Building Scene Models by Completing and Hallucinating Depth and Semantics

Miaomiao Liu[1], Xuming He[1], Mathieu Salzmann[2]

[1]Data61, CSIRO, and ANU, Australia [2]CVLab, EPFL, Switzerland
[1]miaomiao.liu@data61.csiro.au, xuming.he@anu.edu.au,
[2]mathieu.salzmann@epfl.ch

Abstract. Building 3D scene models has been a longstanding goal of computer vision. The great progress in depth sensors brings us one step closer to achieving this in a single shot. However, depth sensors still produce imperfect measurements that are sparse and contain holes. While depth completion aims at tackling this issue, it ignores the fact that some regions of the scene are occluded by the foreground objects. Building a scene model would therefore require to hallucinate the depth behind these objects. In contrast with existing methods that either rely on manual input, or focus on the indoor scenario, we introduce a fully-automatic method to jointly complete and hallucinate depth and semantics in challenging outdoor scenes. To this end, we develop a two-layer model representing both the visible information and the hidden one. At the heart of our approach lies a formulation based on the Mumford-Shah functional, for which we derive an effective optimization strategy. Our experiments evidence that our approach can accurately fill the large holes in the input depth maps, segment the different kinds of objects in the scene, and hallucinate the depth and semantics behind the foreground objects.

1 Introduction

Building 3D models of real scenes has been a longstanding goal of computer vision. While impressive results can be achieved with multi-view and video-based approaches [1–4], the progress of depth sensors and their decreasing prices make them an attractive alternative, able to capture 3D in a single shot [5]. Unfortunately, even the best depth sensors still provide imperfect measurements. In particular, these measurements are often sparse and contain large holes due to various factors, such as reflective surfaces or too-distant portions of the scenes.

Overcoming these limitations has therefore recently become a popular research topic. For instance, *depth super-resolution* [6–11] tackles the sparseness issue and attempts to densify the observed depth data. Typically, however, existing methods assume that the measurements are regularly spaced, and are thus ill-suited to handle large holes. By contrast, *depth completion* or *inpainting* [12, 13] are designed to handle irregular measurements and fill holes in the input depth maps by leveraging RGB image information, or fusing multiple depth measurements [14]. These methods, however, simply complete the observed data.

As a consequence, they are ill-suited to build a model of a scene, where one is *not* interested in modeling the foreground objects. To address this problem, one should truly *hallucinate* the depth behind the observed foreground objects.

Only little work has been done to tackle the task of depth hallucination from a noisy depth map and its corresponding RGB image [12, 13, 15, 16], and existing methods typically work under additional assumptions. For example, [12, 13] rely on a user-defined foreground mask to hallucinate the background depth. The method in [15] relies on a layered depth model simply assuming that each layer is a smoothly varying surface, thus not considering semantics or image information. While [16] exploits image and semantics, it relies on CAD models to represent the foreground objects. Furthermore, both methods were designed for the indoor scenario, and are thus ill-suited to handle complex outdoor scenes.

By contrast, in this paper, we introduce a fully automatic approach to performing depth completion and hallucination for general (outdoor) scenes in a single shot. To this end, we develop a two-layer scene model accounting for the visible information and the hidden one. In each layer, we jointly estimate the depth and the semantics of the scene. Not only does this let us leverage depth to detect the foreground objects, but it also allows us to exploit the dependencies between depth and semantics to improve completion and hallucination. As evidenced by Fig. 1, our approach lets us accurately fill the large holes in the input depth maps, segment the different kinds of objects observed in the scene, and hallucinate the depth and semantics behind the foreground objects.

Specifically, we rely on the assumptions that depth is piecewise planar, semantics piecewise constant, and that the discontinuities of both modalities should largely coincide. We show that these assumptions can be formalized with a single Mumford-Shah functional. We then formulate the task of jointly completing and hallucinating depth and semantics as a discrete-continuous optimization problem whose variables encode a foreground-background mask and two layers of depth and semantics information: one for the data that is visible in the image/depth map and one for the data that is hidden behind the foreground. Following an alternating optimization strategy, we show that each type of variables has an elegant solution; the discrete ones can be computed via simple thresholding, and the continuous ones via a primal-dual algorithm implemented on the GPU. Altogether, this provides us with an effective framework to build scene models from a single noisy depth map and its corresponding RGB image despite the presence of undesirable foreground objects.

We demonstrate the effectiveness of our approach on two datasets, *i.e.*, KITTI [17] and Stixel [18]. Our experiments evidence that our method can produce accurate models of complex outdoor scenes without requiring any manual intervention. This, we believe, constitutes a significant step towards making 3D scene modeling in real, dynamic environments practical.

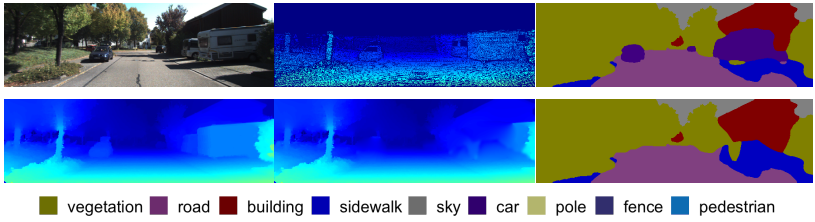

Fig. 1. Our approach. Given an input RGB image and a noisy, incomplete depth map, we complete and hallucinate depth and semantic to produce a complete scene model. First row: Input RGB image, incomplete depth measurements, estimated semantics; Second row: completed depth for the visible layer, hallucinated depth and semantics for the hidden layer.

2 Related Work

With access to depth sensors becoming easier everyday, increasingly many methods rely on depth as input for various applications, such as autonomous driving [17], augmented reality [19] and personal robotics [20]. Unfortunately, depth sensors are not perfect; they typically produce relatively sparse measurements with large holes.

Depth super-resolution attempts to overcome the sparseness issue by generating a high-resolution depth map from a low-resolution one. This is typically achieved via Markov Random Fields [6, 21, 7], bilateral filtering [22], layered representations [23], patch-based approaches [10, 11], or depth transfer [8, 9]. These approaches, however, inherently assume to have access to regularly-spaced depth measurements, and thus cannot handle large holes in depth maps.

By contrast, depth completion techniques have been designed to work with irregular measurements and to fill in large holes. In this context, Liu et al. [24] combine a modified fast matching method with guided filtering to inpaint Kinect depth maps. In [25], image segmentation is exploited to complete range data. Herrera et al. [26] propose an MRF with second-order prior to inpaint piecewise planar depth maps. In [27], depth completion is formulated within a total variation framework where image cues guide the completion process. A different approach to depth completion consists of treating a depth map as an intensity image, and rely on standard image inpainting algorithms, such as [28] and [29]. All the above-mentioned methods focus on depth completion form a single view and aim at completing the visible scene information only.By contrast, some approaches have proposed to exploit multiple views [30, 14] and thus can handle the fact that parts of the scene are hidden in some views, albeit not all of them. Similarly, great progress has been made in building complete scene models by fusing multiple noisy depth maps [31–33]. These methods, however, assume to have access to multiple input depth images.

Only little work has been done on the problem of building a complete scene model in one shot, despite the presence of occluding objects. Guo and Hoiem [34] focus on semantic labeling of unseen surfaces without depth information. In the

context of stereo matching, Bleyer et al. [35] introduce a method that hallucinates depth in the regions that are occluded in one view, but not in both. In [12, 13], while the goal is indeed to replace the depth of foreground objects with that of the background, the methods assume to be given a perfect foreground mask, defined by a user. As a consequence, these approaches truly perform depth completion, albeit without the knowledge of the RGB intensity behind the foreground mask. By contrast, [15] and [16] work without any manual input. However, in both cases, the methods were designed for the indoor scenario, and are thus ill-suited to model complex outdoor scenes, which are typically much more challenging.

In this paper, we introduce a fully-automatic approach to jointly completing and hallucinating depth and semantics. A key component of our approach is the use of a Mumford-Shah functional [36], which defines a non-convex energy function that encourages piece-wise constant solutions. Strekalovskiy and Cremers [37] develop a real-time primal-dual algorithm for minimizing the Mumford-Shah functional with a single variable, which we use and extend in this paper. Furthermore, our work relies on the piece-wise planar world assumption [38]. Despite its simplicity, it has been widely adopted in modeling outdoor man-made scenes [39, 40]. Our work also relates to 3D scene understanding, where joint semantics and depth prediction has been explored, e.g., [41]. However, to the best of our knowledge existing methods do not recover hidden surfaces.

3 Our Approach

Given partial depth measurements and a corresponding intensity image, our goal is to produce a complete scene model with background depth and semantics at every pixel, including those that are hidden by foreground objects. To this end, we need to simultaneously perform depth completion, reason about semantics, and hallucinate the background scene behind the foreground objects.

To achieve this, we introduce a two-layer scene representation modeling the visible information and the hidden one. Each layer consists of two modalities: depth and semantics. The resulting model is encoded by a discrete-continuous optimization problem. In Section 4, we develop an optimization procedure to minimize the corresponding energy, thus allowing us to jointly complete and hallucinate depth and semantics.

3.1 A Visible Layer for Semantics-aware Depth Completion

We first focus on modeling the scene that is visible in the input data. We assume that the underlying scene is piecewise planar and the corresponding semantic label map piecewise constant. Furthermore, we rely on the intuition that the depth discontinuities are often aligned with the boundaries of semantic classes, which lets us exploit the semantics to further regularize depth completion.

Let I be an input image of size $m \times n$ and $\mathbf{x} \in \Omega$ denote a pixel location on the two dimensional image plane Ω. We associate each pixel with two variables encoding depth value and semantic label, respectively. The semantic label

$\mathbf{s}^v(\mathbf{x}) \in \mathbb{R}^L$ is represented as an L-dimensional vector for L classes. As for depth, in this work, we make use of a disparity-based representation.[1] The motivation behind this is the following: Let $y^v(\mathbf{x}) \in \mathbb{R}$ be the disparity value at pixel $\mathbf{x}$. This disparity value can be equivalently encoded by plane parameters $\mathbf{u}^v(\mathbf{x}) \in R^3$, since we can write $y^v(\mathbf{x}) = \mathbf{p}(\mathbf{x})^T\mathbf{u}^v(\mathbf{x})$, where $\mathbf{p}(\mathbf{x}) = (\mathbf{x}^T, 1)^T$ is the homogeneous coordinate representation of $\mathbf{x}$. Then, our piecewise planar assumption of the depth map, which is equivalent to a piecewise planar assumption of the disparity map, can be encoded by a piecewise *constant* assumption on the plane parameters. This therefore allows us to define a unified Mumford-Shah functional on $\mathbf{u}^v$ and $\mathbf{s}^v$, which simultaneously encodes our two initial assumptions.

The Mumford-Shah functional [36] was originally introduced to compute a piecewise smooth approximation of observed data. In our context, let us denote by $\{y^o(\mathbf{x})\}_{\mathbf{x}\in\Omega}$ the incomplete disparity measurements, with disparity observation mask $\{d(\mathbf{x})\}_{\mathbf{x}\in\Omega}$, where $d(\mathbf{x}) = 1$ if the disparity measurement at pixel location $\mathbf{x}$ is valid, and 0 otherwise. Furthermore, let $\mathbf{s}^o(\mathbf{x})$ be a noisy label probability distribution at pixel $\mathbf{x}$, obtained by any image-based semantic labeling method. Our goal therefore is for our visible layer to fit the observed data, and thanks to our change of variable, that both $\mathbf{u}^v$ and $\mathbf{s}^v$ are piecewise constant while having their discontinuities aligned. This can be expressed by a coupled Mumford-Shah functional of the form

$$E_v(\mathbf{u}^v, \mathbf{s}^v) = E_d(\mathbf{u}^v, \mathbf{s}^v) + E_{r,v}(\mathbf{u}^v, \mathbf{s}^v) \;, \tag{1}$$

where $E_d(\mathbf{u}^v, \mathbf{s}^v)$ is the data fidelity term, and $E_{r,v}(\mathbf{u}^v, \mathbf{s}^v)$ denotes the regularization term that jointly encodes the piecewise constant and aligned discontinuities assumptions. We now describe these two energy terms in details.

Data term. The data term encourages the disparity and semantic label predictions to be consistent with the incomplete disparity measurements and the noisy semantic label probabilities. This can be expressed as

$$E_d(\mathbf{u}^v, \mathbf{s}^v) = \sum_{\mathbf{x}\in\Omega} d \cdot (\mathbf{p}^T\mathbf{u}^v - y^o)^2 + \eta_d \sum_{\mathbf{x}\in\Omega} \|\mathbf{s}^v - \mathbf{s}^o\|^2 \;. \tag{2}$$

where η_d is a weight that balances the influence of depth and semantics.

Regularization term. The regularization term encourages both $\mathbf{u}^v$ and $\mathbf{s}^v$ to be piecewise constant while having their discontinuities aligned. Following the Mumford-Shah formalism, we express this as

$$E_{r,v}(\mathbf{u}^v, \mathbf{s}^v) = \eta_{rv} \sum_{\mathbf{x}\in\Omega} \min(\alpha_1 \|\mathbf{K}\mathbf{u}^v\|^2 + \|\mathbf{K}\mathbf{s}^v\|^2, \lambda_1) \;, \tag{3}$$

where η_{rv} and α_1 are parameters controlling the strength of the smoothness and of the coupling between the two modalities and λ_1 is the truncation parameter. Here, we further rely on the oriented gradient operator $\mathbf{K}$ of [27], which computes an image-adaptive gradient for each channel of $\mathbf{u}^v$ and $\mathbf{s}^v$. More specifically, the oriented gradient operator $\mathbf{K}$ at location $\mathbf{x}$ is defined by $T_I(\mathbf{x})\nabla$, where T_I is an

[1] Note that using disparity instead of depth does not really come at any loss of generality, since they simply are the inverse of each other, up to a constant. If provided with depth measurements for the image pixels, one can therefore easily convert them to pseudo-disparities.

image-based anisotropic diffusion tensor. This tensor is defined as

$$T_I = \exp(-\beta|\nabla I|^\gamma)\mathbf{n}\mathbf{n}^T + \mathbf{n}^\perp\mathbf{n}^{\perp T}, \tag{4}$$

where $\mathbf{n} = \frac{\nabla I}{|\nabla I|}$ and $\mathbf{n}^\perp$ is the normal vector to the image gradient. Note that T_I is a symmetric matrix, and hence $\mathbf{K} = T_I(\mathbf{x})\nabla$ is a linear operator.

3.2 Adding a Hidden Layer for Depth and Semantics Hallucination

Recall that our goal is to produce a complete scene model from incomplete depth measurements. While the functional introduced in the previous section can complete the missing depth it still only represents the visible information. As such, it is unable to infer the scene depth and semantics behind the foreground objects. To address this limitation, we incorporate a hidden layer that focuses on modeling and hallucinating the depth and semantics of the background scene.

Formally, we split the semantic class set $\mathcal{L}$ into two subsets, one for the foreground classes $\mathcal{L}_f$ and the other for the background ones $\mathcal{L}_b$. At each pixel location $\mathbf{x}$, we introduce two additional variables, $\mathbf{u}^h(\mathbf{x}) \in \mathbb{R}^3$ and $\mathbf{s}^h(\mathbf{x}) \in \mathbb{R}^L$, which encode the (potentially occluded) disparity value and semantic label of the hidden scene layer at $\mathbf{x}$. Furthermore, we define a binary variable $m(\mathbf{x})$ indicating the foreground class mask (i.e., where the hidden layer is invisible). In other words, for the pixels where $m(\mathbf{x}) = 1$, there are neither disparity measurements nor semantic predictions for the hidden layer variables $\mathbf{u}^h(\mathbf{x})$ and $\mathbf{s}^h(\mathbf{x})$. Note that this binary variable is not strictly necessary, since this information can be extracted from the semantics variables. However, as will be discussed in Section 4, introducing it makes the resulting problem easier to optimize.

To hallucinate the depth and semantics of the hidden scene layer, we rely on the following assumptions/constraints: In the parts of the image that correspond to foreground, 1) the hidden layer should be jointly piecewise constant in $\mathbf{u}^h$ and $\mathbf{s}^h$; 2) given training data, the hidden layer variables should follow the data statistics; 3) In the parts of the image that correspond to background, the visible and hidden layers should agree; 4) The mask and the visible semantics should be coherent. Below, we formalize these assumptions by defining a corresponding set of energy terms and linear constraints.

1) Piecewise constancy. Similarly to the visible layer, we define a regularization term $E_{r,h}(\mathbf{u}^h, \mathbf{s}^h, m)$ that encourages $\mathbf{u}^h$ and $\mathbf{s}^h$ to be piecewise constant and have aligned discontinuities. Here, however, we only enforce this term on the foreground regions, i.e., where $m(\mathbf{x}) = 1$. This can be expressed as

$$E_{r,h}(\mathbf{u}^h, \mathbf{s}^h, m) = \eta_{rh} \sum_{\mathbf{x}} m \cdot \min(\alpha_2 \|\nabla \mathbf{u}^h\|^2 + \|\nabla \mathbf{s}^h\|^2, \lambda_2) \,, \tag{5}$$

where η_{rh} and α_2 are parameters controlling the strength of the smoothness and of the coupling between the two modalities, and λ_2 is the truncation parameter. As there are no image cues for the hidden layer in the foreground regions, we use the standard gradient to penalize the discontinuities.

2) Training data statistics. Given training data, we compute an average disparity map for each background class $k \in \mathcal{L}_b$, denoted by $\{y_k^s(\mathbf{x})\}_{\mathbf{x}\in\Omega}$. We refer the reader to Section 5 for the details of this process. We then encourage the

disparity and semantics of the hidden layer to be consistent with this statistics, which can be expressed as

$$E_s(\mathbf{u}^h, \mathbf{s}^h, m) = \eta_s \sum_{\mathbf{x}} m \cdot \sum_{k \in \mathcal{L}_b} \mathbf{s}_k^h (\mathbf{p}^T \mathbf{u}^h - y_k^s)^2 \,. \tag{6}$$

where η_s is a weight defining the influence of this term.

3) Agreement between the two layers. These constraints can be directly expressed as

$$\mathbf{u}^h(\mathbf{x}) = \mathbf{u}^v(\mathbf{x}), \quad \mathbf{s}^h(\mathbf{x}) = \mathbf{s}^v(\mathbf{x}), \quad \forall \mathbf{x} \mid m(\mathbf{x}) = 0 \,, \tag{7}$$

4) Coherent mask and visible semantics. We encourage the mask and the visible semantics to agree by penalizing the discrepancy between the total probability mass of foreground classes predicted by $\mathbf{s}^v$ and the mask variable at every pixel. This can be written as

$$E_c(\mathbf{s}^v, m) = \eta_c \sum_{\mathbf{x}} \Big(\sum_{k \in \mathcal{L}_f} \mathbf{s}_k^v - m + b \Big)^2 \,. \tag{8}$$

where η_c is a weighting parameter and b is a bias for the foreground mask.

Altogether, our two-layer approach to completing and hallucinating depth and semantics can be expressed as the discrete-continuous optimization problem

$$\begin{aligned} \min_{\mathbf{u}^v, \mathbf{s}^v, \mathbf{u}^h, \mathbf{s}^h, m} \quad & E_d + E_{r,v} + E_{r,h} + E_s + E_c \\ \text{s.t.} \quad & \mathbf{u}^h(\mathbf{x}) = \mathbf{u}^v(\mathbf{x}),\ \mathbf{s}^h(\mathbf{x}) = \mathbf{s}^v(\mathbf{x}) \quad \forall \mathbf{x} \mid m(\mathbf{x}) = 0 \\ & \sum_k \mathbf{s}_k^v(\mathbf{x}) = 1,\ \mathbf{s}_j^v(\mathbf{x}) \geq 0,\ \sum_k \mathbf{s}_k^h(\mathbf{x}) = 1,\ \mathbf{s}_j^h(\mathbf{x}) \geq 0, \quad \forall \mathbf{x},\ j \\ & m(\mathbf{x}) \in \{0, 1\}, \quad \forall \mathbf{x} \end{aligned} \tag{9}$$

where E_d, $E_{r,v}$, $E_{r,h}$, E_s, E_c are defined in Eqs. (2), (3), (5), (6) and (8), respectively. The first two constraints come from Eq. (7), and the third and fourth ones encode the simplex domain of probability distributions, and the fifth one the binary nature of the foreground mask m.

4 Optimizing our Two-Layer Model

The optimization problem encoding our two-layer problem, defined in Eq. (9), is challenging to solve, since it has a large number of coupled discrete and continuous variables. Fortunately, given the disparity and semantics, optimizing the mask is straightforward; the optimal mask value at each pixel can be computed in a closed form. Furthermore, when the mask variables are given, the energy functional decomposes into two subproblems: one for the visible layer, and one for the hidden one. These subproblems correspond to multi-modal versions of the Mumford-Shah functional. An efficient first-order primal-dual algorithm was introduced by [37] to tackle the single-modality case. We show that this algorithm can be extended to address the multi-modal scenario.

We therefore adopt an alternating procedure to minimize Eq. (9). This procedure consists of three steps repeated iteratively. In the first and second step, we optimize w.r.t. the visible and hidden layer, respectively, and, in the third step, we update the mask variables. Since our procedure decreases the energy

functional in every cycle, it converges to a local minimum. Below, we first review the first-order primal-dual algorithm of [37] for solving the Mumford-Shah functional and then discuss the solution to each step of our minimization strategy.

Primal-Dual Algorithm for the Mumford-Shah Functional. The primal-dual algorithm in [37] aims to solve a non-convex optimization problem of form

$$\min_{\mathbf{y}} D(\mathbf{y}) + R(\mathbf{A}\mathbf{y}) \ , \tag{10}$$

where $D(\cdot)$ usually denotes a data fidelity term, and $R(\cdot)$ is the regularization term encouraging piecewise smoothness in the Mumford-Shah functional. Let $\mathbf{A}$ denote a linear operator, which can be the gradient operator ∇, or an oriented gradient operator $\mathbf{K}$ additionally encoding image gradient information.

The primal-dual formulation introduces a dual variable $\mathbf{q}$ and solves the equivalent saddle-point problem

$$\min_{\mathbf{y}} \max_{\mathbf{q}} \ \ D(\mathbf{y}) + < \mathbf{q}, \mathbf{A}\mathbf{y} > - R^{\star}(\mathbf{q}). \tag{11}$$

where R^* is the conjugate of the regularization term. Following the fast Mumford-Shah method of [37], the primal-dual update equations can be written as

$$\mathbf{q}^{n+1} = prox_{\sigma_n, R^{\star}}(\mathbf{q}^n + \sigma_n \mathbf{A}\bar{\mathbf{y}}^n), \quad \mathbf{y}^{n+1} = prox_{\tau_n, D}(\mathbf{y}^n - \tau_n \mathbf{A}^{-1}\mathbf{q}^{n+1}), \tag{12}$$

$$\theta_n = \tfrac{1}{\sqrt{1+4\tau_n}}, \ \tau_{n+1} = \theta_n \tau_n, \ \sigma_{n+1} = \tfrac{\sigma_n}{\theta_n}. \tag{13}$$

$$\bar{\mathbf{y}}^{\mathbf{n+1}} = \mathbf{y}^{n+1} + \theta_n(\mathbf{y}^{n+1} - \mathbf{y}^n), \tag{14}$$

where $prox_{\cdot,\cdot}(\cdot)$ denotes the proximal operator. The convergence [42] of this primal-dual procedure for a convex problem depends on the parameter values τ and σ, which must satisfy $\tau\sigma\|\mathbf{A}\|^2 \leq 1$. For non-convex functional, [37] shows the algorithm generates a bounded solution with empirically convergence.

Our procedure uses a similar primal-dual procedure to optimize the subproblems corresponding to the visible and hidden layers. These subproblems have a specific functional form for D and R. Moreover, they rely on two modalities, $\mathbf{u}$ and $\mathbf{s}$. Below, we develop our algorithms for the visible and hidden layers, respectively. We only provide the formulation of D and R as in Eq. (10) and refer the reader to the supplementary for the details of the proximal operators.

4.1 Optimization w.r.t. the Visible Layer $\mathbf{s}^v$, $\mathbf{u}^v$

In this step, we fix the variables in the hidden layer $\mathbf{u}^h$, $\mathbf{s}^h$ and the foreground mask m, and optimize the subproblem defined on the visible layer. We also relax the consistent constraints of Eq. (9) at this step. We will enforce the constraints after optimizing w.r.t the visible and hidden layer. The resulting subproblem can thus be written as

$$\min_{\mathbf{u}^v, \mathbf{s}^v} \qquad E_d(\mathbf{u}^v, \mathbf{s}^v) + E_{r,v}(\mathbf{u}^v, \mathbf{s}^v) + E_c(s^v, m). \tag{15}$$

Note that the subproblem objective can be written in the standard Mumford-Shah functional form when it is optimized w.r.t. either $\mathbf{u}^v$ or $\mathbf{s}^v$. Therefore, to optimize this subproblem with the primal-dual algorithm, we further divide the task into two steps.

Optimizing $\mathbf{u}^v$ with fixed $\mathbf{s}^v$. By fixing the semantic variable $\mathbf{s}^v$, we can write the objective in Eq. (15) in the standard Mumford-Shah form, with

$$D_{\mathbf{u}^v}(\mathbf{u}^v) = \sum_{\mathbf{x}} \|d(\mathbf{p}^T\mathbf{u}^v - y^o)\|^2 \,, \tag{16}$$

$$R_{\mathbf{u}^v}(\mathbf{K}\mathbf{u}^v) = \eta_{rv} \sum_{\mathbf{x}\in\Omega} \min(\alpha_1\|\mathbf{K}\mathbf{u}^v\|^2 + e_{uv}, \lambda_1) \,, \tag{17}$$

where $e_{uv} = \|\mathbf{K}\mathbf{s}^v\|^2$. Here, $\|\mathbf{K}\mathbf{u}\|^2 := \sum_j \|\mathbf{K}\mathbf{u}_j\|^2$ denotes the Euclidean norm, where $\mathbf{u}_j$ is the j-th channel in the multi-channel variable $\mathbf{u}$.

Optimizing $\mathbf{s}^v$ with fixed $\mathbf{u}^v$. We then fix the disparity variable $\mathbf{u}^v$, and write the objective in Eq. (15) in the standard form, which yields

$$D_{\mathbf{s}^v}(\mathbf{s}^v) = \sum_{\mathbf{x}} \eta_d\|(\mathbf{s}^v - \mathbf{s}^o)\|^2 + \eta_c \sum_{\mathbf{x}} (\mathbf{f}^T\mathbf{s}^v - m + b)^2 \,, \tag{18}$$

$$R_{\mathbf{s}^v}(\mathbf{K}\mathbf{s}^v) = \eta_{rv} \sum_{\mathbf{x}\in\Omega} \min(\alpha_1 e_{sv} + \|\mathbf{K}\mathbf{s}^v\|^2, \lambda_1) \,, \tag{19}$$

where $e_{sv} = \|\mathbf{K}\mathbf{u}^v\|^2$, and $\mathbf{f}$ is a binary vector with 1s in the position corresponding to the foreground classes and 0 everywhere else.

4.2 Optimization w.r.t. the Hidden Layer $\mathbf{s}^h$, $\mathbf{u}^h$

Let us now fix the disparity and semantics of the visible layer $\mathbf{u}^v$, $\mathbf{s}^v$ and the foreground mask $\mathbf{m}$, and optimize the functional w.r.t. the hidden layer variables $\mathbf{u}^h$, $\mathbf{s}^h$. We consider the following equivalent subproblem

$$\min_{\mathbf{u}^h,\mathbf{s}^h} \qquad E_s(\mathbf{u}^h, \mathbf{s}^h, m) + E_{r,h}(\mathbf{u}^h, \mathbf{s}^h, m) + E_p(\mathbf{u}^h, \mathbf{s}^h) \tag{20}$$

Where $E_p(\cdot)$ is a regularization term with the following form:

$$E_p(\mathbf{u}^h, \mathbf{s}^h) = \gamma_{uh} \sum_{\mathbf{x}} (1-m)(\mathbf{p}^T\mathbf{u}^h - \mathbf{p}^T\mathbf{u}^v)^2 + \gamma_{sh} \textstyle\sum_{\mathbf{x}} (1-m)(\mathbf{s}^h - \mathbf{s}^v)^2 \tag{21}$$

Here γ_{uh} and γ_{sh} are large weights (usually 1000), and we essentially use a soft version of consistency constraints to regularize the problem, which empirically produces a more stable optimization step. Similar to the visible layer, we divide the optimization of this subproblem into two steps.

Optimizing $\mathbf{u}^h$ with fixed $\mathbf{s}^h$. Fixing the semantic variable $\mathbf{s}^h$, and writing the objective in Eq. (20) in the standard form yields

$$D_{\mathbf{u}^h}(\mathbf{u}^h) = \gamma_{uh} \sum_{\mathbf{x}} (1-m)(\mathbf{p}^T\mathbf{u}^h - \mathbf{p}^T\mathbf{u}^v)^2 + m\ \eta_s \sum_j s_j^h (\mathbf{p}^T\mathbf{u}^h - y_j^s)^2 \,, \tag{22}$$

$$R_{\mathbf{u}^h}(\nabla\mathbf{u}^h) = \eta_{rh} m \min(\alpha_2\|\nabla\mathbf{u}^h\|^2 + e_{uh}, \lambda_2) \,, \tag{23}$$

where $e_{uh} = \|\nabla\mathbf{s}^h\|^2$.

Optimizing $\mathbf{s}^h$ with fixed $\mathbf{u}^h$. We then fix the disparity variable $\mathbf{u}^h$, and write the objective in Eq. (20) in the standard form, which yields

$$D_{\mathbf{s}^h}(\mathbf{s}^h) = \gamma_{sh} \sum_{\mathbf{x}} (1-m)(\mathbf{s}^h - \mathbf{s}^v)^2 + m\ \eta_s \sum_j s_j^h (\mathbf{p}^T\mathbf{u}^h - y_j^s)^2 \,, \tag{24}$$

$$R_{\mathbf{s}^h}(\nabla\mathbf{s}^v) = \eta_{rh} m \sum_{\mathbf{x}\in\Omega} \min(\alpha_2 e_{sh} + \|\nabla\mathbf{s}^v\|^2, \lambda_2) \,, \tag{25}$$

where $e_{sh} = \|\nabla\mathbf{u}^h\|^2$.

4.3 Adding Constraints and Updating the Foreground Mask m

After computing the visible and hidden variables without the constraints, we now project them onto the constraint set defined in Eq. (9). The projection onto the consistent constraint set is computed as $\mathbf{s}^v = \mathbf{s}^h = \frac{\mathbf{s}^v + \mathbf{s}^h}{2}$ and $\mathbf{u}^v = \mathbf{u}^h = \frac{\mathbf{u}^v + \mathbf{u}^h}{2}$. For semantics $\mathbf{s}^v$, $\mathbf{s}^h$, we then project them onto the probability simplex.

Given the semantic and disparity variables in the visible and hidden layers, the foreground mask variables are decoupled into a set of independent variables for each location $\mathbf{x}$. The problem can then be re-written as

$$\min_{m} \sum_{\mathbf{x}} w(\mathbf{x}) m(\mathbf{x}) \ , \ \text{ s.t. } \ m(\mathbf{x}) \in \{0, 1\}, \tag{26}$$

where the weight $w(\mathbf{x})$ is given by

$$w(\mathbf{x}) = \eta_{rh} \cdot \min(\alpha_2 \|\nabla \mathbf{u}^h\|^2 + \|\nabla \mathbf{s}^h\|^2, \lambda_2) + \eta_s \sum_j s_j^h (\mathbf{p}^T \mathbf{u}^h - y_j^s)^2 + \eta_c \big(1 - 2(\mathbf{f}^T \mathbf{s}^v + b)\big). \tag{27}$$

Ultimately, $m(\mathbf{x}) = 1$ if $w(\mathbf{x}) < 0$, and 0 otherwise.

5 Experiments

To demonstrate the effectiveness of our approach, we evaluated our method on two publicly available outdoor datasets: KITTI [17] and Stixel [18]. Below, we discuss our results on both datasets.

5.1 Experimental Setup

Initialization. We used SLIC [43] to produce an over-segmentation of the image, and fit a plane to each superpixel using the corresponding sparse depth observations. The resulting plane parameters are used as initialization for $\mathbf{u}^v$ for each pixel in the superpixels. For large holes where no observations were available in the superpixels, we initialized the plane parameters to zero.

We adopted the FCN-32s model [44] followed by smoothing via a fully-connected CRF [45], which allowed us to initialize $\mathbf{s}^v$ and foreground mask $\mathbf{m}$, as well as provides the observations $\mathbf{s}^o$. We initialize $\mathbf{u}^h$ and $\mathbf{s}^h$ from $\mathbf{u}^v$ and $\mathbf{s}^v$ and set the foreground regions to 0.

Ground-truth for the hidden layer. To the best of our knowledge, no ground-truth is available for the hidden layer variables. In order to provide a quantitative evaluation, we generated the ground truth in two different ways: (1) Manual annotation. We first annotated the hidden semantic labels, based on which we then filled in the hidden depth using the planes fitted to the superpixels around the true foreground mask. (2) Image and depth composition. We overlaid an object from an image (foreground image) on a background image of unoccluded scene. Since the camera intrinsics are roughly the same for both images, the depth map would be consistent after adding the object in the same location as in the foreground image.

Co-occurence statistics. To obtain the class-dependent disparity statistics $\{y_k^s\}$ in Eq. (6), we followed the intuition that semantics are often highly correlated with image location, which was exploited, for example, in [46] for depth

prediction. To this end, we follow a superpixel-based approach. For each superpixel j in the test image, we take the plane parameters of the corresponding pixels in all the training images. For each class k, we then cluster these plane parameters, and take the cluster center with largest size. We finally generate y_k^s as the disparity obtained from the plane parameters of this center.

Baselines. Note that our scene model consists of two layers. For the visible layer, depth estimation translates to the usual depth completion problem. We therefore compare the results of our visible layer with the of the classical method of [29], and with the more recent technique of [27].

For the hidden layer, since no other has tackled the outdoor scenario in a fully-automatic manner, we rely on the following two-stage strategy. We first generate a foreground mask using the state-of-the-art semantic labeling method, FCN-32s model [44], followed by a smoothing with a fully-connected CRF [45]. Let us denote by *Fg-Mask* this foreground mask and by *Bg-Mask* the remaining image pixels. In Bg-Mask, the appearance is known, and thus the same depth completion methods as before can be employed. In Fg-Mask, however, no appearance information about the background is available. We therefore apply the technique of [28] to inpaint this area, which, to the best of our knowledge, remains the most mature method when it comes to depth completion without intensity information. This yields two baselines, which we will refer to as Baseline-1 (semantic segmentation followed by [29] + [28]) and Baseline-2 (semantic segmentation followed by [27] + [28]). To compare the different algorithms, we make use of the following metrics:1) *visible-rmse*: the-root-mean-square-error (rmse) for the entire depth map; 2)*hidden-rmse*: the rmse for the depth map hallucinated underneath the ground truth foreground mask.

5.2 Results on KITTI

As a first dataset, we utilized three subsets of the KITTI data annotated with semantic labels and/or disparity maps, and provided by (i) Ladický et al. [47], i.e., 60 aligned images, with dense disparity map and accurate semantic labels; (ii) Xu et al. [48], i.e., 107 images with accurate semantic labels; and (iii) Ros et al. [49], i.e., 146 images with accurate semantic labels. Note that only Ladický et al. [47] provide ground-truth disparity maps. However, this subset is constrained in terms of the scene types it depicts, i.e. mostly residential areas. To make our evaluation more meaningful, we therefore only used 40 images of the first subset as test images, complemented by 14 images from the other subsets. To obtain the ground-truth disparity maps for these 14 images, we employed the MC-CNN-acrt stereo matching algorithm [50], which ranks at the top in the KITTI stereo challenge. To avoid biasing our conclusions with these different types of ground-truth, we report results on the entire set, $test - 54$, and on the two subsets, $sub - 40$ and $sub - 14$, respectively. We also partitioned the data according to Manhattan (MH: 35 images) vs Non-Manhattan (NMH: 19 images) scenes, and further evaluate our method on two different scene structures. The remaining images from the three subsets were split into 200 for training and 59 for validation. For semantics, we mapped different label annotations to 9 classes

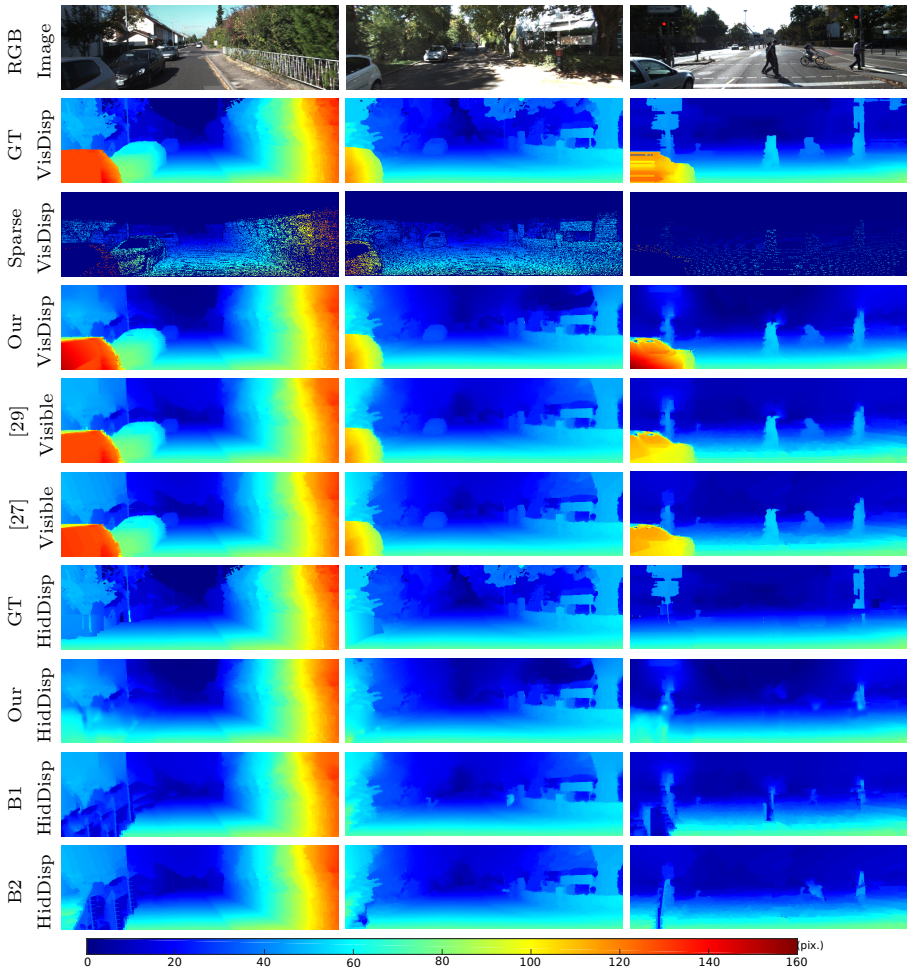

Fig. 2. Qualitative results on the KITTI dataset. For the disparity values, red denotes large values, and blue denotes small disparity values. **From top to bottom:** RGB image, ground-truth visible disparity map, sparse observations with large holes, our completed disparity map, two baselines for the visible layer, ground truth disparity for the hidden layer, our disparity for the hidden layer, and two baselines for the hidden layer. Note that our method can remove the foreground as well as accurately fill in the background disparity behind the foreground objects. Compared to the baselines, our approach can better complete the disparity for the visible and hidden layers.

and fine-tuned the FCN-32s of [44] to these 9 classes using the training data. We then define *car* and *pedestrian* as foreground classes.

In Table 1, we compare the results of our approach with the baselines for both the visible and hidden layers using the manually annotated ground truth. Note that we outperform the baselines in most cases. In particular, our approach

visible-rmse	test-54	sub-40	sub-14	MH	NMH	**hidden-rmse**	test-54	sub-40	sub-14	MH	NMH
Ours	**5.15**	**5.53**	**4.07**	**4.88**	5.66	Ours	**10.56**	**10.43**	**11.08**	**10.1**	**13.4**
[29]	5.42	5.67	4.68	5.49	**5.28**	Baseline-1	13.29	11.85	17.92	11.7	21.4
[27]	5.38	5.60	4.77	5.41	5.34	Baseline-2	12.53	11.37	16.34	11.3	19.1

Table 1. Depth estimation. Quantitative comparison with several baselines for the visible and hidden depth, respectively.

veg	road	building	sky	sidewalk	polar	fence	class-avg.	pixel-avg
73.6	51.52	85.07	16.64	16.97	3.61	0.51	35.42	50.08

Table 2. Estimating hidden semantics. Per-class and overall accuracy of our approach.

yields a large improvement in the hidden regions of the image. This evidences that our two-layer model is well-suited for the task of hallucinating depth, and thus constitutes a significant step towards being able to build scene models despite the presence of occluding foreground objects. Note that the fact that our model also yields more accurate depth estimates in the visible regions than state-of-the-art depth completion methods also suggests that it effectively leverages the visible information. Additionally, we created a test set of 14 images using the composition strategy described in the previous section, which gives us access to the ground-truth hidden depth. Note that the 14 images were chosen to respect the scene type ratio of the original test data. The resulting hidden-rmse of our method is 7.72, which is superior to Baseline-1 (9.76) and Baseline-2 (10.94). Fig. 2 provides a qualitative comparison of our results with the ground truth and the baselines.

In Table 2, we show the results of our semantics labeling estimates for the hidden regions. Here, since no baseline is available for this task, we only report the results of our approach. These results show that, while hallucinating small classes, such as fence and poles, remains challenging, our model yields good accuracy on the more common and larger classes. Note that effectively handling the small classes in outdoor semantic labeling is known to be difficult even when leveraging visible information. Finally, we observed that the semantic labeling accuracy in the visible layer did not significantly change compared to our initialization. In particular, we obtained 88.51% per pixel accuracy and 67.28% average per class accuracy. In Fig. 3, we provided the qualitative results for semantic segmentation on KITTI dataset.

To further illustrate the effect of our approach on the visible semantics, we initialized our algorithm with the results of FCN-32s only. The per-pixel and per-class accuracies of FCN-32s were 87.86% and 69.98%, respectively. Our method improved the per-pixel accuracy to 88.5% and left the per-class one virtually unchanged (69.81%). This also resulted in an improved visible-rmse of 5.01.

5.3 Results on Stixel

As a second experiment, we employed the Stixel dataset. This dataset contains 500 images with corresponding noisy depth (disparity) maps and semantics, par-

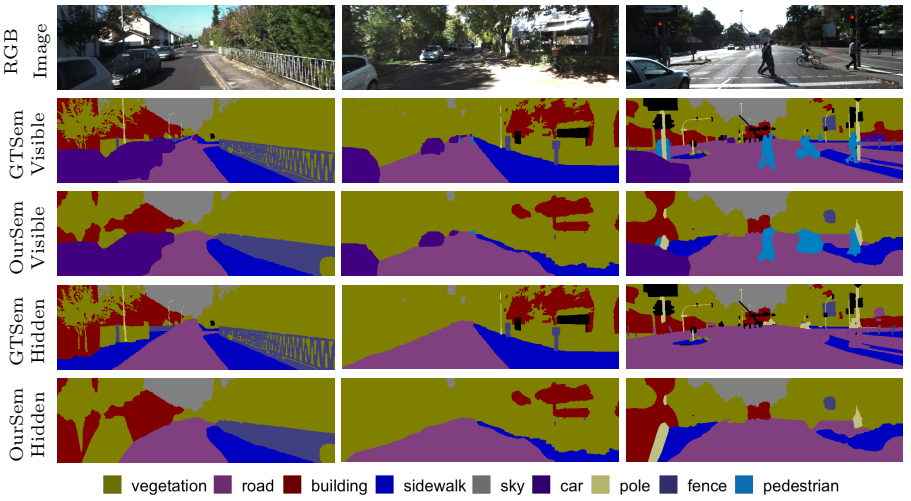

Fig. 3. Qualitative results for semantic segmentation on the KITTI dataset. From top to bottom: RGB image, ground truth results and our results, ground truth disparity for the hidden layer, our disparity for the hidden layer, Baseline 1 and Baseline 2, ground truth semantics for the hidden layer, and our estimated semantics for the hidden layer.

titioned into 300 training images and 200 test images. Note that the disparity provided in this dataset was computed using a semi-global matching algorithm. Since ground-truth disparity is only partially available for this dataset, it is therefore not possible to generate the ground-truth disparity for the foreground mask as before. We therefore only provide a qualitative comparison of our approach with with the baselines. There are 5 semantic classes in the dataset. We define *car* and *pedestrian* as the foreground class. The qualitative results of this dataset are shown in the Supplementary Material (Fig. 4). Note that, again, we can see that our approach produces more accurate disparity maps.

6 Conclusion

We have introduced a fully-automatic approach to jointly completing and hallucinating depth and semantics from an incomplete depth map and an RGB image. To this end, we have developed a two-layer model, encoding both the visible information and the information hidden behind the foreground objects. Furthermore, we have designed an effective strategy to optimize our two-layer model. Our experiments have evidenced that our approach can accurately fill the large holes in the input depth map, produce a semantic segmentation of the observed scene, and hallucinate the depth and semantics behind the foreground objects. In the future, we plan to extend our method to accumulate the information observed in a video sequence of a dynamic scene.

References

1. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from internet photo collections. IJCV (2008)
2. Cornelis, N., Leibe, B., Cornelis, K., Van Gool, L.: 3d urban scene modeling integrating recognition and reconstruction. IJCV (2008)
3. Hane, C., Zach, C., Cohen, A., Angst, R., Pollefeys, M.: Joint 3d scene reconstruction and class segmentation. In: CVPR. (2013)
4. Kundu, A., Li, Y., Dellaert, F., Li, F., Rehg, J.M.: Joint semantic segmentation and 3d reconstruction from monocular video. In: ECCV 2014. (2014)
5. Shao, L., Han, J., Kohli, P., Zhang, Z.: Computer vision and machine learning with RGB-D sensors. Springer (2014)
6. Diebel, J., Thrun, S.: An application of markov random fields to range sensing. In: NIPS. (2005)
7. Park, J., Kim, H., Tai, Y.W., Brown, M., Kweon, I.: High quality depth map upsampling for 3d-tof cameras. In: ICCV. (2011)
8. Aodha, O.M., Campbell, N.D., Nair, A., Brostow, G.: Patch based synthesis for single depth image super-resolution. In: ECCV. (2012)
9. Hornácek, M., Rhemann, C., Gelautz, M., Rother, C.: Depth super resolution by rigid body self-similarity in 3d. In: CVPR. (2013)
10. Kiechle, M., Hawe, S., Kleinsteuber, M.: A joint intensity and depth co-sparse analysis model for depth map super-resolution. In: ICCV. (2013)
11. Lu, S., Ren, X., Liu, F.: Depth enhancement via low-rank matrix completion. In: CVPR. (2014)
12. Wang, L., Jin, H., Yang, R., Gong, M.: Stereoscopic inpainting: Joint color and depth completion from stereo images. In: CVPR. (2008)
13. Doria, D., Radke, R.J.: Filling large holes in lidar data by inpainting depth gradients. In: CVPR Workshops. (2012)
14. Dolson, J., Baek, J., Plagemann, C., Thrun, S.: Upsampling range data in dynamic environments. In: CVPR. (2010)
15. Zach, C.: Dual decomposition for joint discrete-continuous optimization. In: AISTATS. (2013)
16. Geiger, A., Wang, C.: Joint 3d object and layout inference from a single rgb-d image. In: Pattern Recognition. Springer (2015) 183–195
17. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR. (2012)
18. Scharwächter, T., Enzweiler, M., Franke, U., Roth, S.: Stixmantics: A medium-level model for real-time semantic scene understanding. In: ECCV. (2014)
19. Glocker, B., Izadi, S., Shotton, J., Criminisi, A.: Real-time rgb-d camera relocalization. In: ISMAR. (2013)
20. Rusu, R.B., Holzbach, A., Diankov, R., Bradski, G., Beetz, M.: Perception for mobile manipulation and grasping using active stereo. In: IEEE-RAS International Conference on Humanoid Robots. (2009)
21. Wang, L., Jin, H., Yang, R., Gong, M.: Stereoscopic inpainting: Joint color and depth completion from stereo images. In: CVPR. (2008)
22. Yang, Q., Yang, R., Davis, J., Nistér, D.: Spatial-depth super resolution for range images. In: CVPR. (2007)
23. Shen, J., Cheung, S.: Layer depth denoising and completion for structured-light rgb-d cameras. In: CVPR. (2013)

24. Liu, J., Gong, X., Liu, J.: Guided inpainting and filtering for kinect depth maps. In: ICPR. (2012)
25. Bhavsar, A.V., Rajagopalan, A.N.: Range map superresolution-inpainting, and reconstruction from sparse data. CVIU **116**(4) (2012)
26. Herrera, D., Kannala, J., Heikkilä, J., et al.: Depth map inpainting under a second-order smoothness prior. In: Image Analysis. Springer (2013)
27. Ferstl, D., Reinbacher, C., Ranftl, R., Rüther, M., Bischof, H.: Image guided depth upsampling using anisotropic total generalized variation. In: ICCV. (2013)
28. Criminisi, A., Perez, P., Toyama, K.: Object removal by exemplar-based inpainting. In: CVPR. (2003)
29. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. In: TOG. (2004)
30. Schuon, S., Theobalt, C., Davis, J., Thrun, S.: Lidarboost: Depth superresolution for tof 3d shape scanning. In: CVPR. (2009)
31. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., et al.: Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In: ACM UIST. (2011)
32. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: IROS. (2012)
33. Zhou, Q.Y., Koltun, V.: Dense scene reconstruction with points of interest. TOG (2013)
34. Guo, R., Hoiem, D.: Beyond the line of sight: labeling the underlying surfaces. In: ECCV. (2012)
35. Bleyer, M., Rother, C., Kohli, P., Scharstein, D., Sinha, S.: Object stereojoint stereo matching and object segmentation. In: CVPR. (2011)
36. Mumford, D., Shah, J.: Optimal approximations by piecewise smooth functions and associated variational problems. Communications on pure and applied mathematics (1989)
37. Strekalovskiy, E., Cremers, D.: Real-time minimization of the piecewise smooth mumford-shah functional. In: ECCV. (2014)
38. Faugeras, O.D., Lustman, F.: Motion and structure from motion in a piecewise planar environment. International Journal of Pattern Recognition and Artificial Intelligence **2**(03) (1988)
39. Baillard, C., Zisserma, A.: Automatic reconstruction of piecewise planar models from multiple views. In: CVPR. (1999)
40. Gallup, D., Frahm, J.M., Pollefeys, M.: Piecewise planar and non-planar stereo for urban scene reconstruction. In: CVPR. (2010)
41. Ladický, L., Sturgess, P., Russell, C., Sengupta, S., Bastanlar, Y., Clocksin, W., Torr, P.H.: Joint optimization for object class segmentation and dense stereo reconstruction. IJCV (2012)
42. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of Mathematical Imaging and Vision (2011)
43. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Suesstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. PAMI (2012)
44. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. (2015)
45. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: NIPS. (2011)
46. Liu, B., Gould, S., Koller, D.: Single image septh estimation from predicted semantic labels. In: CVPR. (2010)

47. Ladický, L., Shi, J., Pollefeys, M.: Pulling things out of perspective. In: CVPR. (2014)
48. Xu, P., Davoine, F., Bordes, J.B., Zhao, H., Denœux, T.: Multimodal information fusion for urban scene understanding. Machine Vision and Applications (2014)
49. Ros, G., Ramos, S., Granados, M., Bakhtiary, A., Vazquez, D., Lopez, A.M.: Vision-based offline-online perception paradigm for autonomous driving. In: WACV. (2015)
50. Zbontar, J., LeCun, Y.: Stereo matching by training a convolutional neural network to compare image patches. JMLR. **17** (2016)