19}$$

It should be noted that a different scoring function $\psi_{ed}$ is used here. In addition to the robust cost function, an initial linear mapping is performed between the estimated concavity/convexity probabilities. This allows us to exploit the probabilistic nature of the scene interpretation, such that the edge categorisation confidence is accounted for.

## 6. Optimisation

We integrate these unary and pairwise cues into a single cost function over the plane primitives:

$$E = \sum_{s_i \in \mathcal{S}} E_{bc}(s_i) + E_{gc}(s_i) + E_{ce}(s_i) + E_{tr}(s_i) + E_{sn}(s_i) + \sum_{s_i \in \mathcal{S}} \sum_{s_j \in \mathcal{S}} E_{cn}(s_i, s_j) + E_{cp}(s_i, s_j) + E_{cl}(s_i, s_j) + E_{ed}(s_i, s_j) , \tag{20}$$

which is minimised to obtain the optimal plane parameters

$$\boldsymbol{\alpha} = \arg\min_{\bar{\boldsymbol{\alpha}}} E\left(\bar{\boldsymbol{\alpha}}\right) . \tag{21}$$

To balance the contribution of all the independent cost functions, each energy is weighted with an appropriate weight. These weights are applied at the same time as the scoring functions $\psi$ and are aggregated in a weighting vector $\boldsymbol{\upsilon}$. These weights are learned from example data. First a superpixel segmentation is performed, and planes are fit to the ground truth depths for each segment. These ground truth plane parameters are referred to as $\boldsymbol{\alpha}_*$. We can then optimise the weights to minimise the difference between the predicted plane parameters and the ground truth:

$$\boldsymbol{\upsilon} = \arg\min_{\bar{\boldsymbol{\upsilon}}} \left| \boldsymbol{\alpha}_* - \arg\min_{\bar{\boldsymbol{\alpha}}} E\left(\bar{\boldsymbol{\alpha}} | \bar{\boldsymbol{\upsilon}}\right) \right| . \tag{22}$$

similar to [59].

With the proposed formulation, all the energies are linear in these $\boldsymbol{\alpha}$ parameters, except for the cost function $\psi$ (which is actually applied by the robust optimiser) and the image lookups in Section 4.1. These are linearised via a first-order Taylor expansion. This formulation is similar to a derivation of the *optical flow constraint* from the brightness constancy constraint in motion estimation [60, 61]. The linearity of the problem then allows a very efficient solution using sparse linear programming.

To illustrate how the image lookups (and thus the resulting match functions) are linearised in terms of $\boldsymbol{\alpha}$, we illustrate the linearisation of the Brightness Constancy cost (introduced in Equation (6))

$$E_{bc}\left(s_i\right) = \sum_{\mathbf{x}^r \in s_i} \psi\left(\mathbf{I}^r\left(\mathbf{x}^r\right) - \mathbf{I}^t\left(\mathrm{H}\left(\mathbf{x}^r | \boldsymbol{\alpha}_i\right)\right)\right) , \tag{23}$$

however, it extends trivially to the other matching costs.

As mentioned at the end of Section 4, this equation omits the conversion from 3 element homogeneous pixel positions ($\mathbf{x}$) to 2D image locations ($\tilde{\mathbf{x}}$). We now include a conversion

$$\tilde{\mathbf{x}} = \mathrm{G}(\mathbf{x}) \tag{24}$$

into the cost function explicitly

$$E_{bc}\left(s_i\right) = \sum_{\mathbf{x}^r \in s_i} \psi\left(\mathbf{I}^r\left(\mathrm{G}\left(\mathbf{x}^r\right)\right) - \mathbf{I}^t\left(\mathrm{G}\left(\mathrm{H}\left(\mathbf{x}^r | \boldsymbol{\alpha}_i\right)\right)\right)\right) . \tag{25}$$

The lookup in the reference image ($\mathbf{I}^r$) does not depend on $\boldsymbol{\alpha}$ and so does not need to be linearised. For the target image ($\mathbf{I}^t$) lookup, we perform a Taylor expansion and drop all terms of quadratic order or higher. If we have a current estimate $\boldsymbol{\alpha}^0$, we can perform the Taylor expansion around this and obtain the parameter update $\Delta\boldsymbol{\alpha}$:

$$\mathbf{I}^t\left(\mathrm{G}\left(\mathrm{H}\left(\mathbf{x}_i^r|\boldsymbol{\alpha}_i^0+\Delta\boldsymbol{\alpha}\right)\right)\right)\approx\mathbf{I}^t\left(\mathrm{G}\left(\mathrm{H}\left(\mathbf{x}_i^r|\boldsymbol{\alpha}_i^0\right)\right)\right)+\mathbf{J}\Delta\boldsymbol{\alpha}\,, \tag{26}$$

where $\mathbf{J}$ is the Jacobian of the combined function.

The function being approximated can be viewed as the composition of 3 functions ($\mathbf{I}^t$, G and H). Thus, we can compute $\mathbf{J}$ in closed form using the total derivative chain rule,

$$\mathbf{J}=\mathbf{J}_I(\mathrm{G}\left(\mathrm{H}\left(\mathbf{x}_i^r|\boldsymbol{\alpha}_i\right)\right))\,\mathbf{J}_G(\mathrm{H}\left(\mathbf{x}_i^r|\boldsymbol{\alpha}_i\right))\,\mathbf{J}_H(\mathbf{x}_i^r)\,. \tag{27}$$

In other words, $\mathbf{J}$ is the matrix product of the three Jacobians ($\mathbf{J}_I$, $\mathbf{J}_G$ and $\mathbf{J}_H$) for the composed sub-functions, with each Jacobian being evaluated at the location output by its preceding sub-function.

To define the first sub-Jacobian ($\mathbf{J}_H\in\mathbb{R}^{3\times3}$), remember that the H function is defined as the application of 3 matrix multiplications to the pixel position. First the inverse of the calibration matrix for the reference camera, second the homography matrix induced by the plane, and finally the intrinsic matrix for the target camera

$$\mathrm{H}\left(\mathbf{x}^r|\boldsymbol{\alpha}_i\right)=\mathbf{K}_t\mathbf{H}_i\mathbf{K}_r^{-1}\mathbf{x}^r=\mathbf{x}^t\,. \tag{28}$$

The matrix $\mathbf{H}_i$ is defined as $\mathbf{H}_i=\mathbf{R}+\mathbf{t}\boldsymbol{\alpha}_i^\top$ and is the only part of the equation which depends on $\boldsymbol{\alpha}$. As such, the sub-jacobian $\mathbf{J}_H$ in terms of $\boldsymbol{\alpha}$ is given by

$$\mathbf{J}_H(\mathbf{x}^r)=\mathbf{K}_t\mathbf{t}(\mathbf{K}_r^{-1}\mathbf{x}^r)^\top\,, \tag{29}$$

the intrinsics of the target camera, and the outer product of the camera baseline with the normalised homogeneous pixel position.

The second sub-Jacobian ($\mathbf{J}_G\in\mathbb{R}^{2\times3}$) is the simplest, given by

$$\mathbf{J}_G(\mathbf{x}^t)=\begin{bmatrix}\frac{1}{w} & 0 & -\frac{u}{w^2}\\ 0 & \frac{1}{w} & -\frac{v}{w^2}\end{bmatrix} \tag{30}$$

where $u,v,w$ are the elements of $\mathbf{x}^t$.

The final sub-Jacobian $\mathbf{J}_I\in\mathbb{R}^{1\times2}$ encodes how the image intensity varies as a result of changes in the pixel position, and is constructed from the $x$ and $y$ gradients of the target image.

$$\mathbf{J}_I(\tilde{\mathbf{x}}_i^t)=\mathrm{I}_\Delta^t(\tilde{\mathbf{x}}_i^t) \tag{31}$$

Given these definitions, we can substitute the approximation of Equation (26) into Equation (25):

$$E_{bc}\left(s_i\right)=\textstyle\sum_{\mathbf{x}_i^r\in s_i}\psi\Big(\ \mathbf{I}^r\left(\mathrm{G}(\mathbf{x}_i^r)\right)-\mathbf{I}^t\left(\mathrm{G}(\mathrm{H}(\mathbf{x}_i^r|\boldsymbol{\alpha}_i^0))\right) -\mathbf{J}_I\mathbf{J}_G\mathbf{J}_H\Delta\boldsymbol{\alpha}\Big). \tag{32}$$

This cost function is linear in terms of $\Delta\boldsymbol{\alpha}$ (ignoring for a moment the robust penalty function $\psi$).

Now all cost functions can be expressed in a form that is linear in terms of $\boldsymbol{\alpha}$. If the robust penalty function $\psi$ is defined as the $\ell_1$ norm, this becomes a linear Least Absolute Deviation regression problem. This can be equivalently formulated as a linear programming problem by introducing slack variables. Reformulating into a linear program also allows us to introduce the additional constraints that all planes must be in front of the camera

$$\mathbf{r}^\top\boldsymbol{\alpha}_i>0,\quad\forall\mathbf{r}\in s_i\,. \tag{33}$$

Further constraints could be included, such as enforcing a depth limit, however we did not find this to be necessary.

Due to the linear approximation described above, we iteratively solve this problem such that $\boldsymbol{\alpha}^{n+1}=\boldsymbol{\alpha}^n+\Delta\boldsymbol{\alpha}$, which is repeated until convergence. In practice we find that the approach generally converges after only 3 iterations.

## 7. Evaluation

The proposed approach is initially evaluated on the recent Middlebury 2014 dataset [51]. This dataset consists of 33 pairs of high definition (≈6 megapixels each) stereo images. In addition, 3 different resolution modes are provided (marked as F, H and Q for full, half and quarter resolution, respectively), rendering this an extremely large scale dataset, with around 350 megapixels of image data. The performance of the various techniques is computed for fully dense estimates, including occluded regions. The default settings for the Middlebury evaluation are to ignore occluded regions using the "nonocc" mask. However, failures in these regions are common and can cause major problems in many applications such as robotic navigation and augmented reality. Extrapolation in occluded regions is also an area where a holistic understanding of the scene may prove extremely valuable. We tabulate the average and RMS error in terms of disparity levels to give an idea of overall accuracy. In addition we tabulate the 99th percentile error (referred to as A99 in the Middlebury2014 benchmark), which provides an indication of the quantity and magnitude of outliers in the reconstruction. This can be seen as a measure of robustness (*i.e.* catastrophically incorrect interpretations of the scene). Lower is better for all performance measures. The system was implemented in Matlab, and all runtime measurements ran on a single Intel Sandy Bridge core at 2.4 GHz and required a maximum of 4 GB of memory.

One advantage of the proposed framework is that there are few free parameters. The optimal weightings $\boldsymbol{\upsilon}$ for the various energy terms are learned. The older Middlebury 2006 [55] dataset was used for this purpose, as both the training and testing sets of Middlebury 2014 are used for evaluation. A matching threshold $\lambda$ of 0.5 was employed for the CNN triangulation. The initial superpixel segmentation was performed using the efficient graph-based approach of

| Technique | Mode | RMS Err. (px) | Avg.Err. (px) | A99 (px) | Time (s) |
|---|---|---|---|---|---|
| CoR [62] | Q | 27.2 | 11.1 | 131 | **9** |
| **CoR + HLSC** | | **26.1** | **10.5** | **113** | 118 |
| MC-CNN [63] | | 34.0 | 11.7 | 160 | 16 |
| **MC-CNN + HLSC** | | 34.5 | 14.4 | 148 | 142 |
| Mesh Stereo [64] | | 32.3 | 13.1 | 156 | 572 |
| **Mesh Stereo + HLSC** | | 31.5 | 13.2 | 139 | 742 |
| CoR [62] | H | 27.6 | 9.9 | 126 | **38** |
| **CoR + HLSC** | | **25.7** | **9.6** | **112** | 1352 |
| MC-CNN [63] | | 36.4 | 11.7 | 176 | 148 |
| **MC-CNN + HLSC** | | 39.0 | 15.0 | 174 | 1483 |
| Mesh Stereo [64] | | 42.0 | 20.4 | 166 | 1067 |
| **Mesh Stereo + HLSC** | | 42.4 | 21.4 | 162 | 2856 |
| CoR [62] | F | 26.7 | 9.6 | 123 | **262** |
| **CoR + HLSC** | | **25.5** | **9.4** | **115** | 9162 |
| **CoR + HLSC** | Test H | **38.9** | **12.8** | **175** | **1972** |

Table 1: Performance of the unified framework, when including High Level Scene Cues (HLSC) with 3 state-of-the-art approaches to bottom-up matching. Bold numbers indicate the best score within that resolution mode.

| Technique | RMS Err. (px) | Avg.Err. (px) | A99 (px) | Time (s) |
|---|---|---|---|---|
| CoR [62] | **3.41** | **0.98** | 16.8 | **17** |
| **CoR + HLSC** | 3.69 | 1.32 | 17.5 | 352 |
| MC-CNN [63] | 3.51 | 0.99 | **15.8** | 133 |
| **MC-CNN + HLSC** | 4.42 | 1.48 | 19.8 | 354 |
| Mesh Stereo [64] | 6.93 | 2.98 | 36.4 | 378 |
| **Mesh Stereo + HLSC** | 6.69 | 2.46 | 32.2 | 528 |

Table 2: Performance of the unified framework, when including High Level Scene Cues (HLSC) with 3 state-of-the-art approaches on the KITTI dataset [52]. Bold numbers indicate the best score.

Felzenszwalb and Huttenlocher [65], for which the default segmentation threshold of 40 was used. The training data for the top-down cues was taken from the NYU depth dataset of urban indoor scenes [58].

### *7.1. The value of high level scene cues*

First we examine the value of the proposed framework to integrating high-level cues into stereo reconstruction. The approach is integrated into a number of leading bottom-up stereo reconstruction algorithms:

**MC-CNN** (Matching Cost Convolutional Neural Network) [63] where deep learning techniques were employed to develop a stereo matching cost function which is then refined via semi-global matching. This is implemented in parallel on the GPU, which makes it extremely fast but unfortunately prevents it running on the *Full* resolution benchmark (due to limitations on GPU memory).

**Mesh Stereo** [64], a technique designed to produce 3D triangular meshes of the environment, primarily for view interpolation. The technique is implemented in C++ but again due to the scaling of memory usage it cannot be run on the *Full* resolution benchmark.

**CoR** (Consensus of Regions) [62] which performs simultaneous optimisation and outlier detection across scales, using a semiglobal matching cost.

The results are presented in Table 1, with bold numbers indicating the best score within the operating mode. The results show that integrating high level scene cues into stereo reconstruction provides significant improvements in the robustness of the obtained reconstructions. As measured by the A99 score, occurrences of unrealistic scene interpretations and outliers is significantly reduced in every operating mode for every evaluated technique. On average the A99 improvement was around 11 % (ranging up to 15 % for Mesh Stereo). We also see smaller gains in terms of reconstruction accuracy (measured by the average and RMS errors), of up to 5 %. The runtimes when integrating high level scene cues obviously include the runtimes for the base technique, and add a fairly constant overhead regardless of the technique. The additional overhead is insignificant for techniques such as Mesh stereo, but for GPU accelerated techniques such as MC-CNN it is more noticeable. The last row of the table shows the performance on the test data. These results are slightly lower than on the training data (despite no learning actually being performed on this data). This is likely due to a bias towards larger disparities in the test data (the average disparity limit is 330 on the training

| | Res. | Adirondack | ArtL | Jadeplant | Motorcycle | MotorcycleE | Piano | PianoL | Pipes |
|---|---|---|---|---|---|---|---|---|---|
| RMS Err. (px) | Q | 13.8 / 2 | 22.4 / 2 | 75.8 / 2 | 26.6 / 2 | 27.0 / 2 | 11.9 / 1 | 17.0 / 1 | 37.4 / 3 |
| | H | 11.8 / 5 | 22.0 / 9 | 76.7 / 2 | 24.9 / 10 | 25.0 / 9 | 10.4 / 2 | 18.0 / 3 | 34.0 / 9 |
| | F | 13.1 / 2 | 19.7 / 4 | 80.3 / 2 | 25.5 / 6 | 25.5 / 5 | 10.5 / 3 | 13.5 / 2 | 30.5 / 5 |
| Avg. Err. (px) | Q | 4.58 / 2 | 11.7 / 2 | 37.6 / 2 | 8.16 / 2 | 8.23 / 2 | 4.87 / 1 | 7.90 / 1 | 17.0 / 4 |
| | H | 3.35 / 5 | 9.7 / 9 | 35.0 / 7 | 6.85 / 10 | 6.87 / 9 | 3.92 / 4 | 7.30 / 3 | 13.8 / 12 |
| | F | 3.38 / 2 | 7.9 / 5 | 36.0 / 3 | 7.01 / 6 | 6.95 / 5 | 3.70 / 2 | 5.64 / 1 | 11.3 / 5 |
| A99 (px) | Q | 65.4 / 2 | 76.1 / 1 | 280 / 1 | 141 / 2 | 141 / 2 | 41.4 / 1 | 66.3 / 1 | 155 / 3 |
| | H | 58.5 / 2 | 86.1 / 7 | 311 / 2 | 139 / 11 | 137 / 9 | 35.3 / 1 | 93.3 / 6 | 146 / 8 |
| | F | 65.9 / 2 | 81.0 / 4 | 321 / 2 | 144 / 6 | 143 / 5 | 36.2 / 2 | 47.5 / 1 | 142 / 5 |

| | Res. | Playroom | Playtable | PlaytableP | Recycle | Shelves | Teddy | Vintage | **Average** |
|---|---|---|---|---|---|---|---|---|---|
| RMS Err. (px) | Q | 24.7 / 2 | 28.7 / 1 | 12.2 / 2 | 19.1 / 4 | 20.5 / 2 | 9.0 / 2 | 50.8 / 3 | 26.1 / 2 |
| | H | 32.1 / 10 | 37.4 / 5 | 11.3 / 6 | 15.9 / 10 | 22.0 / 11 | 11.1 / 6 | 47.0 / 13 | 25.7 / 7 |
| | F | 29.5 / 6 | 29.0 / 2 | 11.7 / 4 | 14.6 / 4 | 19.5 / 3 | 16.5 / 7 | 50.5 / 7 | 25.5 / 4 |
| Avg. Err. (px) | Q | 8.5 / 2 | 10.0 / 1 | 4.97 / 2 | 5.21 / 2 | 10.9 / 2 | 3.76 / 2 | 13.6 / 1 | 10.5 / 2 |
| | H | 10.1 / 10 | 16.6 / 8 | 3.90 / 6 | 3.55 / 7 | 11.7 / 11 | 3.02 / 7 | 14.6 / 7 | 9.6 / 6 |
| | F | 9.7 / 6 | 14.7 / 2 | 3.61 / 4 | 3.36 / 3 | 12.7 / 7 | 4.68 / 6 | 15.4 / 4 | 9.4 / 3 |
| A99 (px) | Q | 122 / 2 | 158 / 1 | 55.3 / 2 | 87.6 / 4 | 90.1 / 2 | 44.5 / 2 | 217 / 3 | 113 / 2 |
| | H | 192 / 10 | 162 / 4 | 51.5 / 5 | 68.0 / 7 | 92.6 / 11 | 43.6 / 4 | 44 / 5 | 112 / 7 |
| | F | 175 / 7 | 129 / 2 | 49.5 / 2 | 64.1 / 3 | 82.2 / 4 | 78.0 / 7 | 182 / 6 | 115 / 4 |

Table 3: The performance for all resolution benchmarks on all sequences. The error value is listed alongside the ranking on that sequence (out of 5, 14 and 8 for the Q, H and F benchmarks respectively).

data and 440 on the test data). This is confirmed by the rankings compared to the rest of the leaderboard. Despite the drops in performance of the proposed technique, the ranking does not change for average or RMS error, and only changes slightly for the A99 error.

When contrasting the different algorithms CoR is generally the most accurate technique both with and without integration in the proposed framework. However, MC-CNN is competitive at the lowest resolutions, where CoR has fewer observations to support the automatic outlier rejection. Of the three techniques, Mesh stereo is generally the least accurate. However, this is unsurprising as the technique primarily focuses on supporting realistic novel viewpoints, rather than refining the input viewpoints. It is interesting to note that this focus does not lead to an improvement in realism or reduction in gross outliers (measured by A99) compared to the other techniques. This is contrary to expectations, as outliers are likely to have a significant effect on interpolated viewpoints.

Table 2 repeats the evaluation on the KITTI dataset [52], which contains around 200 outdoor urban environments viewed from a vehicle. The findings are similar, again CoR generally performs the best, although MC-CNN is closer in this case, and actually outperforms CoR in terms of A99 score. In this case the High Level Scene Cues make the most significant difference to Mesh Stereo. However, the observed improvement in robustness is slightly more modest than for the Middlebury dataset and performance decreases slightly for CoR. This is likely due to the sparse ground truth for the KITTI dataset (collected via LIDAR) which covers only around 30 % of the scene. The sparse ground truth slightly biases the evaluation away from areas of common reconstruction errors, such as occlusion boundaries.

We now examine the performance of the proposed framework in detail on each sequence from the Middlebury 2014, for each resolution mode. To aid clarity, we include only the top performing variant of the proposed integrated technique above (**CoR + HLSC**) which also enables us to perform the remaining evaluations even in *Full* mode. In Table 3, it can be seen that including high level scene cues helps make the algorithm particularly robust, leading to the top ranking in terms of A99 score on the majority of scenes. This helps reduce outliers in areas of low texture information such as uniformly painted surfaces, which are common in urban environments. It is particularly illuminating to note the performance of the proposed technique in the special scenarios (denoted with suffixes). For example the dataset contains 2 scenes (*ArtL* and *PianoL*) which include lighting changes between the two views. There is also one scene (*MotorcycleE*) with a significantly different exposure level in the second view. The proposed technique proves to be extremely resilient to these challenging inconsistencies compared to the traditional bottom-up reconstruction approaches. For example, performance on the *Motorcycle* sequence with and without the exposure change are roughly the same, which leads to a consistent improvement in the ranking across all the error measures, because competing approaches are more adversely affected. The lighting change on the *Piano* sequence causes a 20 % performance drop in our algorithm, however this is again lower than the drop experienced by competing techniques, and actually leads to a moderate increase in ranking. The final scene suffix P indicates "perfect" camera calibration. Although

| Technique | RMS Err. (px) | Avg.Err. (px) | A99 (px) | Time (s) |
|---|---|---|---|---|
| **Top-down (ours)** | 163.6 / 7 | 155.3 / 7 | 279 / 7 | 68 / 4 |
| **Bottom-up (ours)** | 27.8 / 3 | 11.6 / 3 | 151 / 3 | 83 / 5 |
| **Full HLSC (ours)** | 26.1 / 2 | 10.5 / 2 | 113 / 2 | 118 / 6 |

Table 4: The performance of the proposed system when including different types of cues. The ranks are out of 7.

| | Omitted energy | RMS Err. (px) | Avg.Err. (px) | A99 (px) | Time (s) |
|---|---|---|---|---|---|
| Bottom-up | $E_{bc}$ | 39.2 | 20.5 | 168 | 91 |
| | $E_{gc}$ | 33.5 | 12.8 | 146 | 94 |
| | $E_{ce}$ | 32.3 | 11.9 | 134 | **81** |
| | $E_{tr}$ | 49.9 | 23.1 | 178 | 121 |
| Top-down | $E_{sn}$ | 28.5 | 13.6 | 125 | 105 |
| | $E_{co}$ | 26.5 | 13.8 | 142 | 106 |
| | $E_{cp}$ | 26.7 | 12.8 | 143 | 98 |
| | $E_{cl}$ | 27.4 | 10.6 | 118 | 109 |
| | $E_{ed}$ | 28.0 | 10.9 | 123 | 95 |
| | - | **26.1** | **10.5** | **113** | 118 |

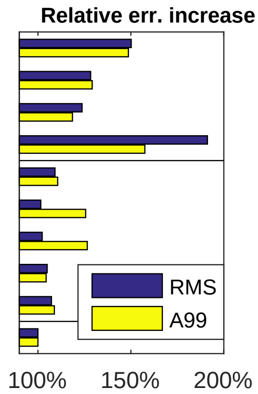



Table 5: Evaluation of the performance of the proposed system when each cue is removed in turn.

this significantly improves the accuracy of the proposed technique, it has little effect on its ranking, suggesting that the gains in accuracy mostly come from the bottom-up system which our technique is integrated with. Because the High Level Scene Cues are based on a holistic understanding of the scene contents, it makes sense that they would exhibit a reduced sensitivity to calibration errors.

In Figure 5 a number of randomly chosen example reconstructions are shown for the *Full* resolution (6 megapixel) mode of Middlebury 2014. A number of results on the KITTI dataset are also shown in Figure 6. As mentioned above, the ground truth for the KITTI dataset does not cover the whole images. For this reason the ground truth images in the last column are dilated to facilitate comparison with the estimated results. Figure 7 allows a qualitative comparison of the reconstructions with and without High Level Scene Cues, by showing the signed error images side by side. It is clear from these results that the benefits of High Level Scene Cues are most prominent in ambiguous areas such as at depth discontinuities. This is particularly evident surrounding the foremost armrest in *Adirondack* and around the Chimney in Teddy.

### *7.2. Examination of subsystems*

In Table 4 we evaluate the contribution of different types of cues within the proposed framework. Using only top-down reasoning with no matching proves to be significantly faster, however the quality of the estimate is poor as finer details are no longer modelled. This is probably because the scene interpretation used to provide the top-down cues is prone to oversmoothing. It is interesting to note that as a consequence, the decrease in robustness (increasing the A99 error by a factor of 3) is significantly lower than the loss of accuracy (increasing the RMS error by a factor of 10). When the technique exploits only bottom-up matching cues, the reconstruction is of higher quality. However, the combination of both the bottom-up and top-down cues performs the best, with around 10 % improvement in all error measures. This demonstrates the complementary nature of the different cues, particularly improving robustness by resolving ambiguities.

In Table 5 we expand on this evaluation, by testing the contribution of every individual energy term from Equation (20). In all cases, removing an energy term causes an increase in the error, meaning that each term encodes useful information for 3D reconstruction, and none of the terms are redundant. To aid visualisation, the increase in error when a cue is removed are plotted visually next to the table. In general we see that the bottom-up terms in the top half of the table have a significant affect on accuracy, with $E_{tr}$ and $E_{bc}$ contributing most strongly to both the RMS and Average error measures. We also note that the top-down terms in the bottom half of the table have a disproportionately large effect on the reconstruction robustness (the yellow bars are longer than the blue for top-down cues, and shorter for bottom-up cues). For instance, while the census transform cost (the least important of the bottom-up cues) is more important than all of the top-down costs in terms of RMS error, it has only average importance in terms of the A99 measure. The connected and coplanar structure costs prove to be the most valuable of the top-down cues.

We also explore the repeatability of the surface normal estimation system. The agreement of the surface normals estimated in the 2 views was quantified by

$$\frac{\mathbf{R}\mathbf{I}_n^r\left(\mathbf{x}^r\right)\cdot\mathbf{I}_n^t\left(\tilde{\mathbf{x}}^t\right)+1}{2}, \tag{34}$$

Figure 5: Randomly chosen *Full* resolution reconstructions from the Middlebury 2014 dataset. One input image (left), the output of our algorithm (middle) and the ground truth reconstruction (right). From top to bottom: *Adirondack*, *ArtL*, *Piano*, *Pipes*, *PlaytableP*, *Recycle*.

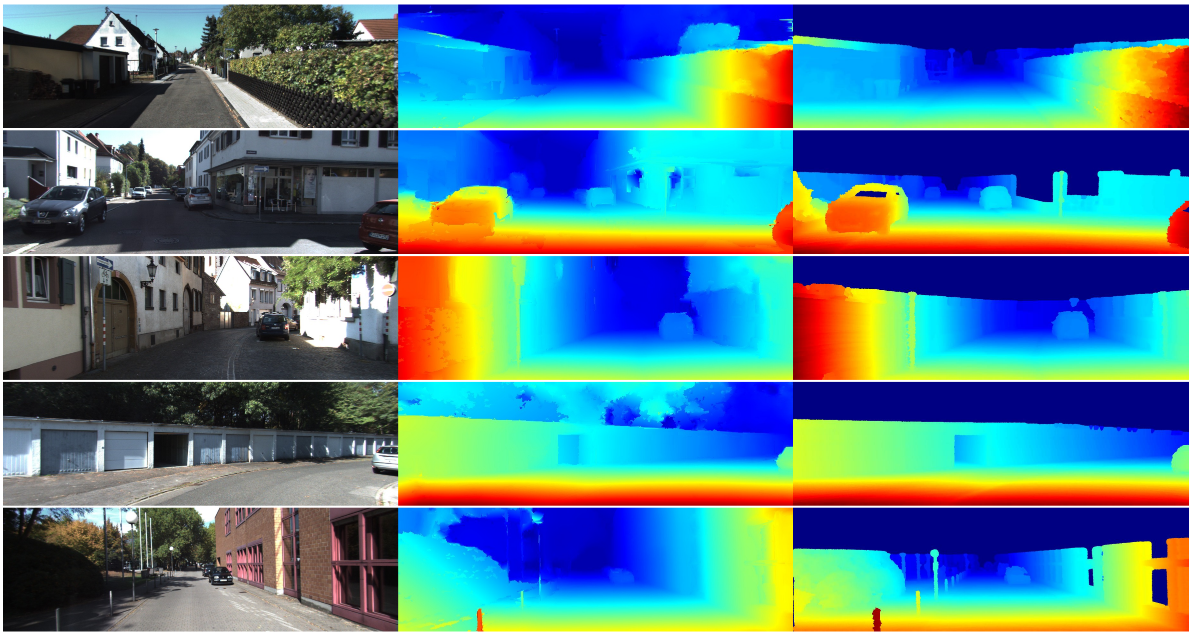


Figure 6: Randomly chosen examples from the KITTI dataset. Input image (left), the output of our algorithm (middle) and the ground truth reconstruction (right, dilated). From top to bottom the frames scenes are: 1, 19, 18, 23, 26.

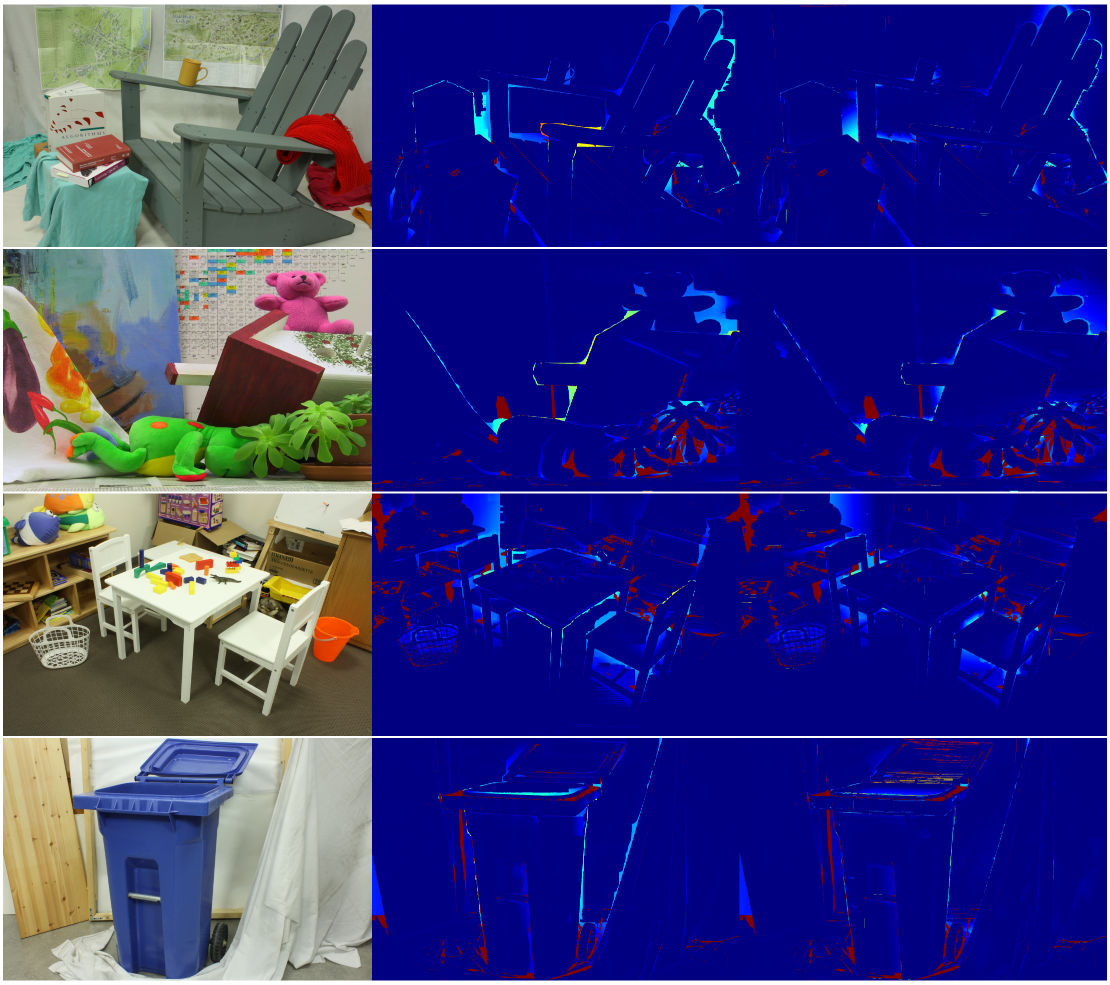


Figure 7: Signed error images for the baseline technique (centre) and when including using High Level Scene Cues (right). From top to bottom the scenes are: *Adirondack*, *Teddy*, *PlaytableP* and *Recycle*

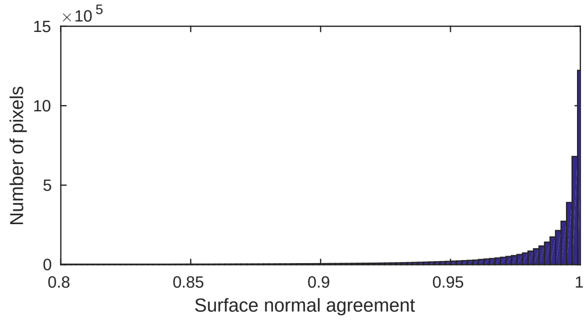



Figure 8: Distribution of consistency between the two viewpoints for estimated surface normals.

where $\tilde{\mathbf{x}}^t$ is the corresponding pixel in the target image according to the ground truth disparity map. In other words, the dot product of the 2 normal vectors in the target co-ordinate frame, normalised between 0 and 1.

The distribution across all pixels from all scenes in the Middlebury 2014 training set is displayed in Figure 8. The average consistency across viewpoints is 0.982, with almost nothing below 0.95. This implies that the large scale of the dataset renders the data driven scene understanding cues extremely reliable in the multiview scenario, even though they are only trained from monocular data.

To measure the sensitivity of the overall reconstruction framework to this, we also computed the correlation between the surface normal agreement and the disparity error (again over all pixels in the training sequences). The resulting correlation coefficient was -0.08, indicating a slight anti-correlation (*i.e.* increased surface normal agreement indicates a reduction in disparity error). This is reasonable as matching estimated surface normals is one of the inputs to the system, however the sensitivity proves very slight due to the influence of other cues.

In Figure 9 we examine the effect of the stereo baseline on the performance of the proposed system. The Middlebury 2014 and KITTI datasets are poorly suited for this evaluation as there is little variation in baseline, and the change in scene clutter between different scenarios has a far more significant effect on performance. Instead we use the Middlebury 2003 dataset which includes a larger array of cameras. We can then use different pairs of images to simulate stereo pairs with different baselines, but all viewing the same scene. The results show that an extremely narrow baseline is the most detrimental, and that good performance can be obtained for a wide range of baselines. However, there is an eventual decay in performance when the baseline becomes too large.

We also evaluate the effect of varying the superpixel segmentation threshold in Figure 10 to vary the size of the superpixels used on the *Quarter* Resolution benchmark. Higher thresholds lead to a smaller numbers of larger superpixels, and can significantly improve the runtime of the algorithm. However, the effect on accuracy is negligible for thresholds of 40 and over. Below 40, the superpixels are often poorly constrained due to their small size, and accuracy suffers. For an in-depth evaluation of different superpixel segmentation schemes and densities, we refer the reader to [37].

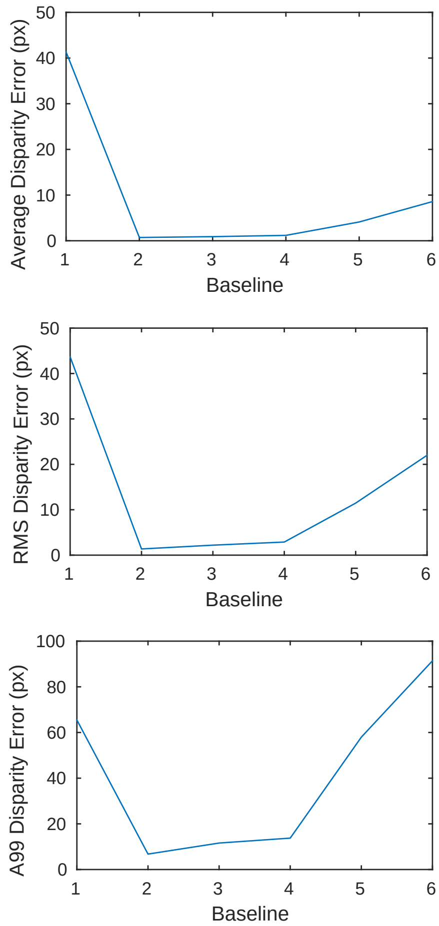



Figure 9: Performance against varying stereo baseline.

## 8. HCI stereo evaluation

The evaluation criteria used by the Middlebury and KITTI benchmarks have been the standard performance measures in the field for many years. However, there are many applications of stereo reconstruction for which these metrics are not suitable. For example robotic navigation applications are often unable to cope with catastrophic mistakes or “outliers” in the scene geometry. The fine grained accuracy of “inlier” regions is generally irrelevant in these situations. The field of Augmented Reality also has similar requirements. Outliers in the reconstruction can prove extremely disorienting for the user, while inaccuracies in fine details are generally not noticeable. This is a particularly important application area given the rapid development of specialist consumer-level hardware.

To address these concerns, there has been recent work by Honauer *et al.* [8] on alternative “geometry-aware” performance analysis for stereo reconstruction. This evaluation

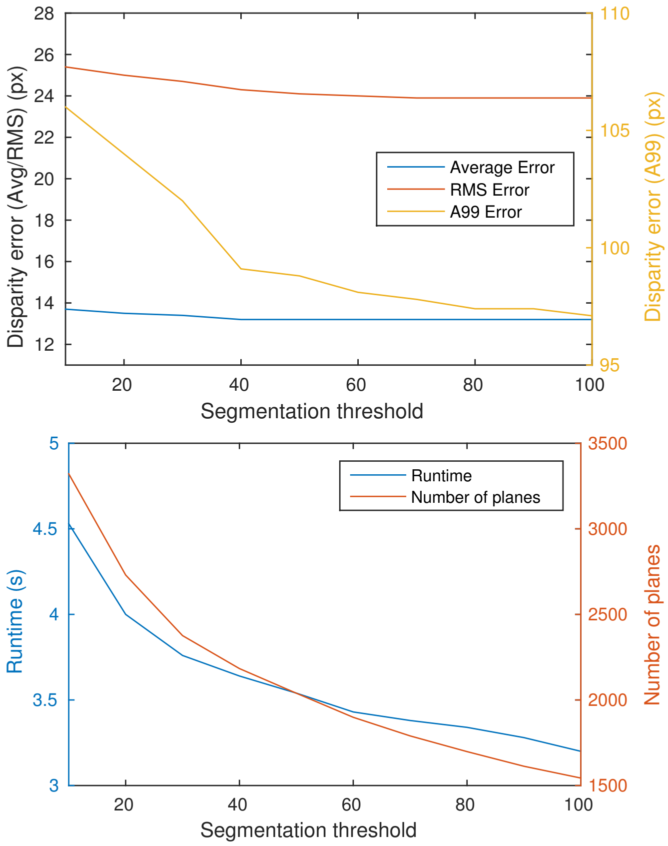



Figure 10: Behaviour of the approach with different segmentation thresholds. Top - Plots of the 3 accuracy characteristics. Bottom - plots of the tradeoff (speed and number of planes). Note that both subfigures display two Y scales.

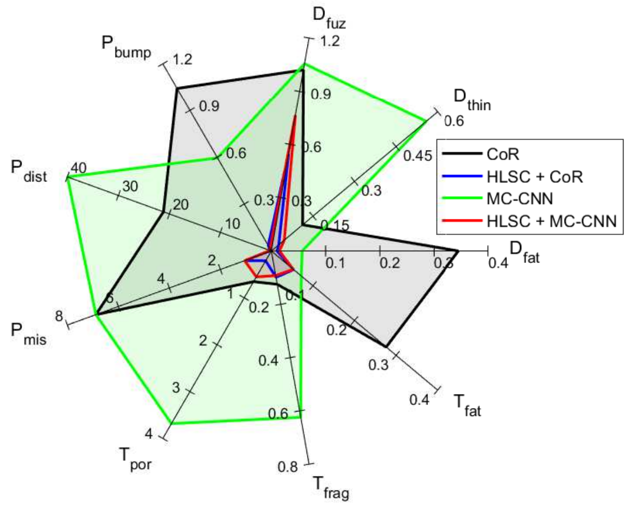



Figure 11: Evaluation of the proposed technique in terms of the HCI stereo metrics [8] averaged over the Middlebury 2014 dataset. For all errors, lower values indicate better performance.

protocol provides a better indication of how techniques perform in a wide range of different stereo applications. In particular, the HCI stereo evaluation focusses on three areas which are traditionally challenging for stereo reconstruction: Depth discontinuities, Planar surfaces and Thin structures. For each of these, three error metrics are proposed.

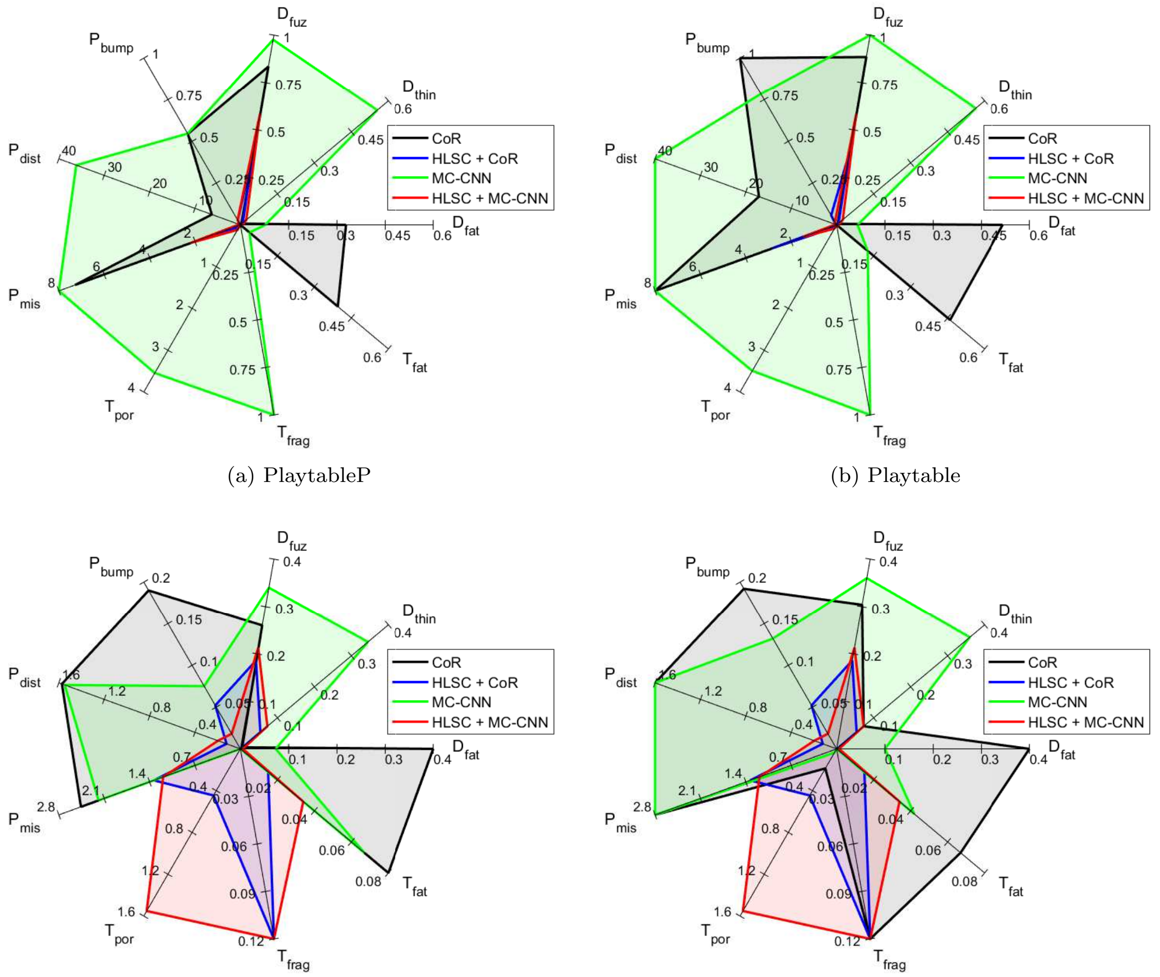
To evaluate depth discontinuities, the first metric is *fuzziness* ($D_{fuz}$) which measures the "sharpness" of edges (*i.e.* whether the depth map smoothly transitions from foreground to background due to oversmoothing). The remaining two performance measures are *foreground fattening* ($D_{fat}$) and *foreground thinning* ($D_{thin}$) which measure the algorithm's biases towards over or under-estimating the size of foreground regions.

The first two metrics for planar surfaces are similar. *Bumpiness* ($P_{bump}$) measures the deviation of the reconstructed surface from a perfectly smooth plane, ignoring errors in its position. *Offset* ($P_{dist}$) measures any bias in the systems positioning of planes (i.e. the accuracy of the location for planar surfaces). Finally *misorientation* ($P_{mis}$) measures how well the system is able to estimate the orientation of planar surfaces (ignoring errors in their position).

For evaluating thin structures, the first error measure is *detail fattening* ($T_{fat}$) similar to $D_{fat}$. The final two metrics look at the distribution of structure, in the case of detail thinning (where part of the thin structure is included in the background). *Porosity* ($T_{por}$) quantifies the amount of the thin structure which is not covered by any part of the estimate, penalising large gaps in the estimate. Finally *Fragmentation* ($T_{frag}$) measures the tendency for a single thin structure to be split into multiple separate structures. For further details of the performance measures (and illustrative examples) please see [8].

The sparse ground truth of the KITTI dataset makes it very difficult to extract thin structures or planar surfaces. As such, the geometry-aware evaluation could only be performed on the Middlebury 2014 training set. As suggested by the benchmark, performance is shown geometrically as a radar plot in Figure 11. For clarity, we focus the evaluation on the top two techniques from the previous section (CoR and MC-CNN), both run on the largest resolution they are capable of. Note that for all errors, lower numbers (i.e. closer to the centre of the radar) indicate better performance. It should be noted that 3 "pixel based" performance measures are also included in [8]. However, these are equivalent to the standard Middlebury/KITTI metrics, and for clarity are not duplicated here.

Clearly the use of top-down scene cues produces reconstructions with significant improvement across all the HCI metrics. The fuzziness of depth discontinuities is only slightly improved, and may be constrained by the quality of the superpixel segmentation step. Specifically, if the image is high resolution or the superpixels are particularly small, it is possible that depth discontinuities may be extracted as separate superpixels which connect the foreground and background. The quality of reconstructed planar surfaces is significantly improved. This is primarily due to the coplanarity cost $E_{cp}$ which combines multiple planar primitives




(a) PlaytableP (b) Playtable

(c) Piano (d) PianoL

Figure 12: Examination of challenging imaging scenarios on the HCI stereo metrics. Each subfigure shows the performance on a single scene from the Middlebury14 dataset.

into larger planar structures if deemed appropriate by the scene reasoning.

It is interesting to observe the relationship between the three *thin structure* metrics. CoR seems to have more problems with fattening of these structures (and fattening of depth discontinuities) but as a result it doesn't lose the structures and performs very well in terms of fragmentation and porosity. Conversely, MC-CNN underestimates or misses thin structures, leading to a very low $T_{fat}$ (and $D_{fat}$) score and very high $T_{por}$ and $T_{frag}$ (and $D_{thin}$) scores. The inclusion of scene understanding cues helps to balance this relationship, providing generally low scores for all three *thin structure metrics*.

Figure 12 looks at the detailed results on a number of scenes from the dataset, in order to explore how the HCI stereo metrics behave in challenging scenarios. The PlaytableP and Playtable scenes compare perfect and imperfect camera calibrations respectively. The Piano and PianoL scenes quantify the effect of consistent and inconsistent lighting (between the stereo pair). It is very interesting to note that these challenges have a very different effect on the HCI metrics than they do on the standard Middlebury metrics discussed under Table 3.

Inconsistent lighting was found to have a pronounced effect on the standard pixelwise metrics in Table 3 causing drops in performance of around 20 % across all techniques. However, in terms of the HCI metrics, errors introduced by lighting inconsistency are less pronounced. Indeed when using high level scene cues, the effect of lighting changes is almost imperceptible. Conversely, the quality of the camera calibration was found to have little impact on performance of the pixelwise metrics, but it has a significant impact on the HCI metrics. In particular, the quality of depth discontinuities is significantly improved when the proposed technique is given a perfect calibration.

## 9. Conclusions

These results demonstrate conclusively that in computer vision, as with the human vision system, understanding and top-down reasoning about the scene is a vital component for providing feasible and robust 3D reconstructions. The combination of bottom-up and top-down visual cues helps to significantly improve the robustness of obtained reconstructions. This is important as, in the case of uncertain scene geometry, a realistic failure (based on the types of scenes that occur in the real world) is preferable to a catastrophic failure with unrealistic artifacts. For example in robotics and autonomous system applications, a single catastrophic scene artifact may cause significant difficulties during path planning, whereas a more realistic failure would generally have little effect on the chosen path.

In addition we have shown that a joint framework, based on oriented planar primitives, is a highly effective representation to unify bottom-up and top-down reconstruction techniques. We have also provided examples of integrating a wide variety of information sources.

Although this approach increases robustness and prevents catastrophic failures, the use of planar primitives makes it impossible to accurately model curved surfaces. Approximating a smooth surface using a piece-wise linear function necessarily leads to residual errors. In a similar vein, the use of superpixels may lead to lower accuracy in areas of fine detail. The result is that the major gains in robustness may also cause a slight reduction in fine-grained accuracy. In the future it may be valuable to explore hybrid or multi-stage techniques to overcome this limitation.

Other possible future work includes the extension of the proposed cue weight learning (Section 6) to include an initial recognition of the environment type (indoor, outdoor urban, forest *etc.*). The weightings used could then be specialised for different environments. This idea could also be applied to the problem of temporal stereo reconstruction, where estimation of the optimal cue weightings could be performed online as the sequence progresses. It may also be interesting to investigate the incorporation of "recognition meets reconstruction" techniques into the proposed framework. These cues have proven very valuable for limited application domains in the literature, but no work has yet looked at use of the inter-relationships between recognised entities. However, without additional work it is not clear what the best way to integrate these cues would be.

*Acknowledgements*

The work presented in this paper was funded by the EPSRC project "Learning to Recognise Dynamic Visual Content from Broadcast Footage" (EP/I011811/1) and the Swiss National Science Foundation project "SMILE - Scalable Multimodal sign language Technology for sign language Learning and assessment". We are greatly indebted to Katrin Honauer for performing a significant portion of the evaluation for this work via the HCI stereo metrics [8]. We also wish to thank the authors of [63, 64, 62] for making their source code available, and particularly David Fouhey for his help regarding monocular scene understanding cues.

[1] R. M. Batson, Photogrammetry with surface-based images, Applied optics 8 (7) (1969) 1315–1322.

[2] B. Julesz, Binocular depth perception of computer-generated patterns, Bell System Technical Journal 39 (5) (1960) 1125–1162.

[3] P. J. Kellman, M. E. Arterberry, The cradle of knowledge: Development of perception in infancy, MIT press Cambridge, MA:, 1998.

[4] X. Wang, D. F. Fouhey, A. Gupta, Designing deep networks for surface normal estimation, in: Proc. CVPR, Boston, USA, 2015, pp. 539–547.

[5] D. F. Fouhey, A. Gupta, M. Hebert, Unfolding an indoor Origami world, in: Proc. ECCV, Zurich, Switzerland, 2014, pp. 687–702.

[6] J. E. Cutting, Perception of space and motion, Elsevier, 1995, Ch. The Integration, Relative Potency, and Contextual Use of Different Information about Depth, p. 69.

[7] S. Hadfield, R. Bowden, Exploiting high level scene cues in stereo reconstruction, in: Proc. ICCV, Santiago, Chile, 2015, pp. 783–791.

[8] K. Honauer, L. Maier-Hein, D. Kondermann, The hci stereo metrics: Geometry-aware performance analysis of stereo algorithms, in: Proc. ICCV, Santiago, Chile, 2015, pp. 2120–2128.

[9] D. G. Lowe, Distinctive image features from scale-invariant keypoints, IJCV 60 (2) (2004) 91 – 110.

[10] H. Bay, T. Tuytelaars, L. Van Gool, Surf: Speeded up robust features, in: Proc. ECCV, Vol. 3951, Graz, Austria, 2006, pp. 404–417.

[11] M. Humenberger, T. Engelke, W. Kubinger, A census-based stereo vision algorithm using modified semi-global matching and plane fitting to improve matching quality, in: CVPR Workshop, 2010, pp. 77–84.

[12] A. Geiger, M. Roser, R. Urtasun, Efficient large-scale stereo matching, in: Proc. ACCV, Queenstown, New Zealand, 2010, pp. 25–38.

[13] Z. Ma, K. He, Y. Wei, J. Sun, E. Wu, Constant time weighted median filtering for stereo matching and beyond, in: Proc. ICCV, Sydney, Australia, 2013, pp. 49–56.

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

[15] K. Lebeda, J. Matas, O. Chum, Fixing the locally optimized RANSAC, in: Proc. BMVC, Surrey, UK, 2012, pp. 1013–1023.

[16] H. Hirschmuller, Stereo processing by semiglobal matching and mutual information, PAMI 30 (2) (2008) 328–341.

[17] R. Spangenberg, T. Langner, R. Rojas, Weighted semi-global matching and center-symmetric census transform for robust driver assistance, in: Computer Analysis of Images and Patterns, 2013, pp. 34–41.

[18] S. Hermann, R. Klette, Iterative semi-global matching for robust driver assistance systems, in: Proc. ACCV, Daejeon, Korea, 2012, pp. 465–478.

[19] G. Schindler, F. Dellaert, Atlanta world: an expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments, in: Proc. CVPR, Vol. 1, Washington, DC, USA, 2004, pp. I–203–I–209 Vol.1.

[20] M. Schönbein, A. Geiger, Omnidirectional 3d reconstruction in augmented manhattan worlds, in: Proc. CVPR, Chicago, USA, 2014, pp. 716–723.

[21] K. Lebeda, S. Hadfield, J. Matas, R. Bowden, Texture-independent long-term tracking using virtual corners, Transactions on Image Processing 25 (2016) 359–371.

[22] C. Zach, T. Pock, H. Bischof, A duality based approach for realtime TV-L1 optical flow, in: Pattern Recognition, Springer, 2007, pp. 214–223.

[23] A. Kumar, S. Haker, C. Vogel, A. Tannenbaum, S. Zucker, Stereo disparity and l1 minimization, in: Conference on Decision and Control, 1997, pp. 1–8.

[24] F. Besse, C. Rother, A. Fitzgibbon, J. Kautz, Pmbp: Patchmatch belief propagation for correspondence field estimation, IJCV 110 (1) (2014) 2–13.

[25] P. Heise, S. Klose, B. Jensen, A. Knoll, Pm-Huber: Patchmatch with Huber regularization for stereo matching, in: Proc. ICCV,

Sydney, Australia, 2013, pp. 2360–2367.

[26] R. Ranftl, T. Pock, H. Bischof, Minimizing TGV-based variational models with non-convex data terms, in: Scale Space and Variational Methods in Computer Vision, Vol. 7893, Springer Berlin Heidelberg, 2013, pp. 282–293.

[27] G. Kuschk, D. Cremers, Fast and accurate large-scale stereo reconstruction using variational methods, in: ICCV workshop, 2013, pp. 700–707.

[28] A.-L. Chauve, P. Labatut, J.-P. Pons, Robust piecewise-planar 3d reconstruction and completion from large-scale unstructured point data, in: Proc. CVPR, San Francisco, CA, USA, 2010, pp. 1261–1268.

[29] D. Gallup, J.-M. Frahm, M. Pollefeys, Piecewise planar and non-planar stereo for urban scene reconstruction, in: Proc. CVPR, San Francisco, CA, USA, 2010, pp. 1418–1425.

[30] K. Yamaguchi, D. McAllester, R. Urtasun, Robust monocular epipolar flow estimation, in: CVPR13 (Ed.), Proc. CVPR, Oregon, USA, 2013, pp. 1862–1869.

[31] C. Zhang, Z. Li, R. Cai, H. Chao, Y. Rui, As-rigid-as-possible stereo under second order smoothness priors, in: Proc. ECCV, Zurich, Switzerland, 2014, pp. 112–126.

[32] K. Wu, M. Levine, Recovering parametric geons from multiview range data, in: Proc. CVPR, Seattle, WA, USA, 1994, pp. 159–166.

[33] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, S. Sinha, Object stereojoint stereo matching and object segmentation, in: Proc. CVPR, Colorado, USA, 2011, pp. 3081–3088.

[34] M. Bleyer, C. Rhemann, C. Rother, Extracting 3D scene-consistent object proposals and depth from stereo images, in: Proc. ECCV, Vol. 7576, Firenze, Italy, 2012, pp. 467–481.

[35] B. Mičušík, J. Košecká, Multi-view superpixel stereo in urban environments, International journal of computer vision 89 (1) (2010) 106–119.

[36] A. Bodis-Szomoru, H. Riemenschneider, L. Van Gool, Fast, approximate piecewise-planar modeling based on sparse structure-from-motion and superpixels, in: Proc. CVPR, Columbus, USA, 2014, pp. 469–476.

[37] A. Bódis-Szomorú, H. Riemenschneider, L. Van Gool, Superpixel meshes for fast edge-preserving surface reconstruction, in: Proc. CVPR, Boston, USA, 2015, pp. 2011–2020.

[38] D. F. Fouhey, A. Gupta, M. Hebert, Data-driven 3D primitives for single image understanding, in: Proc. ICCV, Sydney, Australia, 2013, pp. 3392–3399.

[39] A. Gupta, A. Efros, M. Hebert, Blocks world revisited: Image understanding using qualitative geometry and mechanics, in: Proc. ECCV, Vol. 6314, Heraklion, Crete, 2010, pp. 482–496.

[40] P. Denis, J. H. Elder, F. J. Estrada, Efficient edge-based methods for estimating manhattan frames in urban imagery, in: Proc. ECCV, Marseille, France, 2008, pp. 197–210.

[41] O. Barinova, V. Lempitsky, E. Tretiak, P. Kohli, Geometric image parsing in man-made environments, in: Proc. ECCV, Springer, Heraklion, Crete, 2010, pp. 57–70.

[42] R. Tal, J. H. Elder, An accurate method for line detection and manhattan frame estimation, in: ACCV 2012 Workshops, Springer, 2013, pp. 580–593.

[43] D. C. Lee, M. Hebert, T. Kanade, Geometric reasoning for single image structure recovery, in: Proc. CVPR, Miami, FL, USA, 2009, pp. 2136–2143.

[44] V. Hedau, D. Hoiem, D. Forsyth, Recovering the spatial layout of cluttered rooms, in: Proc. ICCV, Kyoto, Japan, 2009, pp. 1849–1856.

[45] V. Hedau, D. Hoiem, D. Forsyth, Thinking inside the box: Using appearance models and context based on room geometry, in: ECCV10 (Ed.), Proc. ECCV, Heraklion, Crete, 2010, pp. 224–237.

[46] C. Hane, C. Zach, A. Cohen, R. Angst, M. Pollefeys, Joint 3D scene reconstruction and class segmentation, in: Proc. CVPR, Oregon, USA, 2013, pp. 97–104.

[47] N. Cornelis, B. Leibe, K. Cornelis, L. Van Gool, 3D urban scene modeling integrating recognition and reconstruction, International Journal of Computer Vision 78 (2-3) (2008) 121–141.

[48] F. Guney, A. Geiger, Displets: Resolving stereo ambiguities using object knowledge, in: Proc. CVPR, Boston, USA, 2015, pp. 4165–4175.

[49] A. Saxena, J. Schulte, A. Y. Ng, Depth estimation using monocular and stereo cues., in: IJCAI, Vol. 7, 2007, pp. 1–8.

[50] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, L. Van Gool, Shape-from-recognition: Recognition enables meta-data transfer, Computer Vision and Image Understanding 113 (12) (2009) 1222–1234.

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

[52] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The KITTI vision benchmark suite, in: Proc. CVPR, Rhode Island, USA, 2012, pp. 3354–3361.

[53] R. Zabih, J. Woodfill, A non-parametric approach to visual correspondence, PAMI (1996) 1–10.

[54] P. Weinzaepfel, J. Revaud, Z. Harchaoui, C. Schmid, Deepflow: Large displacement optical flow with deep matching, in: Proc. ICCV, Sydney, Australia, 2013, pp. 1385–1392.

[55] H. Hirschmüller, D. Scharstein, Evaluation of cost functions for stereo matching, in: Proc. CVPR, Minneapolis, MN, USA, 2007, pp. 1–8.

[56] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University press, 2000.

[57] R. Koch, M. Pollefeys, L. Van Gool, Multi viewpoint stereo from uncalibrated video sequences, in: Proc. ECCV, Freiburg, Germany, 1998, pp. 55–71.

[58] P. K. Nathan Silberman, Derek Hoiem, R. Fergus, Indoor segmentation and support inference from rgbd images, in: Proc. ECCV, Firenze, Italy, 2012, pp. 746–760.

[59] A. Saxena, M. Sun, A. Ng, Make3D: Learning 3D scene structure from a single still image, PAMI 31 (5) (2009) 824–840.

[60] B. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: Proceedings of the 7th International Joint Conference on Artificial Intelligence, 1981, pp. 674–679.

[61] S. Hadfield, R. Bowden, Scene flow estimation using intelligent cost functions, in: Proc. BMVC, Nottingham, UK, 2014, pp. 1–8.

[62] A. Chakrabarti, Y. Xiong, S. J. Gortler, T. Zickler, Low-level vision by consensus in a spatial hierarchy of regions, in: Proc. CVPR, Boston, USA, 2015, pp. 4009–4017.

[63] J. Žbontar, Y. LeCun, Computing the stereo matching cost with a convolutional neural network, in: Proc. CVPR, Boston, USA, 2015, pp. 1592–1599.

[64] C. Zhang, Z. Li, Y. Cheng, R. Cai, H. Chao, Y. Rui, Meshstereo: A global stereo model with mesh alignment regularization for view interpolation, in: Proc. ICCV, Santiago, Chile, 2015, pp. 2057–2065.

[65] P. F. Felzenszwalb, D. P. Huttenlocher, Efficient graph-based image segmentation, IJCV 59 (2) (2004) 167–181.

**Simon Hadfield** received his PhD from the University of Surrey in 2013 where he also received an MEng (distinction) in Electronic and Computer Engineering. He has been awarded the DTI MEng Prize, for the top performance by an MEng graduate, and the Associateship of the University of Surrey (AUS) for his industrial placement with the Home Office Scientific Development Branch. He is an RA2 at the Centre for Vision Speech and Signal Processing (USurrey).

**Karel Lebeda** received his BSc and MSc (both with honours) from Czech Technical University in Prague in 2010 and 2013. Currently, he is pursuing a PhD degree at the Centre for Vision Speech and Signal Processing at the University of Surrey. His research interests include visual object tracking and 3D computer vision.

**Richard Bowden** received a BSc and MSc from the Universities of London and Leeds and a PhD from Brunel University which was awarded the Sullivan Doctoral Thesis Prize. He is Professor of computer vision and machine learning at the University of Surrey leading the Cognitive Vision Group within the Centre for Vision Speech and Signal Processing and was awarded a Royal Society Leverhulme Trust Senior Research Fellowship. His research centres on the use of computer vision to locate, track, and understand humans. He is an associate editor for the journals IVC and TPAMI.