Incremental Dense Multi-modal 3D Scene Reconstruction

Ondrej Miksik Yousef Amar Vibhav Vineet Patrick Pérez Philip H. S. Torr

***Abstract*— Aquiring reliable depth maps is an essential prerequisite for accurate and incremental 3D reconstruction used in a variety of robotics applications. Depth maps produced by affordable Kinect-like cameras have become a de-facto standard for indoor reconstruction and the driving force behind the success of many algorithms. However, Kinect-like cameras are less effective outdoors where one should rely on other sensors. Often, we use a combination of a stereo camera and lidar, however, process the acquired data in independent pipelines which generally leads to sub-optimal performance since both sensors suffer from different drawbacks. In this paper, we propose a probabilistic model that efficiently exploits complementarity between different depth-sensing modalities for incremental dense scene reconstruction. Our model uses a piecewise planarity prior assumption which is common in both the indoor and outdoor scenes. We demonstrate the effectiveness of our approach on the KITTI dataset, and provide qualitative and quantitative results showing high-quality dense reconstruction of a number of scenes.**

I. Introduction

Aquiring reliable depth maps is an essential prerequisite for accurate and incremental 3D reconstruction used in a variety of robotics applications, including navigation [1], [2], object recognition [3], [4], wearable and/or assistive technology [5], and grasping [6]. Depth maps produced by affordable Kinect-like cameras have become a de-facto standard for indoor perception [7], [8] and the driving force behind the success of many algorithms. However, Kinect-like cameras are less effective outdoors where one should rely on other sensors. With the advent of an increasingly wide selection of sensing modalities (*e.g.* 2D/3D laser range finders, optical cameras, stereo/depth cameras, flash ladars, radars, etc.), it is now common to obtain multiple observations of a given scene; a typical example are sensors mounted on (un)manned vehicles [1]. Using observations from different modalities is generally advantageous as they are complementary but at the same time challenging since there often is no one-to-one correspondence across modalities.

Let us consider, for instance, an optical camera and a lidar, as illustrated in Fig. 1. The camera has a limited dynamic range (Fig. 1, top) and many parts of perceived scene can easily be saturated (specular highlights, reflections, overexposure, …). Stereo matching algorithms generally fail in such areas where they are unable to predict any meaningful depth, resulting in large holes in the dense depth maps (Fig. 1, 3rd row). This reconstruction problem is ill-posed even for images without any illumination artifacts due to ambiguity in dense correspondence matching (textureless areas, repetitive patterns, …) and performance is usually determined by a trade-off between accuracy and efficiency. Fast algorithms typically use only (non-regularized) per-pixel predictions with heuristic postprocessing reducing noise [15], while accurate but slow methods rely on (semi)global optimization enforcing smoothness and ordering constraints [10], [11]. Moreover, most algorithms operate on a per-frame basis, which reduces their efficiency and temporal consistency. In contrast, lidars (Fig. 1, 2nd row) are often able to sense in areas in which RGB video information is not exploitable and provide more accurate/reliable measurements. However, lidars often have smaller field-of-view than cameras, depth readings are limited to a certain maximum range and are obtained at much slower temporal rate (except with most expensive systems, which are not suitable for many applications).

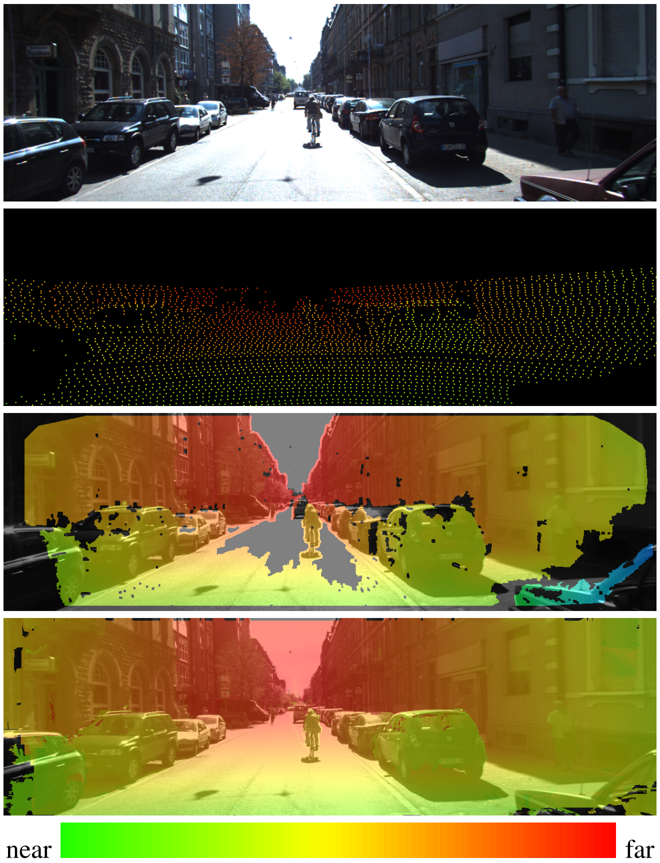

Fig. 1. One view of stereo camera with multiple reflections, specularities and over-exposed zones (top), 3D point cloud captured by a Velodyne HDL-64E laser scanner (2nd row), stereo reconstruction [9] (3rd row), and output from our system (bottom), as seen from a moving platform on-the-fly.

O. Miksik, Y. Amar and P. Torr are with the University of Oxford, UK.
V. Vineet is with Stanford, California, US.
P. Pérez is with Technicolor R&I, Cesson Sévigné, FR.
{ondra.miksik, yousefamar, vibhav.vineet}@gmail.com

This research was supported by Technicolor, EPSRC, the Leverhulme Trust and the ERC grant ERC-2012-AdG 321162-HELIOS.

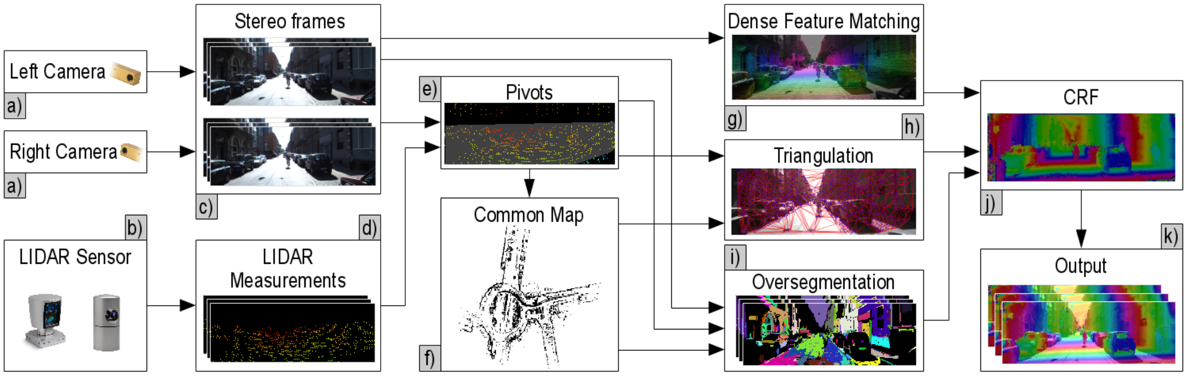

Fig. 2. Overview of our system: (a) given a pair of cameras and (b) lidar, we (c) capture stereo images and (d) 3D point cloud, (e) generate an initial set of pivots and (f) project them on a common map. Given the pivots within the current frustum and stereo images, we evaluate (g) unary potential and piecewise planar term based on (h) the Delaunay triangulation of pivots and (i) oversegmentation over which we (j) define a pairwise CRF to (k) infer the final solution.

Processing data from different modalities in independent pipelines generally leads to sub-optimal performance. In this paper, we propose a model that efficiently exploits complementarity between different depth-sensing modalities for incremental dense scene reconstruction. For ease of exposition, we demonstrate our method on stereo camera and lidar measurements, however the method is general and can accomodate other sensors (*e.g.* radar for obstacle detection, etc.). We directly integrate the lidar data into stereo reconstruction algorithm to predict accurate depth maps and we show that the superior results can be obtained even with second-class, cheap sensors (Fig. 1, bottom).

At the core of our system is a pairwise conditional random field (CRF) that captures interactions between the pixels and efficiently combines information from the stereo camera and lidar (Fig. 2 (k)). To this end, we assume that the sparse yet trustworthy lidar 3D measurements and 3D points generated by robustly matched and triangulated sparse 2D keypoints (Fig. 2 (e)), that we call *pivots*, are accurate enough to provide partial prior knowledge about the scene. To exploit this prior knowledge in our model, we drastically reduce the unary costs attached to these points, so that the optimal depth assignments are attracted towards pivots' depth, and pivots *guide* dense matching. Our unary potentials (Fig. 2 (g)) are based on dense matching of 2D features along the epipolar lines and a piecewise-planar prior defined by various groupings of pivots (Fig. 2 (h, i)). Such prior typically models only small scene fragments and/or does not respect object boundaries well [9]. Thus, we define the groupings of pivots over a multiscale hierarchy of oversegmented regions that provide knowledge about potential object boundaries and model planarity over larger surfaces such as the whole road region or table top. Pivots also help to disambiguate dense matching by constraining the searched range which results in more confident unary predictions and their faster evaluation. Our pairwise terms propagate information into uncertain (*e.g.* saturated) areas and enforce smoothness among the neighbouring pixels (including the lidar data). Note that we do not introduce any hard constraints forcing variables at pivots' coordinates to take the estimated depth, hence, this leaves the chance for recovery if the pivot is assigned incorrect measurement.

Further, we project the pivots on a common map (Fig. 2 (f)) to maintain the temporal consistency and not to discard any measurements. Hence all measured data are available to the algorithm on request (and not just the latest sensor readings). To maintain the computational and memory complexity, we use a sparse hash-table-driven data structures that ignore unoccupied space and swap/stream map data between device and host memories as needed to fit the data into GPU memory and process only the data within a current frustum (as in [12]).

In order to efficiently infer the approximate maximum posterior marginal (MPM) solution [13], we use a mean-field inference technique that refines the marginals of a node with a bilateral filter that is suitable for parallel implementation. This allows us to run inference each frame (as only a few mean-field update iterations are required), which is of utmost importance in most of the robotics settings where output is required at real-time or interactive rates. The system outputs a per-pixel probability distribution instead of a single label, which is desirable in robotics as it allows probabilistic interpretation in other subsystems. All parts of our system are trivially parallelizable, hence suitable for GPU implementation.

It should be noted, that our approach is not specific to this application, can be used with multiple sensors and/or other modalities, naturally accomodates other priors and can be extended to handle other tasks such as semantic and/or motion segmentation, etc.

II. Related work

Dense deth map estimation from stereo images is one of the most studied problems in computer vision [14]. In general, fast methods usually treat each pixel independently, capture context only in a very small area and smoothness is often achieved through heuristical postprocessing [15]. Algorithms that rely on (semi)global optimization capture the structure [9], [16], encode higher order constraints (*e.g.* to model slanted planes) [10] and use segmentation [11], [17], [18]. However, these methods do not exploit complementarity and partial knowledge about the scene obtained from *different* modalities. Torr and Criminisi [19] proposed pivoted dynamic programming in rectified image pairs, which attempts to attract the optimal disparity path along a scanline towards the prior disparity at matched keypoints.

All the above-mentioned methods process data on per-frame basis resulting in temporally inconsistent prediction. It has been shown, that maintaining pivots across the video sequence improves temporal consistency of stereo algorithms [5], however their approach assumes a user in the loop (works only with laser path forming a hull).

Other approaches focus on inpainting [20], [21] of the Kinect depth maps, however, they require fairly dense depth from active sensors. Diebel and Thrun [22] proposed an MRF model for upsampling of laser measurements with enforced smoothness across areas with constant color. Though their method uses both the laser and color data, they do not use planarity prior as we do in proposed method. Further, they do not consider motion while processing depth data. As we show in the experiment section, both these two techniques are necessary for high accuracy and efficiency. Dolson *et al.* [23] proposed filtering framework for dynamic scenes. Badino *et al.* [24] showed how to integrate sparse lidar measurement directly into disparity estimation. Though this method also tries to solve problem similar to our there are some key differences which are necessary for achieving high accuracy efficiently. First we propose to use a region based planar prior which is necessary to model planarity over large regions such as road, table-top etc. Further we solve the energy minimization problem in a mean-field framework which is naturally parallizable compared to the dynamic programming based method of Badino *et al.* [24].

On application side, Munoz *et al.* [25] proposed 2D-3D (camera-lidar) co-inference for semantic segmentation. Held *et al.* [26] combined lidar and optical camera for object tracking, Premebida *et al.* [27] combined the same modalities for pedestrian detection and Arnab *et al.* [28] combined audio-visual cues for semantic segmentation.

III. Dense Multi-modal Depth-Map Estimation

Our system exploits partial prior knowledge about the scene provided by relatively sparse but accurate 3D measurements, called *pivots*. Hence, the first step is to project the lidar measurements into the camera coordinate system. Since lidars often have smaller field-of-view than cameras, we augment these points by robustly matched keypoints. Next, we use dense matching, and piecewise planar prior to evaluate the CRF potentials and run the inference. The following subsections assume synchronized data and process them per-frame. We relax this assumption in III-G.

A. Setting the stage

In our setup, we assume that all sensors are calibrated. In case of cameras, this comprises of: 1) intrinsic camera calibration to compute the geometric parameters of each camera lens (focal length, principal point, radial and tangential distortion); 2) stereo calibration to compute the geometric relationship between the two cameras, expressed as a rotation matrix and a translation vector; 3) stereo rectification to correct the camera image planes such that they are scanline-aligned and disparity computation is simplified. Without loss of generality, the reference camera coordinate system has origin in the top-left corner of the left camera (more details in [29]).

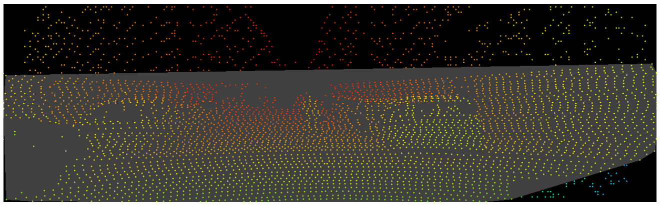

Fig. 3. Pivots – gray area contains lidar measurements, outside this region we perform sparse feature matching.

The laser scanner is registered with respect to the reference camera coordinate system. In this section, we also assume the cameras and laser scanner are synchronized and data are "untwisted" in case of spinning lidars. The optimization is carried out in the disparity image space; with conversion to depth $z_i = bf/d_i$ with baseline b, focal lenght f and disparity at i-th pixel d_i.

B. Pivots

In order to disambiguate dense matching, we first define a set $\mathcal{P}$ of confident 3D points capturing partial prior knowledge about the scene, so called *pivots*. Each pivot $p = \{x_p, y_p, d_p\}$ is represented by coordinates $\{x_p, y_p\} \in \mathbb{N}^2$ and disparity $d_p \in \mathbb{N}$ defining the displacement of the corresponding matching point along the epipolar line in the right image.

In our case, a first natural choice of pivots is the set $\mathcal{P}_0$ consisting of all lidar measurements projected into the image plane as we assume relatively high precision of laser scanning. However, it is often the case that lidars have smaller field-of-view than cameras (Fig. 1 (b)), do not return any measurement on areas subject to reflections or located past the maximum range limit. Hence, we augment the initial set of pivots by a set of robustly matched keypoints $\mathcal{K}$, *i.e.* $\mathcal{P} = \mathcal{P}_0 \cup \mathcal{K}$. See section IV-A for implementation details.

C. Model

We define a random field over random variables $\mathcal{X} = \{X_1, ..., X_N\}$, conditioned on data $\mathcal{I} = \{\mathbf{I}^{(l)}, \mathbf{I}^{(r)}, \mathcal{P}\}$ consisting of a pair of 2D images $\mathbf{I}^{(l)}, \mathbf{I}^{(r)}$ and pivots $\mathcal{P}$. We assume that each discrete random variable X_i is associated with a pixel $i \in \mathcal{N} = \{1...N\}$ in the image of the reference camera (left) and takes a label $d_i \in \mathbb{N}$ from an ordered finite disparity label set $\mathcal{D} = \{d_1, ..., d_D, d_{D+1}\}$. A dummy label d_{D+1} with some constant cost is added to indicate invalid depth (outliers/occlusions). We formulate the problem of assigning disparity labels to the pixels as one of solving a densely-connected, pairwise Conditional Random Field

(CRF)

$$\begin{aligned} P(\mathbf{X}|\mathcal{I}) &= \frac{1}{Z(\mathcal{I})}\exp(-E(\mathbf{X}|\mathcal{I})) \\ E(\mathbf{X}|\mathcal{I}) &= \sum_{i\in\mathcal{N}}\psi_u(X_i) + \sum_{i<j\in\mathcal{N}}\psi_p(X_i,X_j), \end{aligned} \tag{1}$$

in which $E(\mathbf{X}|\mathcal{I})$ is the energy associated with a configuration $\mathbf{X} = (X_1 \dots X_N)$, conditioned on the data $\mathcal{I}$, $Z(\mathcal{I}) = \sum_{\mathbf{X}'}\exp(-E(\mathbf{X}'|\mathcal{I}))$ is the (data-dependent) partition function and $\psi_u(\cdot)$ and $\psi_p(\cdot,\cdot)$ are the unary potential and pairwise potential functions, respectively, both implicitly conditioned on the data $\mathcal{I}$. This model is not constrained to our particular application and can be extended to, *e.g.* joint depth prediction and semantic or motion segmentation, etc.

D. Unary potential

Our unary potential function is inspired by guided dense stereo matching proposed by Torr and Criminisi [19] and large-scale stereo estimation algorithm of Geiger *et al.* [9] and consists of two terms, (1) feature matching and (2) piecewise-planar term that we included directly into the unary potential function.

Let $\mathbf{f}_i \in \mathbb{R}^R$ be an image dependent feature vector (pixel intensity or patch descriptor) for pixel $i = \{x_i, y_i\} \in \mathbb{N}^2$ and superscripts $^{(l)}$,$^{(r)}$ denote left and right images, respectively.

1) Feature Matching: We express the contribution of the data term as a constrained Laplace distribution capturing cost for 1D dense feature matching along the epipolar line

$$\psi^d(\cdot|d_i,\mathcal{F}) \propto \begin{cases} \exp\left(-\beta\|\mathbf{f}_i^{(l)} - \mathbf{f}_{i-d_i}^{(r)}\|_1\right), & \forall d_i \in \bar{\mathcal{D}}_i \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\bar{\mathcal{D}}_i \subseteq \mathcal{D}$ (defined below) is a subset of disparity levels that for each pixel i reduces the searched range and implicitly introduces the epipolar constraint $y_i^{(l)} = y_i^{(r)}$.

2) Piecewise-planar term: We define our prior exploiting partial knowledge about scene provided by pivots $\mathcal{P}$ to be proportional to a sampled Gaussian

$$\psi^p(\cdot|d_i,\mathcal{P}) \propto \begin{cases} \exp\left(-\frac{[d_i-\mu(\tau_t,i)]^2)}{2\sigma^2}\right), & \text{if } d_i \in \bar{\mathcal{D}}_i \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where σ are constants set by cross-validation determining our belief into plane τ defined by lidar measurements (σ_l), robustly matched keypoints (σ_k) or both (σ_{lk}), respectively, and $\bar{\mathcal{D}}_i = \{|d_i - \mu(\cdot)| < 3\sigma \vee d_i \in N_P\}$ is a subset of disparity levels for which the equation is evaluated. We evaluate only disparities within 3σ from the mean to gain speed. The condition $d_i \in N_P$ enables the prior to locally extend its range to better handle disparity discontinuities in places where the linearity assumption might be violated. We define $\mu(\cdot)$ to be a piecewise linear function

$$\mu(\tau_t, f_i^{(l)}) = a_t x_i + b_t y_i + c_t \tag{4}$$

interpolating the subsets $\mathcal{T} = \{\tau_1, ..., \tau_T\}$ of pivots in $\mathcal{P}$. We use two types of partitioning of pivots set $\mathcal{P}$:

Delaunay triangulation: it partitions set of pivots $\mathcal{P}$ into a set of non-overlapping triangles $\mathcal{T}_D \subseteq \mathcal{T}$, *i.e.* $\cup_{t\in\mathcal{T}_D}\tau_t = \mathcal{P}$, hence captures a coarse estimate of a 3D structure. For each triangle τ_t, we form a linear system of equations and solve for the plane parameters $\{a_t, b_t, c_t\}$ by SVD. Hence the mode μ of the proposed distribution is a linear combination of the pivots in triangle τ_t.

Oversegmentation: The Delaunay triangulation partitions pivots into non-overlapping triangles, however, such triangles may cover multiple objects. Also, if the pivots are imprecise (often happens with real-world measurements), such prior may result into *e.g.* non-coplanar neighbouring planes on a flat surface. We overcome both issues by partitioning pivots $\mathcal{P}$ into sets $\mathcal{T}_O \subseteq \mathcal{T}$ defined by object-aware segments – these are often sensitive to potential object boundaries and often contain many pivots, hence "regularize" priors defined by non-overlapping triangles.

A natural question is how to define grouping of image pixels. In contrast to object recognition, even if we had a method that could perfectly segment the objects from each other, it would not be enough for disparity estimation, since a single object often consists of many shapes/parts. Hence we use a multi-scale over-segmentation (details in sec. IV-A) to define such regions and RANSAC with least squares refinement to robustly fit a plane (*i.e.* estimate $\{a_t, b_t, c_t\}$) into a subset of pivots associated with each segment $\tau_t \in \mathcal{T}_O$.

3) Unary potential function: Combining feature matching term (Eq. 2) and piecewise-planar term (Eq. 3) together, taking the negative logarithm and introducing a "discount" function Ω for pivots yield

$$\psi_u(\cdot) = \Omega_i\left[\beta\|\mathbf{f}_i^{(l)} - \mathbf{f}_{i-d_i}^{(r)}\|_1 + \sum_{\tau_t\subseteq\mathbb{I}[i\in\mathcal{T}]}\frac{[d_i - \mu(\tau_t,i)]^2}{2\sigma^2}\right] \tag{5}$$

where $\mathbb{I}[\cdot]$ is an indicator function returning all subsets $\tau_t \subseteq \mathcal{T}$ that contain pixel $\{x_i, y_i\}$, and discount function

$$\Omega_i = \begin{cases} \omega, & \text{if } \{x_i, y_i, d_i\} = p \in \mathcal{P} \\ 1, & \text{otherwise} \end{cases} \tag{6}$$

drastically reduces the cost of configurations assigning measured depth at pivots $p \in \mathcal{P}$ by some constant ω. Using different constants ω for pivots obtained by lidar (ω_l) and robust keypoint matching (ω_k) allows us to model our belief into precision of these measurements.

Note, that we do not introduce any hard constraints forcing variables at pivots' coordinates to take the measured disparity, hence, this leaves the chance for recovery if pivot has assigned incorrect disparity. Also, the piecewise planar term can be replaced by a set of functions with Minimum Description Lenght (MDL) prior to better model non-planar surfaces such as conics, etc.

E. Pairwise potentials

The pairwise potential function $\psi_p(\cdot,\cdot)$ enforces consistency over pairs of random variables and thus generally leads to a smooth output. In our application, we use a weighted mixture

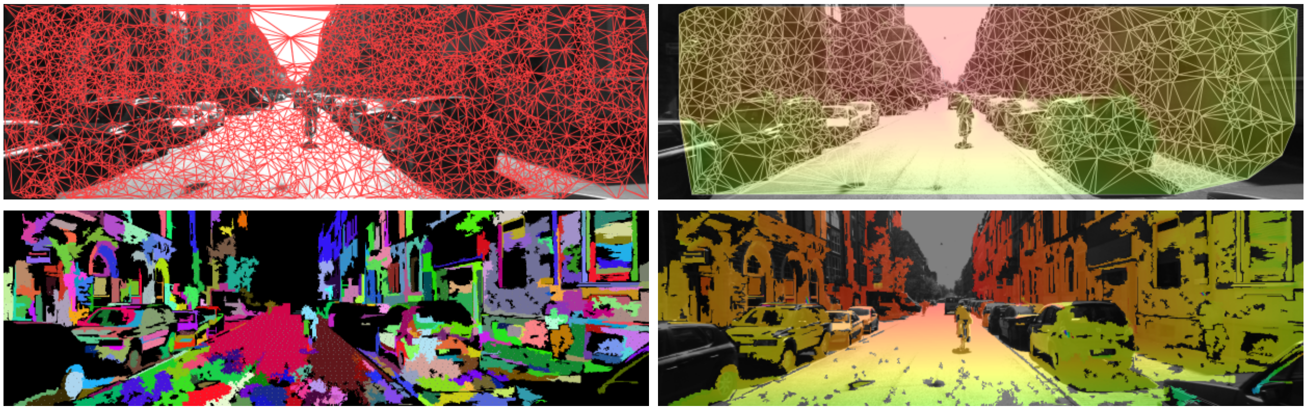

Fig. 4. Piecewise planar prior defined on regions obtained with Delaunay triangulation (top) and multiscale over-segmentation, here we show the 3rd level (bottom).

of Gaussian kernels (with unit covariance matrix) that depend on appearance features

$$\psi_{ij}(d,d') = \Delta(d,d') \sum_{m=1}^{M} w^{(m)} k^{(m)}(\bar{\mathbf{f}}_i^{(m)}, \bar{\mathbf{f}}_j^{(m)}) \tag{7}$$

where weights associated with m-th kernel $w^{(m)}$ are obtained by cross-validation, $\bar{\mathbf{f}}_i^{(m)}, \bar{\mathbf{f}}_j^{(m)}$ are the 2D features extracted from image data $\mathbf{I}^{(l)}$ at the i^{th} and j^{th} pixels (respectively) and $\Delta(d,d')$ is the compatibility function. We use a combination of the Gaussian kernel

$$k^{(1)}(\bar{\mathbf{f}}_i^{(1)}, \bar{\mathbf{f}}_j^{(1)}) = w^{(1)} \exp\left(-\frac{\|\mathbf{c}_i - \mathbf{c}_j\|_2^2}{2\theta_\gamma}\right) \tag{8}$$

removing small isolated areas and bilateral kernel

$$k^{(2)}(\bar{\mathbf{f}}_i^{(2)}, \bar{\mathbf{f}}_j^{(2)}) = w^{(2)} \exp\left(-\frac{\|\mathbf{c}_i - \mathbf{c}_j\|_2^2}{2\theta_\alpha} - \frac{\|\mathbf{I}_i^{(l)} - \mathbf{I}_j^{(l)}\|_2^2}{2\theta_\beta}\right) \tag{9}$$

enforcing neighbouring pixels with similar appearance to take the same label. Parameters θ controls the spatial extent of the kernel, $\mathbf{c} = \{x_i, y_i\}$ are pixel cordinates and $\mathbf{I}$ is color. This form of potential introduces a small fronto-parallel bias, which can be overcome by higher-order potentials. We decided not to use higher-orders, as it would make the inference slower; instead we directly included the "slanted" areas prior directly into our unary potentials. We use the standard truncated L_1 or L_2 compatibility functions

$$\Delta(d,d') = \min(\|d - d'\|_\Gamma, \alpha) \tag{10}$$

where $\|\cdot\|_\Gamma$ is the L_1 or L_2 norm, respectively, and α is the clipping parameter.

F. *Efficient inference*

One of the most popular approaches for multi-label CRF inference has been graph-cuts based α-expansion [30], which finds the maximum a posteriori (MAP) solution. However, graph-cuts leads to slow inference and is not easily parallelisable. Given the form of the energy function defined above, we follow the mean-field based optimization method, that has been shown to be very efficient for pairwise CRFs in 2D image segmentation [13].

In the mean-field framework, we approximate the true distribution $P(\mathbf{X})$ by a family of $Q(\mathbf{X})$ distributions that factorize as the product of all components' marginals (components are independent) $Q(\mathbf{X}) = \prod_i Q_i(X_i)$. The mean-field inference then attempts to minimize the KL-divergence $D_{\mathrm{KL}}(Q||P)$ between the tractable distribution Q and true distribution P. Under this assumption, the fixed point solution of the KL-divergence, under constraint that $Q(\mathbf{X})$ and $Q(X_i)$ are valid distributions, leads to the iterative mean-field update (refer to [31] for more details):

$$\begin{aligned} Q_i(X_i = d) =& \frac{1}{Z_i} \exp\{-\psi_u(X_i) - \\ & \sum_{d' \in \bar{\mathcal{D}}_i} \sum_{j \neq i} Q_j(X_j = d')\psi_p(X_i, X_j)\} \end{aligned} \tag{11}$$

where $Z_i = \sum_{X_i = d \in \bar{\mathcal{D}}_i} \exp\{-\psi_u(X_i) - \sum_{d' \in \bar{\mathcal{D}}_i} \sum_{j \neq i} Q_j(X_j = d')\psi_p(X_i, X_j)\}$ is a constant normalizing the marginal at pixel i. The complexity of the mean-field update is $\mathcal{O}(N^2)$.

It has been shown, that the time-consuming pairwise update of densely connected CRFs can be efficiently approximated by a filter-based variant. Such approach is particularly attractive for tasks with a small number of labels and a constant label over large areas as it allows to capture long-range interactions (*e.g.* object segmentation). However, for disparity estimation, we often have large state space and neighbouring pixels tend to take different labels (typically slanted areas). Hence we further exploit partial prior knowledge about the scene, and evaluate the pairwise updates only for labels within the range defined by prior (*i.e.* states with evaluated unary potential) plus some small slack λ_s (*e.g.* 5 disparity labels) allowing to handle imprecise pivots, *i.e.* $d' \in (\mathcal{D} \cup \lambda_s)$. The algorithm is inherently parallel, runs for a fixed number of iterations, and the MPM solution is extracted by choosing $x_i \in \operatorname{argmax}_d Q_i(x_i = d)$ from soft predictions at the final iteration.

G. *Temporal Sequences of Images*

Often, robotic platforms perceive a gradually changing scene with multiple sensors operating at different rates (*e.g.* cameras at 25Hz, lidar at 15Hz). So far, our system has required synchronized sensors and processed only the latest batch of data. Discarding all previous measurements results into temporaly inconsistent predictions (even for static scenes due to noise) and need for all sensors to operate at rate of the slowest sensor.

It has been shown, that providing pivots consistent over the temporal sequence stabilizes the predicted disparity [5]. To this end, we replace per-frame keypoint matching by more robust temporal matching, *i.e.* the per-frame robustly matched features are propagated over time with mutual exclusive check, and project both, the lidar readings and matched keypoints on a common map. Consequently, all the measurements are available to the algorithm on a request and we do not discard any. The only assumption is, that the 6DoF pose is available (can be obtained with IMU/VO). Our map is represented by a sparse hash-table-driven data structure that ignores unoccupied space. Further, we swap/stream map data between device and host memories as needed to fit the data into GPU memory and process only the data within a current frustum [12], [2]. This results into more stable set of keypoints over the temporal sequence of images.

IV. Experiments

A. Implementation details

In this section, we provide implementation details of our approach. Pivots from different modalities can be defined and modeled in numerous ways. Our implementation relies on a simple yet reasonable assumption that lidar measurements are generally more accurate than feature matching. Hence, we perform sparse feature matching only in areas that are not covered by lidar measurements (such areas are discovered by simple dilation of lidar measurements). Though a variety of fast feature detectors and descriptors has been proposed [32], we follow [9] who showed that matching of the points sampled on a regular grid using the L_1 distance between the descriptors consisting of concatenated horizontal and vertical Sobel responses is both, fast and stable. To impose no restrictions on the disparities, we allow a large disparity 1D search range along the epipolar line. Non-stable keypoints are eliminated by mutual exclusive check [33] and the best to the second best match ratio. We also remove all keypoints which exhibit disparity values dissimilar from all surrounding support points. For videos, we use the Fovis visual odometry library [34] to estimate 6DoF pose and per-frame feature matching (for pivots) is replaced by features tracked by Fovis to increase temporal consistency and robustness.

In principle, our framework can be used with any superpixel grouping algorithm (k-means, mean-shift, slic, …). Our implementation uses multi-scale (4 levels) oversegmentation by Felzenswalb and Huttenlocker [35] since it is fast and it is easy to control size of segments.

B. Dataset and baselines

We demonstrate the effectiveness of our reconstruction from different modalities for both per-frame and video sequences. We evaluate our system on the KITTI dataset [36], which contains a variety of outdoor sequences, including a city, road and campus. All sequences were captured at a resolution of 1241 × 376 pixels using stereo cameras (with baseline 0.54m) mounted on the roof of a car. The cameras were calibrated and captured images rectified. The car was also equipped with a spinning Velodyne HDL-64E laser scanner (LIDAR). All sensors were synchronized, the dataset was captured at 10Hz and cameras triggered when lidar was rotated forward.

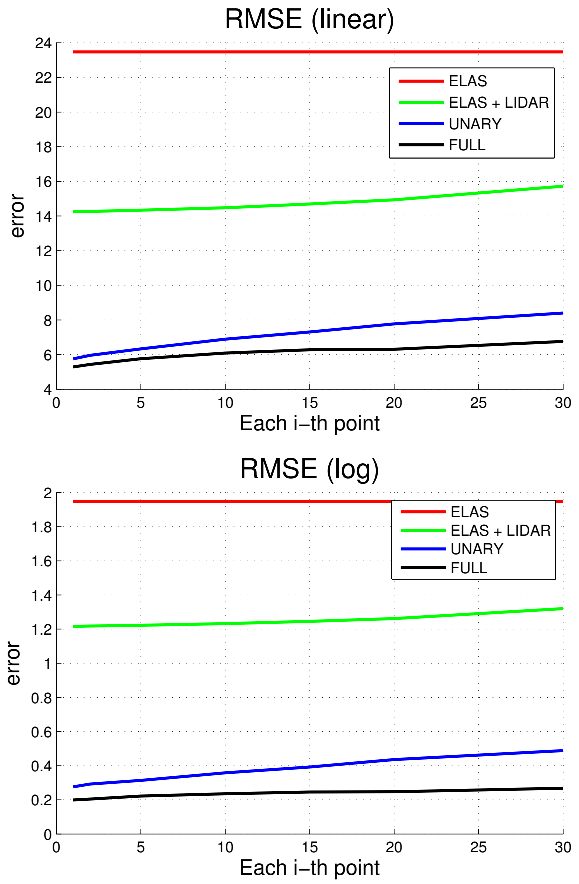

Fig. 5. Quantitative results: RMSE linear (top), RMSE log (bottom). See text for details.

The KITTI dataset is very challenging since it contains numerous changes in lighting conditions resulting in textureless areas, repetitive patterns (road, facades, …), etc. We report both, qualitative and quantitative results evaluated on sequences exhibiting above-mentioned challenging conditions and show substantial improvement with respect to our baselines. The first baseline is the disparity matching algorithm (from passive stereo cameras) by Geiger *et al.* [9] since part of our unary potentials follow this approach. Obviously, comparison with respect to the algorithm relying purely on data from cameras is not fair as this baseline use less data. Hence, the second baseline is a modified version that uses exactly the same set of support points as our approach.

C. Qualitative results

First, we show some qualitative results for our algorithm. In Fig. 6, we highlight the ability of our approach not only to estimate disparity in saturated zones (*e.g.* filled

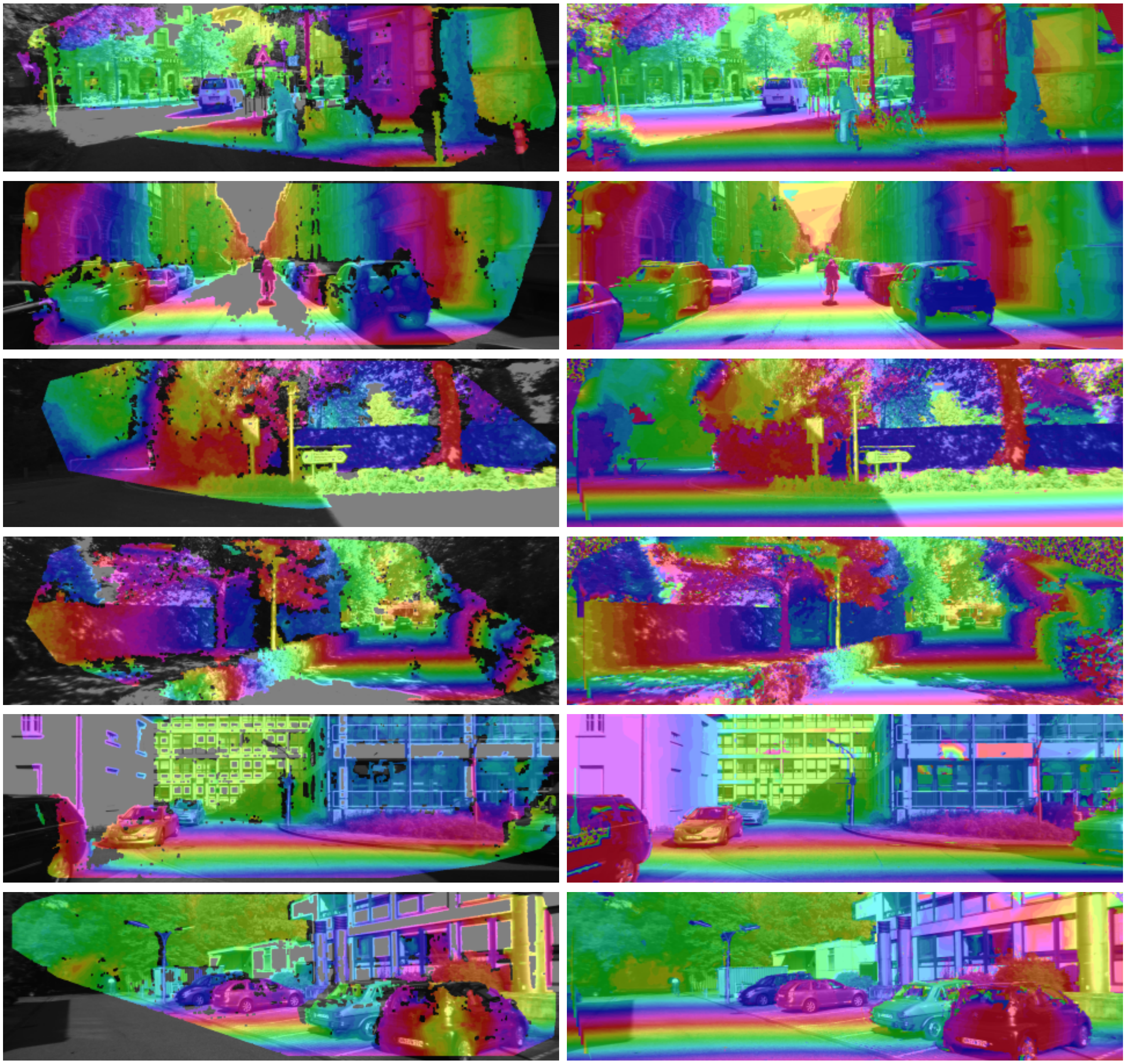

Fig. 6. Qualitative results - left: Geiger *et al.* [9], right: proposed. Cyclic colormap to enhance details.

holes in disparity images), but also to improve accuracy in areas with repetitive patterns (road surfaces under the cars, etc) and also to accurately recover thinner objects such as walking pedestrian. Note in particular that with lidar data, and segment-based prior, the discontinuity in depth better follows the object boudaries.

D. Quantitative results

Next, we quantitatively evaluate the accuracy of our approach. We assess the overall performance by linear and logarithmic root mean square error (RMSE) that are standard metrics defined as $\text{RMSE}_{\text{linear}} = \sqrt{\frac{1}{N}\sum_{d\in N} \|d_i - d_i^*\|^2}$ and $\text{RMSE}_{\text{log}} = \sqrt{\frac{1}{N}\sum_{d\in N} \|\log d_i - \log d_i^*\|^2}$, where d_i is the predicted disparity and d_i^* is the ground-truth. In spirit of disparity evaluation on the KITTI dataset, we use the lidar measurements as a ground-truth (as we do not have any other, more accurate and dense data). It is natural, that our approach performs well in these point measurements. However, our goal is to demonstrate that competitive performance can be achieved with worse sensors. Hence we reduce the number of lidar measurements that we use as pivots, *i.e.* we use each 2nd, 5th, 10th, etc. point and evaluate with respect to the unused points. Our approach significantly outperforms both baselines (elas [9], elas+lidar) and inference helps to get better results (unary vs full), see Fig. 5 (x axis denotes how many points we preserve from lidar measurements, *e.g.* 10 means that we keep each 10th point). The error increases very slowly, which suggests that even with significantly worse sensors we are able to maintain the desired precision – 18000 lidar measurements can be decreased to only 900 points without significant drop in performance.

V. Discussion

Despite very encouraging results, our system is not without limitations. In particular, processing temporal sequences assumes the mapped pivots correspond to the static parts of a scene. Though we have not included it into our system, the pivots corresponding to moving objects can be marked by motion or semantic segmentation [2] (which can potentially be included into our energy function) and excluded from mapping. Also, the quality of estimated depth on temporal sequences depends on estimated pose, however, this is not a limitation in practice as we anyway need accurate pose for 3D reconstruction.

For ease of exposition, we have not used any probabilistic model of lidar and/or camera taking sensor noise, resolution, etc. into account, however, both can be easilly included into our energy function.

VI. Conclusion

In this paper, we have proposed a probabilistic model that efficiently exploits complementarity between different depth-sensing modalities for online dense scene reconstruction. Our model uses planarity prior which is common in both the indoor and outdoor scenes. We demonstrated the effectiveness of our approach on the KITTI dataset, and provide qualitative and quantitative results showing high-quality dense reconstruction and labeling of a number of scenes. More importantly, we show that we are able to get very high quality reconstruction using colour data and only a few hundreds of lidar points. We are planning to incorporate higher order terms to enforce slanted planarity priors as part of future work.

References

[1] C. Urmson et al., “Autonomous driving in urban environments: Boss and the urban challenge,” *JFR*, 2008.

[2] V. Vineet, O. Miksik, M. Lidegaard, M. Nießner, S. Golodetz, V. A. Prisacariu, O. Kähler, D. W. Murray, S. Izadi, P. Perez, and P. H. S. Torr, “Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction,” in *ICRA*, 2015.

[3] S. Song and J. Xiao, “Sliding shapes for 3d object detection in depth images,” in *ECCV*, 2014.

[4] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, “Learning rich features from RGB-D images for object detection and segmentation,” in *ECCV*, 2014.

[5] O. Miksik, V. Vineet, M. Lidegaard, R. Prasaath, M. Nießner, S. Golodetz, S. L. Hicks, P. Perez, S. Izadi, and P. H. S. Torr, “The semantic paintbrush: Interactive 3d mapping and recognition in large outdoor spaces,” in *ACM CHI*, 2015.

[6] E. Potapova, K. M. Varadarajan, A. Richtsfeld, M. Zillich, and M. Vincze, “Attention-driven object detection and segmentation of cluttered table scenes using 2.5d symmetry,” in *ICRA*, 2014.

[7] T. Whelan, H. Johannsson, M. Kaess, J. J. Leonard, and J. McDonald, “Robust Real-Time Visual Odometry for Dense RGB-D Mapping,” in *ICRA*, 2013, pp. 5724–5731.

[8] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, “Indoor segmentation and support inference from rgbd images,” in *ECCV*, 2012.

[9] A. Geiger, M. Roser, and R. Urtasun, “Efficient Large-Scale Stereo Matching,” in *ACCV*, 2010.

[10] O. J. Woodford, P. H. S. Torr, I. D. Reid, and A. W. Fitzgibbon, “Global stereo reconstruction under second order smoothness priors,” *PAMI*, 2009.

[11] S. N. Sinha, D. Scharstein, and R. Szeliski, “Efficient high-resolution stereo matching using local plane sweeps,” in *CVPR*, 2014.

[12] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger, “Real-time 3D Reconstruction at Scale using Voxel Hashing,” *TOG*, 2013.

[13] P. Krähenbühl and V. Koltun, “Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials,” in *NIPS*, 2011.

[14] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” *IJCV*, 2001.

[15] C. Je and H.-M. Park, “Optimized hierarchical block matching for fast and accurate image registration,” *Signal Processing: Image Communication*, 2013.

[16] H. Hirschmller, “Stereo processing by semiglobal matching and mutual information,” *PAMI*, 2008.

[17] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, and S. N. Sinha, “Object stereo - joint stereo matching and object segmentation.” in *CVPR*, 2011.

[18] K. Yamaguchi, D. A. McAllester, and R. Urtasun, “Efficient joint segmentation, occlusion labeling, stereo and flow estimation,” in *ECCV*, 2014.

[19] P. H. S. Torr and A. Criminisi, “Dense stereo using pivoted dynamic programming,” in *Image and Vision Computing*, 2004.

[20] D. H. C., J. Kannala, L. Ladicky, and J. Heikkilä, “Depth map inpainting under a second-order smoothness prior,” in *SCIA*, 2013.

[21] G. Payen de La Garanderie and T. Breckon, “Improved depth recovery in consumer depth cameras via disparity space fusion within cross-spectral stereo,” in *BMVC*, 2014.

[22] J. Diebel and S. Thrun, “An application of markov random fields to range sensing,” in *NIPS*, 2005.

[23] J. Dolson, J. Baek, C. Plagemann, and S. Thrun, “Upsampling range data in dynamic environments.” in *CVPR*, 2010.

[24] H. Badino, D. Huber, and T. Kanade, “Integrating lidar into stereo for fast and improved disparity computation,” in *3DIMPVT*, 2011.

[25] D. Munoz, J. A. Bagnell, and M. Hebert, “Co-inference machines for multi-modal scene analysis,” in *ECCV*, 2012.

[26] D. Held, J. Levinson, and S. Thrun, “Precision tracking with sparse 3d and dense color 2d data,” in *ICRA*, 2013.

[27] C. Premebida, J. Carreira, J. Batista, and U. Nunes, “Pedestrian detection combining rgb and dense lidar data,” in *IROS*, 2014.

[28] A. Arnab, M. Sapienza, S. Golodetz, J. Valentin, O. Miksik, S. Izadi, and P. H. S. Torr, “Joint object-material category segmentation from audio-visual cues,” in *BMVC*, 2015.

[29] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[30] Y. Boykov, O. Veksler, and R. Zabih, “Fast Approximate Energy Minimization via Graph Cuts,” *PAMI*, vol. 23, no. 11, 2001.

[31] D. Koller and N. Friedman, *Probabilistic Graphical Models - Principles and Techniques*. MIT Press, 2009.

[32] O. Miksik and K. Mikolajczyk, “Evaluation of local detectors and descriptors for fast feature matching,” in *ICPR*, 2012.

[33] D. Nister, O. Naroditsky, and J. Bergen, “Visual odometry,” in *CVPR*, 2004.

[34] A. S. Huang, A. Bachrach, P. Henry, M. Krainin, D. Maturana, D. Fox, and N. Roy, “Visual Odometry and Mapping for Autonomous Flight Using an RGB-D Camera,” in *ISRR*, 2011.

[35] P. F. Felzenswalb and D. P. Huttenlocker., “Efficient graph-based image segmentation.” in *IJCV*, 2004.

[36] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The KITTI vision benchmark suite,” in *Proc. CVPR*, 2012.