Semantic 3D Reconstruction with Continuous Regularization and Ray Potentials Using a Visibility Consistency Constraint

Nikolay Savinov, Christian Häne, Ľubor Ladický and Marc Pollefeys
ETH Zürich, Switzerland
{nikolay.savinov,christian.haene,lubor.ladicky,marc.pollefeys}@inf.ethz.ch

Abstract

We propose an approach for dense semantic 3D reconstruction which uses a data term that is defined as potentials over viewing rays, combined with continuous surface area penalization. Our formulation is a convex relaxation which we augment with a crucial non-convex constraint that ensures exact handling of visibility. To tackle the non-convex minimization problem, we propose a majorize-minimize type strategy which converges to a critical point. We demonstrate the benefits of using the non-convex constraint experimentally. For the geometry-only case, we set a new state of the art on two datasets of the commonly used Middlebury multi-view stereo benchmark. Moreover, our general-purpose formulation directly reconstructs thin objects, which are usually treated with specialized algorithms. A qualitative evaluation on the dense semantic 3D reconstruction task shows that we improve significantly over previous methods.

1. Introduction

One of the major goals in computer vision is to compute dense 3D geometry from images. Recently, also approaches that jointly reason about the geometry and semantic segmentation have emerged [10]. Due to the noise in the input data often strong regularization has to be performed. Optimizing jointly over 3D geometry and semantics has the advantage that the smoothness for a surface can be chosen depending on the involved semantic labels and the normal direction to the surface. Eventually, this leads to more faithful reconstructions that directly include a semantic labeling.

Posing the reconstruction task as a volumetric segmentation problem [4] is a widely used approach. A volume gets segmented into occupied and free space. In case of dense semantic 3D reconstruction, the occupied space label is replaced by a set of semantic labels [10]. To get smooth, noise-free reconstructions, the final labeling is normally determined by energy minimization. The formulation of the energy comes with the challenge that the observations are given in the image space but the reconstruction is volumetric. Therefore each pixel of an image contains information about a ray composed out of voxels. This naturally leads to energy formulations with potential functions that depend on the configuration of a whole ray. Including such potentials in a naive way leads to (on current hardware) infeasible optimization problems. Hence, many approaches try to approximate such a potential. One often utilized strategy is to derive a per-voxel unary potential (cost for assigning a specific label to a specific voxel). However, this is only possible in a restricted setting and under a set of assumptions that often do not hold in practice. By modeling the true ray potential, more faithful reconstructions are obtained [26]. Thus, efficient ways to minimize energies with ray potentials, while at the same time being able to benefit from the joint formulation of 3D modeling and semantic segmentation, are desired.

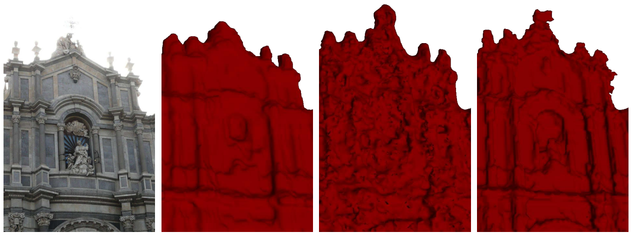


Figure 1: Left to right: example image, close-ups of [10], [26] and our proposed approach.

In this work, we propose an energy minimization strategy for ray potentials that can be directly used together with continuously inspired surface regularization approaches and hence does not suffer from metrication artifacts [12], common to discrete formulations on grid graphs. When using our ray potential formulation for dense semantic 3D reconstruction, it additionally allows for the usage of class-specific anisotropic smoothness priors. Continuously inspired surface regularization approaches are formulated as convex optimization problems. We identify that a convex relaxation for the ray potential is weak and unusable in practice. We propose to add a non-convex term that handles visibility exactly and optimize the resulting energy by linearly majorizing the non-convex part. By regularly re-estimating the linear majorizer during the optimization, we devise an energy minimization algorithm with guaranteed convergence.

1.1. Related Work

Visibility relations in 3D reconstructions were studied for computing a single depth map out of multiple images [14, 15]. To generate a full consistent 3D model from many depth maps, a popular approach is posing the problem in the volume [4]. To handle the noise in the input data a surface regularization term is added [19, 38]. A discrete graph-based formulation is used in [19] and continuous surface area penalization in [38]. One of the key questions is how to model the data term. Starting from depth maps, [19, 38] model the 2.5D data as per-voxel unary potentials. Such a modeling utilizes information contained in the depth map only partially. Using a discrete graph formulation [16, 33] propose to model the free space between the camera center and the measured depth with pairwise potentials.

Another approach to modeling the data term is to directly use a photo-consistency-based smoothness term [29, 11, 13]. To resolve the visibility relations, image silhouettes are used. This is done either in the optimization as a constraint, meaning that the reconstruction needs to be consistent with the image silhouettes [29, 13], or by deriving per-voxel occupancy probability [11]. Silhouette consistency is achieved through a discrete graph-cut optimization in [29], and with a convex-relaxation-based approach in the continuous domain in [13]. The resulting relaxed problem in the latter case is not tight and hence does not generate binary solutions. Therefore, to guarantee silhouette consistency a special thresholding scheme is required. Handling visibility has also been done in mesh-based photo-consistency minimization [5].

To fully address the 2.5D nature of the input data, the true ray potential should be used, meaning the data cost depends on the first occupied voxel along the ray. What happens behind is unobserved and hence has no influence. This was formulated in [25, 9, 21, 22, 31] as a problem of finding a voxel labeling in terms of color and occupancy such that the first occupied voxel along a ray has a similar color as the pixel it projects to. One of the limitations all these works share is that they only compare colors of single pixels, which often does not give a strong enough signal to recover weakly textured areas. We use depth maps that are computed based on comparing image patches and interpret them as noisy input data containing outliers. We use regularization to handle the noise and outliers in the input data, but in contrast to other approaches with ray potentials that use a purely discrete graph-based formulation [9, 21, 22] our proposed method is the first one that combines true ray potentials with a continuous surface regularization term. This allows us to set a new state of the art on two commonly used benchmark datasets. Unlike in any previous volumetric depth map fusion approach, thin surfaces do not pose problems in our formulation, due to an accurate representation of the input data.

Most earlier formulations of ray potentials are for purely geometry-based 3D reconstruction. Ours is more general and also allows to incorporate semantic labels. [26] shows that by using a discrete graph-based approach the true multi-label ray potential can be used as data term. Several artifacts present in the unary potential approximation [10] can be resolved using a formulation over rays. However, utilizing a discrete graph-based approach it is not directly possible to use the class-specific anisotropic regularization proposed in [10]. We bridge this gap and show how the full multi-label ray potential can be used together with continuously inspired anisotropic surface regularization [2, 37].

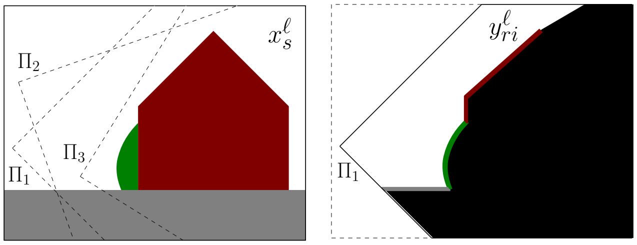
2. Formulation

In this section we will introduce the mathematical formulation that we are using to represent the dense semantic 3D reconstruction as an optimization problem. The problem is posed over a 3D voxel volume $\Omega \subset \mathbb{N}^3$. We denote the label $f = 0$ as the free space label and introduce the set $\mathcal{L} = \{0, 1, \dots, L\}$ of L semantic labels, which represent the occupied space, plus the free space label. The final goal of our method is to assign a label $\ell \in \mathcal{L}$ to each of the voxels. The label assignment is formalized using indicator variables $x_s^\ell \in \{0, 1\}$ indicating if label ℓ is assigned at voxel $s \in \Omega$, $(x_s^\ell = 1)$ or not.

We denote the vector of all per-voxel indicator variables as $\mathbf{x}$. Finally, the energy that we are minimizing in this paper has the form

$$E(\mathbf{x}) = \psi_R(\mathbf{x}) + \psi_S(\mathbf{x})$$
$$\text{subject to} \sum_{\ell \in \mathcal{L}} x_s^\ell = 1, \qquad x_s^\ell \in \{0, 1\}, \tag{1}$$

where $\sum_{\ell \in \mathcal{L}} x_s^\ell = 1$ guarantees that exactly one label is assigned to each voxel. The objective contains two terms. The term $\psi_R(\mathbf{x})$, the ray potential, contributes the data to the optimization problem. This is in contrast to many other formulations where the data term is specified as local per-voxel preferences for the individual labels. The second term $\psi_S(\mathbf{x})$ is a smoothness term, which penalizes the surface area of the transitions between different labels. For this term, we utilize formulations originating from convex continuous multi-label segmentation [2]. As we will see in Sec. 4 this smoothness term allows for class-specific anisotropic penalization of the interfaces between all labels. Due to the continuous nature of the regularization term, it does not suffer from metrication artifacts like most of the graph-based formulations. The straightforward way to utilize such a smoothness term would be to use a convex relax-



Global View

View from Camera Π_1

Figure 2: Variable Types: (left) the global x^ℓ_s, indicate the label assigned to each voxel, (right) the per-ray variables y^ℓ_{ri} describe the visible surface.

ation of the ray potential. Unfortunately, convex relaxations of the ray potential do not seem to lead to strong formulations (*c.f.* Fig.4). In this paper we show how to resolve this problem by adding a non-convex constraint. In Sec. 3 we introduce the convex relaxation formulation of the ray potential and its non-convex extension. The regularization term and optimization strategy are discussed in Sec. 4.

3. Ray Potential

The main idea of the ray potential [26] is that for each ray, originating from an image pixel, a cost is induced that only depends on the position of the first non-free space label along the viewing ray or the ray is all free space. This means that the potential can take only linearly many (in the number of voxels along a ray) values, which is the reason why optimization of such potentials is tractable. Note that this is not a restriction we impose, it represents the fact that it is impossible to see behind occupied space. We denote the cost of having the first occupied space label at position i with label $\ell \in \mathcal{L}$ as c^ℓ_{ri} and the cost of having the whole ray free space as c^f_r.

The vector of indicator variables x^ℓ_s belonging to ray $r \in \mathcal{R}$ is denoted as $\mathbf{x}_r$. To index positions along a ray, we denote $s_{ri} \in \Omega$ as the positions of all the voxels belonging to ray $r \in \mathcal{R}$, where $i \in \{0, \cdots, N_r\}$ denotes the position index along the ray. Note that there exists only one x^ℓ_s variable per label for each voxel $s \in \Omega$, if s_{ri} evaluates to the same position for different rays it refers to the same variable. Now we can state the ray potential part of the energy as a sum of potentials over rays

$$\psi_R(\mathbf{x}) = \sum_{r\in\mathcal{R}} \psi_r(\mathbf{x}_r) \qquad (2)$$

$$\psi_r(\mathbf{x}_r) = \left(\sum_{\ell\in\mathcal{L}\backslash\{f\}} \sum_{i=0}^{N_r} c^\ell_{ri}\left(\min_{j\le i-1} x^f_{s_{rj}}\right) x^\ell_{s_{ri}}\right) + c^f_r \min_{j\le N_r} x^f_{s_{rj}}$$

with $\mathcal{L}\backslash\{f\}$ meaning the set of all labels excluding the free space label. The term $(\min_{j\le i-1} x^f_{s_{ri}}) x^\ell_{s_{ri}}$ is 1 iff the first occupied label along the ray r is ℓ at position i. Similarly, $\min_{j\le N_r} x^f_{s_{rj}}$ equals 1 iff the whole ray r is free space. Thus, in Eq. 2 only one term is non-zero, and its coefficient equals the desired cost of the ray configuration.

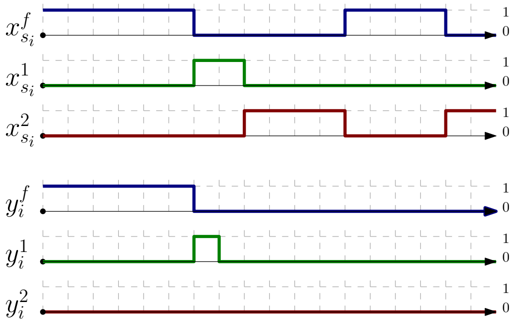


Figure 3: Example of variable assignments along a single viewing ray for a three-label problem.

To make the derivations throughout the paper compact, we omit the last term without loss of generality by shifting the costs by a constant, $c^\ell_{ri} \leftarrow c^\ell_{ri} - c^f_r$ and $c^f_r \leftarrow 0$.

3.1. Visibility Variables

Before we state a convex relaxation of the ray potential, which we eventually augment with a non-convex constraint, we rewrite the potential using visibility variables. First, we introduce the visibility variables y^ℓ_{ri} indicating that the ray r only contains free space up to the position $i-1$ and the label assigned at position i is $\ell \in \mathcal{L}$.

$$y^\ell_{ri} = \min(y^f_{r,i-1}, x^\ell_{s_{ri}}) \qquad (3)$$

To anchor the definition we assume that the -1$^{\text{st}}$ voxel of the ray has free space assigned, $y^f_{r,-1} = 1$. Note that if we insert all the nested definitions for a free space variable we get $y^f_{ri} = \min_{j\le i} x^f_{s_{rj}}$. Note that this variables are per-ray local variables and multiple ones can exist per voxel in case multiple rays cross that voxel in contrast to the global per-voxel variables x_s, which exist only once per voxel (*c.f.* Fig. 2). In the remainder of Sec. 3 we will drop the index r at most places for better readability. We now state a reformulation of the ray potential as

$$\psi_r(\mathbf{x}_r) = \sum_{\ell\in\mathcal{L}} \sum_{i=0}^{N} c^\ell_i y^\ell_i \qquad (4)$$

$$\text{subject to} \quad y^\ell_i = \min(y^f_{i-1}, x^\ell_{s_i}) \;\; \forall \ell \in \mathcal{L}, \forall i,$$

Here we introduced costs along the ray also for the free space label $c^f_i = 0$, $\forall i$. This does not change the potential but is required for our next step, where we reformulate the non-convex equality constraints as a series of inequality constraints. To make sure that the corresponding equalities are still satisfied in the optimum, we show that the costs c^ℓ_i can be replaced by non-positive ones without changing the

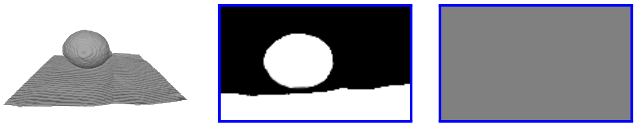


Figure 4: Evaluation of the convex relaxation for two-label problem: **(left)** a reconstruction of the model obtained by our non-convex procedure, slices through the volume (0 black, 1 white, 0.5 grey) in the non-convex formulation **(middle)** and the convex formulation **(right)**.

minimizer of the energy. This means that the y_i^ℓ are bounded from above by the linear inequality constraints and are tight from below through the minimization of the cost function, so the resulting constraints model the same optimization problem. The inequality constraints read as follows

$$0 \le y_i^\ell \le y_{i-1}^f, \quad y_i^\ell \le x_{s_i}^\ell \qquad \forall \ell \in \mathcal{L} \tag{5}$$

To derive the transformation to non-positive costs, we first notice that after applying $\min(y_{i-1}^f, \cdot)$ to both sides of the constraint $\sum_{\ell \in \mathcal{L}} x_{s_i}^\ell = 1$ from Eq. 1, we can plug in the constraints of Eq. 4 to obtain

$$y_{i-1}^f = \sum_{\ell \in \mathcal{L}} y_i^\ell. \tag{6}$$

Intuitively, this means if position $i-1$ is in the observed visible free space then the next position is either free space or one of the occupied space labels and if $i-1$ is in the occupied space then all the y_i^ℓ are 0 (see Fig. 3). The cost transformation is done for every ray separately. Starting with the last position $i = N$, we add the following expression, which always evaluates to 0, to the ray potential.

$$\left(\max_{\ell' \in \mathcal{L}} c_i^{\ell'} \right) \left(y_{i-1}^f - \sum_{\ell \in \mathcal{L}} y_i^\ell \right) = 0 \tag{7}$$

This moves one non-negative term to the previous position and make all the c_i^ℓ for the current position non-positive.

$$\begin{aligned} c_{i-1}^f &\leftarrow c_{i-1}^f + \max_{\ell' \in \mathcal{L}} c_i^{\ell'} \\ c_i^\ell &\leftarrow c_i^\ell - \max_{\ell' \in \mathcal{L}} c_i^{\ell'} \quad \forall \ell \in \mathcal{L} \end{aligned} \tag{8}$$

This is done iteratively for all $i \in \{N, \dots, 0\}$, leaving just a constant, which can be omitted.

3.2. Convex Relaxation and Visibility Consistency

So far our derivation has been done using binary variables $x_s^\ell \in \{0, 1\}$ and hence also all the $y_i^\ell \in \{0, 1\}$. To minimize the energy, we relax this constraint by replacing $x_s^\ell \in \{0, 1\}$ with $x_s^\ell \in [0, 1]$ in Eq. 1. This directly leads to a convex relaxation of the ray potential. Unfortunately, this relaxation is weak and therefore inapplicable in practice. In Fig. 4 we evaluate the convex relaxation on a two-label example (Lemon dataset), using surface area penalization via a total variation (TV) smoothness prior. The convex relaxation fails entirely, producing variable assignments to the x_s^ℓ that are 0.5 up to machine precision and hence no meaningful solution can be extracted. A comparison of the energies reveals that there is a significant difference between the non-convex and the convex solution (626614 and 431893, respectively), which indicates that the relaxed problem is far from the original binary one. Most importantly, our earlier convex formulation [26] shares this behavior of not making a decision for any voxel, when run without initialization on a two-label problem. The aspects of initialization, heuristic assignment of unassigned variables, move making algorithm, and a coarse-to-fine scheme are essential elements of the algorithm in [26].

The reason for the weak relaxation is that Eq. 6 is unsatisfied for the solution of the convex relaxation. This equation ensures that the per camera local view is consistent with the global model (*c.f.* Fig. 2). Concretely, the equation states that the change in visibility is directly linked to the cost that can be taken by the potential *e.g.* a surface can only be placed iff the occupancy along the ray changes. Hence we propose a formulation that directly enforces this constraint, which we will call visibility consistency constraint. Eq. 6 can be reformulated using the definition of y_i^f as

$$\begin{aligned} \sum_{\ell \in \mathcal{L} \setminus \{f\}} y_i^\ell = y_{i-1}^f - y_i^f &= y_{i-1}^f - \min(y_{i-1}^f, x_{s_i}^f) \\ &= \max(0, y_{i-1}^f - x_{s_i}^f). \end{aligned} \tag{9}$$

This means that we can only have an occupied space label $\ell \in \mathcal{L} \setminus \{f\}$ assigned as the visible surface at position i, if position i does not have free space assigned and $y_{i-1}^f = 1$ and hence the whole ray from the camera center to the position $i-1$ has free space assigned (see Fig. 3).

Since we minimize the objective with non-positive c_i^ℓ, the visibility consistency constraint is equivalent to the inequality

$$\sum_{\ell \in \mathcal{L} \setminus \{f\}} y_i^\ell \le \max(0, y_{i-1}^f - x_{s_i}^f). \tag{10}$$

Our final formulation for the ray potential is

$$\begin{aligned} \psi_r(\mathbf{x}_r) = \sum_{\ell \in \mathcal{L}} \sum_{i=0}^{N} c_i^\ell y_i^\ell \\ \text{s.t.} \;\; y_i^\ell \le y_{i-1}^f, \; y_i^\ell \le x_{s_i}^\ell, \; y_i^\ell \ge 0 \;\; \forall \ell \in \mathcal{L}, \forall i \\ \sum_{\ell \in \mathcal{L} \setminus \{f\}} y_i^\ell \le \max(0, y_{i-1}^f - x_{s_i}^f) \qquad \forall i \end{aligned} \tag{11}$$

The above potential is non-convex because of the non-convex inequality which describes visibility consistency.

We follow the strategy of using a surrogate convex constraint for the non-convex one that majorizes the objective of the non-convex program. The majorization, as we will see in Sec. 4, happens during the iterative optimization. Therefore, at each iteration, we have a current assignment to the variables, which we denote by $\mathbf{x}^{(n)}$ and $\mathbf{y}^{(n)}$. Here we introduced the notation that variable assignments at iteration n are denoted with a superscript (n). Replacing

$$\begin{aligned} &\sum_{\ell \in \mathcal{L} \setminus \{f\}} y_i^\ell \leq \max\{0, y_{i-1}^f - x_{s_i}^f\} \\ \text{by} \quad &\sum_{\ell \in \mathcal{L} \setminus \{f\}} y_i^\ell \leq g(x_{s_i}^f, y_{i-1}^f | x_{s_i}^{f,(n)}, y_{i-1}^{f,(n)}) \end{aligned} \tag{12}$$

with the linear majorizer,

$$\begin{aligned} &g(x_{s_i}^f, y_{i-1}^f | x_{s_i}^{f,(n)}, y_{i-1}^{f,(n)}) \\ &\qquad = \begin{cases} 0 & \text{if } y_{i-1}^{f,(n)} \leq x_{s_i}^{f,(n)} \\ y_{i-1}^f - x_{s_i}^f & \text{if } y_{i-1}^{f,(n)} > x_{s_i}^{f,(n)} \end{cases} \end{aligned} \tag{13}$$

leads to a surrogate linear (and therefore convex) ray potential, which we will denote by $\psi_r^{(n)}(\mathbf{x}, \mathbf{y} | \mathbf{x}^{(n)}, \mathbf{y}^{(n)})$. The variables $\mathbf{x}^{(n)}$ and $\mathbf{y}^{(n)}$ denote the position of the linearization. We handle the corner case where both branches are feasible to always take the first branch. In numerical experiments we observed that this choice is not critical, it makes no significant difference which branch is used in this case.

Next we state a Lemma that will be a crucial part of the optimization strategy detailed in Sec. 4.

Lemma 1. *Given $\mathbf{x}^{(n)}$, with $\mathbf{x}^{(n)} \geq 0$ point-wise, we can find $\tilde{\mathbf{y}}^{(n)}$ such that all the constraints of the ray potential Eq. 11 are fulfilled and the value of the potential is minimal.*

Intuitively, the lemma states that given the global per-voxel variable assignments x_s^ℓ, an assignment to the per-ray variables y_i^ℓ can be found. This is not surprising given that the whole information about the scene is contained in the variables x_s^ℓ (*c.f*. Fig. 2). We prove the lemma by giving a construction.

Proof. We provide an algorithm that computes $\tilde{\mathbf{y}}^{(n)}$ for each ray individually. First we set $\tilde{y}_i^{f,(n)} = \min_{j \leq i} x_{s_j}^{f,(n)}$, which satisfies $\tilde{y}_i^{f,(n)} \leq \tilde{y}_{i-1}^{f,(n)}$, $\tilde{y}_{i\ell}^{(n)} \geq 0$. Now we iteratively increase $\tilde{y}_i^{\ell,(n)}$ such that $\sum_{\ell \in \mathcal{L} \setminus \{f\}} \tilde{y}_i^{\ell,(n)} \leq \max(0, \tilde{y}_{i-1}^{f,(n)} - x_{s_i}^f)$ and $\tilde{y}_i^{\ell,(n)} \leq x_{s_i}^\ell$. For an optimal assignment we do this procedure in an increasing order of c_i^ℓ. The observation holds by construction. □

4. Energy Minimization Strategy

Before we discuss the proposed energy minimization, we complete the formulation by including the regularization term.

4.1. Regularization Term

There are several choices of regularization terms for continuously inspired multi-label segmentation that can be inserted into our formulation [36, 2, 30, 37]. They are all convex relaxations and are originally posed in the continuum and discretized for numerical optimization. The main differences are the strength of relaxation and generality of the allowed smoothness priors. We directly describe the strongest, most general version, which allows for non-metric and anisotropic smoothness [37]. We only state the smoothness term and explain the meaning of the individual variables. For a thorough mathematical derivation we refer the reader to the original publications [37, 10].

$$\begin{aligned} \psi_S(\mathbf{x}, \mathbf{z}) &= \sum_{s \in \Omega} \psi_s(\mathbf{x}, \mathbf{z}) \quad \text{with} \\ \psi_s(\mathbf{x}, \mathbf{z}) &= \sum_{\ell, m : \ell < m} \phi_s^{\ell m}(z_s^{\ell m} - z_s^{m \ell}) \\ \text{s.t. } x_s^\ell &= \sum_m \left(z_s^{\ell m}\right)_k, x_s^\ell = \sum_m \left(z_{s-e_k}^{m\ell}\right)_k, \forall k, z_s^{\ell m} \geq 0. \end{aligned} \tag{14}$$

The variables $z_s^{\ell m} \in \mathbb{R}^3$ describe the transitions between the assigned labels. They indicate how much change there is from label ℓ to label m along the direction they point to and are hence called label transition gradients. For example, if there is a change from label ℓ to label m at voxel s along the first canonical direction, the corresponding $z_s^{\ell m}$ is $[1, 0, 0]^T$. The $z_s^{\ell m}$ need to be non-negative in order to allow for general, non-metric smoothness priors [37]. Therefore the difference $z_s^{\ell m} - z_s^{m\ell}$ is used to allow for arbitrary transition directions. The variable e_k denotes the canonical basis vector for the k-th component, *i.e.* $e_1 = [1, 0, 0]^T$. $\phi_s^{\ell m} : \mathbb{R}^3 \rightarrow \mathbb{R}_0^+$ are convex positively 1-homogeneous functions that act as anisotropic regularization of a surface between label ℓ and m. Note that the regularization term takes into account label combinations. This enables us to select class-specific smoothness priors, which depend on the surface direction and the involved labels and are inferred from training data [10]. For example, a surface between ground and building is treated differently from a transition between free space and building. The following lemma will be necessary for our optimization strategy.

Lemma 2. *Given $\mathbf{x}^{(n)}$, $\mathbf{z}^{(n)}$, with $x_s^{\ell,(n)} \geq 0\ \forall s, \ell$ an assignment $\tilde{\mathbf{z}}^{(n)}$ can be determined that fulfills the constraints of the regularization term.*

For the full proof of the lemma we refer the reader to the supplementary material, here we only state the main idea of the proof. In a first step we project our current solution onto the space spanned by the equality constraints. This leads to an initialization of the $\tilde{\mathbf{z}}^{(n)}$ which fulfills the equality constraints but might lead to negative assignments to the $z_s^{\ell m}$.

To get a non-negative solution, we notice that as long as there is a $z_s^{\ell',m'}$ which is negative we can find ℓ'' and m'' such that we can increase $z_s^{\ell',m'}$ by ϵ along with changing $z_s^{\ell'',m'}, z_s^{\ell',m''}, z_s^{\ell'',m''}$ by the same ϵ in order not to affect the equality constraints.

4.2. Optimization

The goal of this section is to minimize the proposed energy using the non-convex ray potential Eq. 11. Optimizing non-convex functionals is an inherently difficult task. One often successfully utilized strategy is the so called majorize-minimize strategy (for example [18]). The idea is to majorize the non-convex functional in some way with a surrogate convex one. Alternating between minimizing the surrogate convex energy, which we will call the minimization step in the following, and recomputing the surrogate convex majorizer, which we will denote the majorization step, leads to an algorithm that decreases the energy at each step and hence converges.

Note that we already discussed the majorization step of the ray potential in Sec. 3, Eq. 13. Together with the regularizer we end up with a surrogate convex but non-smooth program.

$$E^{(n)}(\mathbf{x},\mathbf{y},\mathbf{z}) = \psi_R^{(n)}(\mathbf{x},\mathbf{y}|\mathbf{x}^{(n)},\mathbf{y}^{(n)}) + \psi_S(\mathbf{x},\mathbf{z}) \quad (15)$$
$$\text{s.t.} \sum_{\ell\in\mathcal{L}} x_s^\ell = 1\ \forall s,\quad x_s^\ell \in [0,1]\ \forall s \in \Omega,\ \forall \ell \in \mathcal{L}.$$

This energy can be globally minimized using the iterative first order primal-dual algorithm [24]. However, there is no guarantee that the energy during the iterative minimization decreases monotonically nor that the constraints are fulfilled before convergence. One solution is to run the convex optimization until convergence however in practice this leads to slow convergence. Therefore, we follow a different strategy where we regularly run the majorization step during the optimization of the energy. Before we can state the final algorithm we present the following lemma.

Lemma 3. *Given $\mathbf{x}^{(n)}, \mathbf{y}^{(n)}, \mathbf{z}^{(n)}$, in the optimization problem Eq. 15, which do not necessarily fulfill the constraints. A feasible solution $\tilde{\mathbf{x}}^{(n)}, \tilde{\mathbf{y}}^{(n)}, \tilde{\mathbf{z}}^{(n)}$ to the ray potential Eq. 11 and the regularization term Eq. 14 can be constructed in a finite number of steps.*

Proof. To fulfill the constraints $\sum_{\ell\in\mathcal{L}} x_s^\ell = 1$ and $x_s^\ell \in [0,1]$ we project the variables $\mathbf{x}^{(n)}$ individually per voxel s to the unit probability simplex [6]. Subsequent application of Lemma 1 and 2 leads to the desired result. □

Our final majorize-minimize optimization strategy can now be stated as follows.

Majorization step Using the current variables $\mathbf{x}^{(n)}, \mathbf{y}^{(n)}, \mathbf{z}^{(n)}$ and Lemma 3 a feasible solution $\tilde{\mathbf{x}}^{(n)}, \tilde{\mathbf{y}}^{(n)}, \tilde{\mathbf{z}}^{(n)}$ can be found. If the new energy is lower or equal than the last known energy the linearization Eq. 13 is applied and the new optimization problem is passed to the minimization step. Otherwise the whole majorization step is skipped and the old variables $\mathbf{x}^{(n)}, \mathbf{y}^{(n)}, \mathbf{z}^{(n)}$ are passed to the minimization step.

Minimization step The primal-dual algorithm [24] is run on the surrogate convex program for a fixed number of p iterations. For guaranteed convergence, the primal dual gap η can be evaluated and the minimization step can be restarted until we have $\eta \leq f(n)$ with a function $f(n) \to 0$ for $n \to \infty$. In practice, we get a good convergence behavior without a restart.

The majorization step either does no changes to the optimization state or finds a better or equal solution with a new linearization of the non-convex part because the linearization does not change the energy. If no changes are done this can be due to two reasons. Either the current solution was worse than the last known one, or the majorization stayed the same. In the latter case the primal-dual gap could reveal convergence and as a consequence we know that we arrived at a critical point or corner point of the original non-convex energy. In any other case the minimization step is run again and due to the convexity of the surrogate convex function a better solution or the optimality certificate will be found. The above procedure could stop at a non-critical point[1] due to the kink in the maximum in Eq. 12. To guarantee convergence to a critical point, the visibility consistency constraint Eq. 6 can be smoothed slightly *e.g.* using [23]. Again, in our experiments this was unnecessary to achieve good convergence behavior.

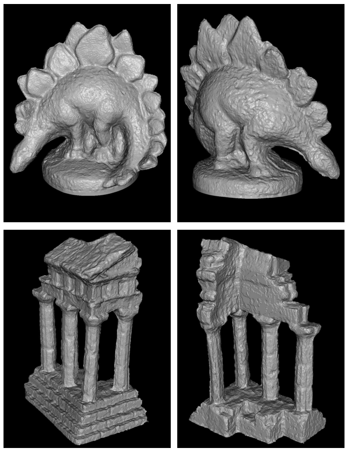
5. Experiments

Before we discuss the experiments, we describe the input data and state the costs c_{ri}^ℓ used for the ray potentials.

5.1. Input Data

We are using our approach for two different tasks: standard dense 3D reconstruction and dense semantic 3D reconstruction. In both cases, the initial input is a set of images with associated camera poses. Those camera poses are either provided with the dataset (as in the Middlebury Benchmark [27] or in the Thin Road Sign dataset [32]) or computed via structure from motion algorithm [3] (as in the semantic reconstruction experiments). We computed the depth maps using plane sweeping stereo for Middlebury Benchmark and semantic reconstruction datasets, while utilizing those already provided with the dataset for the experiment with Thin Road Sign. The patch similarity measure

[1] similar to block-coordinate descent based message passing algorithms



	Temple Full Acc / Comp	Temple Ring Acc / Comp	Temple S. Ring Acc / Comp	Dino Full Acc / Comp	Dino Ring Acc / Comp	Dino S. Ring Acc / Comp
Our Method	0.41 / 99.7	0.5 / 99.5	0.69 / 97.8	**0.26** / 99.8	**0.25** / 99.9	0.34 / 99.7
Galliani et al.[8]	0.39 / 99.2	0.48 / 99.1	0.53 / 97.0	0.31 / 99.9	0.3 / 99.4	0.38 / 98.6
Zhu et al.[39]		**0.4** / 99.2	**0.45** / 95.7		0.38 / 98.3	0.48 / 95.4
Li et al.[20]		0.73 / 98.2	0.66 / 97.3		0.28 / 100	**0.3** / 100
Wei et al. [34]	**0.34** / 99.4			0.42 / 98.1		
Xue et al.						**0.3** / 99.1
Furukawa et al. [7]	0.49 / 99.6	0.47 / 99.6	0.63 / 99.3	0.33 / 99.8	0.28 / 99.8	0.37 / 99.2

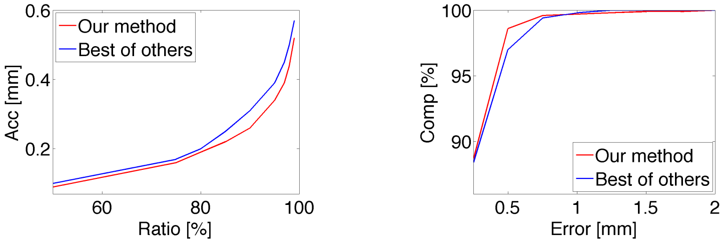


Figure 5: Middlebury Multi-View Stereo Benchmark: **(left)** Reconstructions of the Dino Ring and Temple Ring datasets computed by our algorithm, (left-most) view 1, (right-most) view 2, **(right, top)** benchmark's competitive methods in inverse chronological order (smaller Acc and higher Comp numbers are better), **(right, bottom)** Acc vs. Ratio (lower curve better) and Comp vs. Error (higher curve better) plots for the Dino Full dataset (for details on these plots see [27]).

for stereo matching was zero-mean normalized cross correlation. For the dense semantic 3D reconstruction experiments, we computed per-pixel semantic labels using [17], trained on the datasets from [1], [28] and [10].

5.2. Ray Potential Costs

In case of a two-label problem, there exists only one single label $\ell \in \mathcal{L}$. This allows us to directly insert the visibility consistency, Eq. 9, into the objective. In this case the majorization can directly be done on the objective instead of the visibility constraint, leading to a more compact optimization problem with a smaller memory footprint. Like [10], we assume exponential noise on the depth maps and define the assignments to the costs c^{ℓ}_{ri}, given the position of the depth measurement along the ray r as i' as

$$c^{\ell}_{ri} := \min\{0, \lambda|i - i'| - K\}. \tag{16}$$

The parameters $\lambda \geq 0$ and K are chosen such that the potential captures the uncertainty of the depth measurement.

For the multi-class case we also assume exponential noise on the depth data and independence between the depth measurement and the semantic measurement. Therefore the combined costs read as

$$c^{\ell}_{ri} := \min\{0, \lambda|i - i'| - K\} + \sigma^{\ell}, \tag{17}$$

with σ^{ℓ} being the response of the semantic classifier for the respective pixel. This is the same potential that [10] approximates with unary potentials.

5.3. Middlebury Multi-View Stereo Benchmark

We evaluate our method for dense 3D reconstuction on the Middlebury benchmark [27]. We ran our algorithm on all 6 datasets (using the same parameters). Two quantitative measures are defined in this benchmark paper: accuracy (Acc) and completeness (Comp). In terms of accuracy our algorithm sets a new state-of-the-art for the Dino Full and Dino Ring datasets (*c.f.* Fig. 5). An actual ranking of the benchmark is difficult because there is no default, commonly accepted, way to combine the two measures. Taking into account both measures we are close to the state-of-the-art on all datasets (results can be found online [2]).

5.4. Street Sign Dataset

A challenging case for volumetric 3D reconstruction are thin objects. When approximating the data term, which is naturally given as a ray potential in the 2.5D input data, by unary or pairwise potentials the data terms from both sides are prone to cancel out. Similarly, when using visual hulls a slight misalignment of the two sides might generate an empty visual hull. These are the reasons why thin objects are considered to be a hard case in volumetric 3D reconstruction. We evaluate the performance of our algorithm for such objects on the street sign dataset from [32]. The dataset consists 50 images of a street sign with corresponding depth maps. As depicted in Fig 6 the thin surface does not pose any problem to our method, thanks to an accurate representation of the input data in the optimization problem. To illustrate the result obtained with a standard volumetric 3D reconstruction algorithm we ran our implementation of the TV-Flux fusion from [35] on the same data. Note that this dataset is particularly hard because the two sides actually interpenetrate as detailed in [32].

[2] http://vision.middlebury.edu/mview/eval/

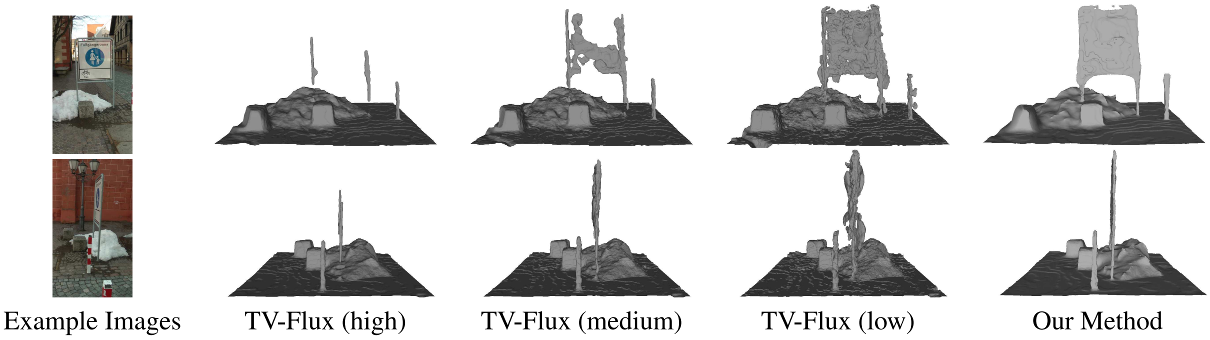


Figure 6: Reconstructions of the street sign dataset from [32] using the TV-Flux fusion from [35] with three different smoothness settings (high/medium/low), and our proposed method.

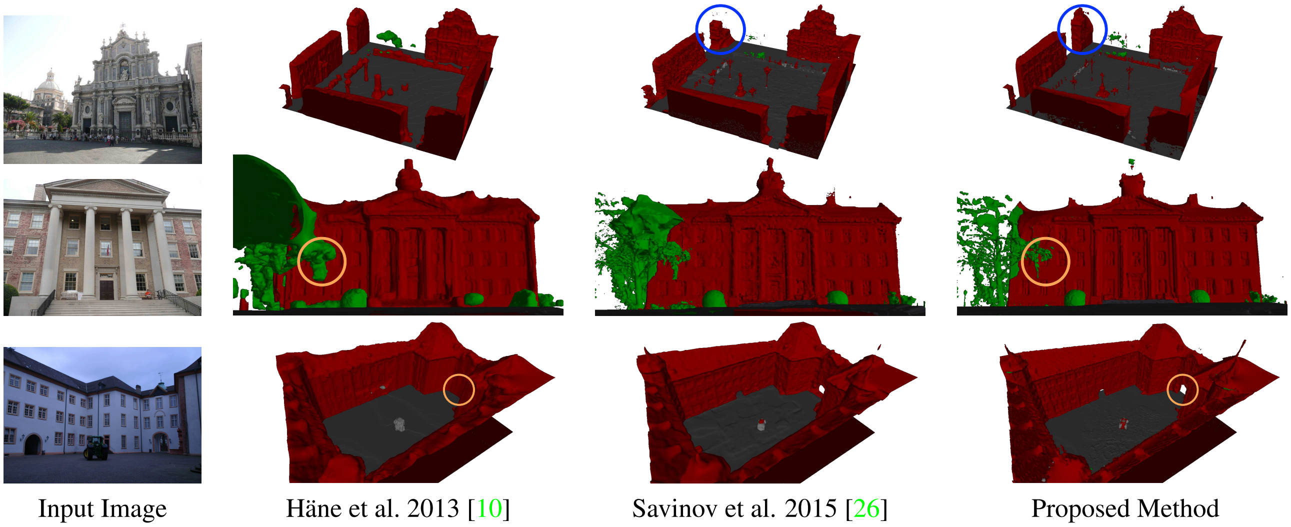


Figure 7: Semantic 3D Reconstructions: we improve in weakly observed areas and resolve unary potential artifacts at the same time. Five semantic labels are used: ground, building, vegetation, clutter, free space.

5.5. Multi-Label Experiments

We evaluate our formulation for dense semantic 3D reconstruction on several real-world datasets. We show our results side-by-side with the method of [10] and [26] in Figs. 7 and 1. Our method uses the same smoothness prior as [10]. For all the datasets we observe that the approximation of the data cost with a unary potential in [10] artificially fattens corners and thin objects (*e.g*. pillars or tree branches). In the close-ups (*c.f*. Fig. 1) we see that such a data term recovers significantly less surface detail with respect to our proposed method. This problem has been addressed in [26], but their discrete graph-based approach suffers from metrication artifacts, cannot be combined with the class-specific anisotropic smoothness prior and does not lead to smooth surfaces (*c.f*. Fig. 1). Moreover, their coarse-to-fine scheme produces artifacts in the reconstructions. Our approach takes the best of both worlds, the ray potential part ensures an accurate position of the observed surfaces, while the anisotropic smoothness prior faithfully handles weakly observed areas.

6. Conclusion

In this paper we proposed an approach for using ray potentials together with continuously inspired surface regularization. We demonstrated that a direct convex relaxation is too weak to be used in practice. We resolved this issue by adding a non-convex constraint to the formulation. Further, we detailed an optimization strategy and gave an extensive evaluation on two-label and multi-label datasets. Our algorithm allows for a general multi-label ray potential, at the same time it achieves volumetric 3D reconstruction with high accuracy. In semantic 3D reconstruction we are able to overcome limitations of earlier methods.

Acknowledgements: We thank Ian Cherabier for providing code for Lemma 2. This work is partially funded by the Swiss National Science Foundation projects 157101 and 163910. Ľubor Ladický is funded by the Max Planck Center for Learning Systems Fellowship.

References

[1] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *European Conference on Computer Vision*, 2008. 7

[2] A. Chambolle, D. Cremers, and T. Pock. A convex approach to minimal partitions. *SIAM Journal on Imaging Sciences*, 2012. 2, 5

[3] A. Cohen, C. Zach, S. N. Sinha, and M. Pollefeys. Discovering and exploiting 3D symmetries in structure from motion. In *Conference on Computer Vision and Pattern Recognition*, 2012. 6

[4] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *Conference on Computer graphics and interactive techniques*, 1996. 1, 2

[5] A. Delaunoy and E. Prados. Gradient flows for optimizing triangular mesh-based surfaces: Applications to 3D reconstruction problems dealing with visibility. *International Journal of Computer Vision (IJCV)*, 2011. 2

[6] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *International conference on Machine learning (ICML)*, 2008. 6

[7] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2010. 7

[8] S. Galliani, K. Lasinger, and K. Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *International Conference on Computer Vision*, 2015. 7

[9] P. Gargallo, P. Sturm, and S. Pujades. An occupancy - depth generative model of multi-view images. In *Asian Conference on Computer Vision (ACCV)*. 2007. 2

[10] C. Häne, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3D scene reconstruction and class segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2013. 1, 2, 5, 7, 8

[11] C. Hernandez, G. Vogiatzis, and R. Cipolla. Probabilistic visibility for multi-view stereo. In *Conference on Computer Vision and Pattern Recognition*, 2007. 2

[12] M. Klodt, T. Schoenemann, K. Kolev, M. Schikora, and D. Cremers. An experimental comparison of discrete and continuous shape optimization methods. In *Computer Vision–ECCV 2008*, pages 332–345. Springer, 2008. 1

[13] K. Kolev and D. Cremers. Integration of multiview stereo and silhouettes via convex functionals on convex domains. In *European Conference on Computer Vision (ECCV)*, 2008. 2

[14] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *European Conference on Computer Vision (ECCV)*. 2002. 2

[15] V. Kolmogorov, R. Zabih, and S. Gortler. Generalized multi-camera scene reconstruction using graph cuts. In *Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, 2003. 2

[16] P. Labatut, J.-P. Pons, and R. Keriven. Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In *IEEE International Conference on Computer Vision*, 2007. 2

[17] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical CRFs for object class image segmentation. In *International Conference on Computer Vision (ICCV)*, 2009. 7

[18] K. Lange, D. R. Hunter, and I. Yang. Optimization transfer using surrogate objective functions. *Journal of computational and graphical statistics*, 2000. 6

[19] V. S. Lempitsky and Y. Boykov. Global optimization for shape fitting. In *Conference on Computer Vision and Pattern Recognition*, 2007. 2

[20] Z. Li, K. Wang, W. Zuo, D. Meng, and L. Zhang. Detail-preserving and content-aware variational multi-view stereo reconstruction. *arXiv preprint arXiv:1505.00389*, 2015. 7

[21] S. Liu and D. B. Cooper. Ray Markov random fields for image-based 3D modeling: Model and efficient inference. In *Conference on Computer Vision and Pattern Recognition*, 2010. 2

[22] S. Liu and D. B. Cooper. Statistical inverse ray tracing for image-based 3D modeling. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014. 2

[23] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 2005. 6

[24] T. Pock and A. Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *2011 IEEE International Conference on Computer Vision (ICCV)*, 2011. 6

[25] T. Pollard and J. L. Mundy. Change detection in a 3-d world. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 2

[26] N. Savinov, L. Ladicky, C. Häne, and M. Pollefeys. Discrete optimization of ray potentials for semantic 3D reconstruction. In *Conference on Computer Vision and Pattern Recognition*, 2015. 1, 2, 3, 4, 8

[27] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'2006)*, volume 1, pages 519–526. IEEE Computer Society, June 2006. 6, 7

[28] J. Shotton, J. Winn, C. Rother, and A. Criminisi. *TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *European Conference on Computer Vision*, 2006. 7

[29] S. N. Sinha, P. Mordohai, and M. Pollefeys. Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh. In *International Conference on Computer Vision*, 2007. 2

[30] E. Strekalovskiy and D. Cremers. Generalized ordering constraints for multilabel optimization. In *International Conference on Computer Vision (ICCV)*, 2011. 5

[31] A. O. Ulusoy, A. Geiger, and M. J. Black. Towards probabilistic volumetric reconstruction using ray potentials. In *International Conference on 3D Vision (3DV)*, 2015. 2

[32] B. Ummenhofer and T. Brox. Point-based 3D reconstruction of thin objects. In *IEEE International Conference on Computer Vision (ICCV)*, 2013. 6, 7, 8

[33] H.-H. Vu, P. Labatut, J.-P. Pons, and R. Keriven. High accuracy and visibility-consistent dense multiview stereo. *Transactions on Pattern Analysis and Machine Intelligence*, 2012. 2

[34] J. Wei, B. Resch, and H. Lensch. Multi-view depth map estimation with cross-view consistency. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014. 7

[35] C. Zach. Fast and high quality fusion of depth maps. In *3D Data Processing, Visualization and Transmission*, 2008. 7, 8

[36] C. Zach, D. Gallup, J.-M. Frahm, and M. Niethammer. Fast global labeling for real-time stereo using multiple plane sweeps. In *International Workshop on Vision, Modeling and Visualization (VMV)*, 2008. 5

[37] C. Zach, C. Hane, and M. Pollefeys. What is optimized in convex relaxations for multilabel problems: Connecting discrete and continuously inspired MAP inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014. 2, 5

[38] C. Zach, T. Pock, and H. Bischof. A globally optimal algorithm for robust TV-L1 range image integration. In *International Conference on Computer Vision (ICCV)*, 2007. 2

[39] Z. Zhu, C. Stamatopoulos, and C. S. Fraser. Accurate and occlusion-robust multi-view stereo. *ISPRS Journal of Photogrammetry and Remote Sensing*, 109:47–61, 2015. 7