Semantic 3D Reconstruction with Finite Element Bases

Audrey Richard[1,†]
audrey.richard@geod.baug.ethz.ch

Christoph Vogel[2,†]
christoph.vogel@icg.tugraz.at

Maroš Bláha[1]
maros.blaha@geod.baug.ethz.ch

Thomas Pock[2,3]
thomas.pock@icg.tugraz.at

Konrad Schindler[1]
konrad.schindler@geod.baug.ethz.ch

[1] Photogrammetry & Remote Sensing
ETH Zurich, Switzerland

[2] Institute of Computer Graphics & Vision
TU Graz, Austria

[3] Austrian Institute of Technology

[†] shared first authorship

Abstract

We propose a novel framework for the discretisation of multi-label problems on arbitrary, continuous domains. Our work bridges the gap between general FEM discretisations, and labeling problems that arise in a variety of computer vision tasks, including for instance those derived from the generalised Potts model. Starting from the popular formulation of labeling as a convex relaxation by functional lifting, we show that FEM discretisation is valid for the most general case, where the regulariser is anisotropic and non-metric. While our findings are generic and applicable to different vision problems, we demonstrate their practical implementation in the context of semantic 3D reconstruction, where such regularisers have proved particularly beneficial. The proposed FEM approach leads to a smaller memory footprint as well as faster computation, and it constitues a very simple way to enable variable, adaptive resolution within the same model.

1 Introduction

A number of computer vision tasks, such as segmentation, multiview reconstruction, stitching and inpainting, can be formulated as *multi-label problems on continuous domains*, by functional lifting [7, 10, 24, 30, 32]. A recent example is semantic 3D reconstruction (*e.g.* [3, 15]), which solves the following problem: Given a set of images of a scene, reconstruct both its 3D shape and a segmentation into semantic object classes. The task is particularly challenging, because the evidence is irregularly distributed in the 3D domain; but it also possesses a rich, anisotropic prior structure that can be exploited. Jointly reasoning about shape and class allows one to take into account class-specific shape priors (*e.g.*, building walls should be smooth and vertical, and vice versa smooth, vertical surfaces are likely to be building walls), leading to improved reconstruction results. So far, models for the mentioned multi-label problems, and in particular for semantic 3D reconstruction, have been limited to axis-aligned discretisations. Unless the scenes are aligned with the coordinate

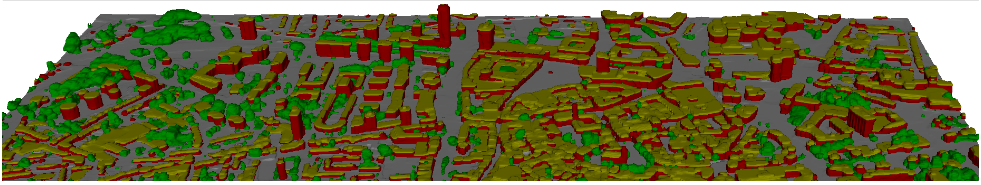

Figure 1: Semantic 3D model, estimated from aerial views with our FEM method.

axes, this leads to an unnecessarily large number of elements. Moreover, since the evidence is (inevitably) distributed unevenly in 3D, it also causes biased reconstructions. Thus, it is desirable to adapt the discretisation to the scene content (as often done for purely geometric surface reconstruction, *e.g.* [21]).

Our formulation makes it possible to employ a finer tesselation in regions that are likely to contain a surface, exploiting the fact that both high spatial resolution and high numerical precision are only required in those regions. Our discretisation scheme leads to a smaller memory footprint and faster computation, and it constitues a very simple technique to allow for arbitrary adaptive resolution levels within the same problem. *I.e.*, we can refine or coarsen the discretisation as appropriate, to adapt to the scene to be reconstructed. While our scheme is applicable to a whole family of finite element bases, we investigate two particularly interesting cases: Lagrange (P1) and Raviart-Thomas elements of first order. We further show that the grid-based voxel discretisation is a special case of our P1 basis, such that minimum energy solutions of "identical" discretisations (same vertex set) are equivalent.

2 Related Work

Since the seminal work [11] volumetric reconstruction from image data has evolved remarkably [9, 13, 17, 18, 19, 25, 39, 42]. Most methods use depth maps or 2.5D range scans for evidence [46, 47], represent the scene via an indicator or signed distance function in the volumetric domain, and extract the surface as its zero level set, *e.g.*, [26, 38].

Joint estimation of geometry and semantic labels, which had earlier been attempted only for single depth maps [22], has recently emerged as a powerful extension of volumetric 3D reconstruction from multiple views [1, 3, 15, 20, 35, 40, 41]. A common trait of these works is the integration of depth estimates and appearance-based labeling information from multiple images, with class-specific regularisation via shape priors.

Multi-label problems are in general NP-hard, but under certain conditions on the pairwise interactions, the original non-convex problem can be converted into a convex one via functional lifting and subsequent relaxation, *e.g.* [7]. This construction was further extended to anisotropic (direction-dependent) regularisers [37]. Moreover, [48] also relaxed the requirement that the regulariser forms a metric on the label set, yet its construction can only be applied after discretisation [24]. In this paper, we consider the relaxation in its most general form [48], but are not restricted to it. The latter construction is also the basis to the model of [15], whose energy model we adapt for our semantic 3D reconstruction method. Their voxel-based formulation can be seen as a special case of our discretisation scheme.

For (non-semantic) surface reconstruction, several authors prefer a data-dependent discretisation, normally a Delaunay tetrahedralisation of a 3D point cloud [16, 21, 43]. The occupancy states of the tetrahedra are found by discrete (binary) labeling, and the final surface is composed of the triangles that separate different labels. Loosely speaking, our proposed methodology can be seen either as an extension of [15] to arbitrary simplex partitions of the domain; or as an extension of [21] to semantic (multi-label) reconstruction.

We note that regular voxel partitioning of the volume leads to a high memory consumption and computation time. Yet, we are essentially reconstructing a 2D manifold in 3D space, and this can be exploited to reduce run-time and memory footprint. [20] use an octree instead of equally sized voxels to adapt to the local density of the input data. [3] go one step further and propose an adaptive octree, where the discretisation is refined on-the-fly, during optimisation. In our framework the energy is independent of the discretisation, it can thus be combined directly with such an adaptive procedure.

Also volumetric fusion via signed distance functions [28] benefits from irregular tesselations of 3D space, *e.g.*, octrees [36] or hashmaps [29]. In contrast to our work, these target real-time reconstruction and refrain from global optimisation, instead locally fusing depth maps. Their input normally is a densely sampled, overcomplete RGB-D video-stream, whereas we deal with noisy and incomplete inputs. To achieve high-quality reconstructions in our setting, we incorporate semantic information, leading to a multi-label problem.

Our work is based on the finite element method (FEM), *e.g.* [5, 33]. Introduced by Ritz [34] more than a century ago, and refined by Galerkin and Courant [8], FEM serves to numerically solve variational problems, by partitioning the domain into finite, parametrised elements. In computer vision FEM has been applied in the context of level-set methods [45] and for Total Variation [2]. To our knowledge, we are the first to apply it to multi-labeling.

3 Method

The multi-labeling problem [7, 24, 37, 48] in the domain $\Omega \subset \mathbb{R}^d$ is defined by finding m labeling functions $x^i : \Omega \rightarrow \{0,1\}, i = 1 \ldots m$ as the solution of:

$$\inf_{x^i} \sum_{i=1}^{m} \int_{\Omega} \rho^i(z) x^i(z) \mathrm{d}z + J(x^i), \quad \text{s. t. } \sum_{i=1}^{m} x^i(z) = 1 \; \forall z \in \Omega, \tag{1}$$

where ρ models the data term for a specific label at location $z \in \Omega$ and J denotes a convex regularisation functional that enforces the spatial consistency of the labels. One prominent example is to chose $J := \|\cdot\|_2$, known as Total Variation, which penalises the perimeter of the individual regions [7, 30]. Note that in the two-label case (Potts model), this relaxation is exact after thresholding with any threshold from the open unit interval [7]. Although we are ultimately interested in non-metric regularisation, we start with the continuous, anisotropic model [37], and postpone the extension to the non-metric case to Sec. 3.5.

3.1 Convex Relaxation

The continuous model allows for an anisotropic regulariser in J: label transitions can be penalised on the area of the shared surface, as well as on the surface normal direction. This is achieved with problem-specific 1-homogeneous functions that emerge from convex sets, so called Wulff-shapes. A relaxation of $x^i(z) \in \{0,1\}$ to $x^i(z) \in [0,1]$ then leads to a convex energy, which can be written as the following saddle point problem, with primal functions x and dual functions λ:

$$\min_{x^i} \max_{\lambda^i} \sum_i \int_{\Omega} \rho^i(z) x^i(z) + \langle x^i(z), \nabla \cdot \lambda^i(z) \rangle \mathrm{d}z, \text{ s.t. } \lambda^i(z) - \lambda^j(z) \in W^{ij}, \sum_{i=1}^{m} x^i(z) = 1, x^i(z) \geq 0. \tag{2}$$

The constraints have to be fulfilled for all $z \in \Omega$. In addition to the primal variables x^i, we have introduced the dual vector-field $\lambda^i : \mathbb{R}^d \rightarrow \mathbb{R}^d$, whose pairwise differences are constrained to lie in the convex sets (Wulff-shapes) W^{ij}. By letting these shapes take an anisotropic form, one can then encode scene structure, *e.g.* [15, 37]. For our problem we demand Neumann conditions at the boundary of Ω, *i.e.* $\langle \lambda^i, \nu \rangle = 0, \forall z \in \partial\Omega$, because the scene will continue beyond our domain (ν is the normal of the domain boundary $\partial\Omega$).

3.2 Finite Element Spaces

Here, we can only informally introduce the basic idea of FEM and explain its suitability for problems of the form (2). We refer to textbooks [5, 12, 23] for a deeper and formal treatment.

One way to solve (2) is to approximate it at a finite number of regular grid points, using finite differences. FEM instead searches for a solution in a finite-dimensional vector space; this *trial* space is a subspace of the space in which the exact solution is defined. To that end, one chooses a suitable basis for the *trial* space, with basis functions of finite support, as well as an appropriate *test* function space. FEM methods then find approximate solutions to variational problems by identifying the element from the *trial* space that is orthogonal to all functions of the *test* function space. For our saddle-point problem, we can instead identify the *trial* space with our primal function space and the *test* space with its dual counterpart. Now, we can apply the same principles, and after discretisation our solution corresponds to the continuous solution defined by the respective basis. As (2) is already a relaxation of the original problem (1), we do not present an analysis of convergence at this point. Instead, the reader is referred to [2] for an introduction to this somewhat involved topic.

In order to choose a space with good approximation properties and suitable basis functions, we tesselate our domain into simplices. More formally, we define $M = \{F, V, S\}$ to be a simplex mesh with vertices $v \in V, v \in \mathbb{R}^d$, faces $f \in F$ defined by d, and simplices $s \in S$, defined by $d+1$ vertices that partition Ω: $\cup_k s_k = \Omega, s_l \cap s_k = f_{l,k} \in F$ – *i.e.* two adjacent simplices share only a single face. In this work, for a specific set of vertices V, we select M to be the corresponding Delaunay tetrahedralisation of Ω and only consider explicit bases. In particular, we focus on the Lagrangian (P1) basis, which we use in the following to derive our framework; and on the Raviart-Thomas (RT) basis. Details for the latter are given in the supplementary material. The main difference between them is that P1 leads to piecewise linear solutions, which must be thresholded, while RT leads to a constant labeling function per simplex, similar to discrete MAP solutions on CRFs. We note that constant labeling can lead to artefacts, such that the adaptiveness of the FEM model becomes even more important.

The idea of both derivations is similar: *(i)* select a basis for our primal (P1) or dual (RT) variable set, *(ii)* find a suitable form via the divergence theorem and Fenchel duality, *(iii)* extend to the non-metric case, following a principle we term ”label mass preservation”.

3.3 Lagrange Elements

The Lagrange $\mathrm{P}^k(M)$ basis functions describe a *conforming* polynomial basis of order $k+1$ on our simplex mesh M, *i.e.* its elements belong to the Hilbert space of differentiable function with finite Lebesgue measure on the domain Ω: $\mathrm{P}^k(M) \subset H^1(\Omega) := \{p \in L^2(\Omega), \nabla p \in (L^2(\Omega)^d)\}$. We are interested in the Lagrange basis of first order, $\mathrm{P}^1(M)$:

$$P^1(M) := \{p : \Omega \to \mathbb{R} | p \in C(\Omega),\, p(x) := \sum_{s \in S} c_s^\top x + d_s, c_s \in \mathbb{R}^d, d_s \in \mathbb{R},\ \text{if } x \in s \text{ and } 0 \text{ else}\}. \quad (3)$$

We construct our linear basis with functions that are defined for each vertex v of a simplex s and can be described in local form with barycentric coordinates:

$$p^1_{s,v}(x) := \alpha_v \text{ with } x = \sum_{v \in s} \alpha_v v,\ \sum_{v \in s} \alpha_v = 1,\ \alpha_v \geq 0 \quad \text{if } x \in s \text{ and } 0 \text{ else}. \quad (4)$$

In each simplex, one can define a scalar field $\phi_s(x) \in \mathbb{R}$ and compute a gradient in this basis that will be constant per simplex s (*cf.* Fig. 2):

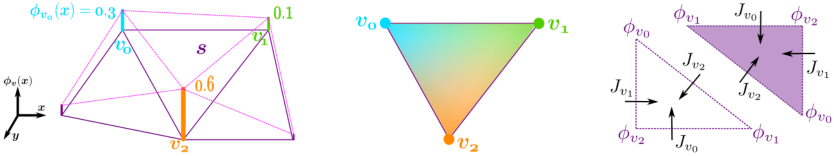

Figure 2: *Left:* Illustration of P1 basis function shape. *Middle:* Scalar field defined as a convex combination of basis coefficients. *Right:* Gradient definition in a simplex (5).

$$\phi_s(x) := \sum_{v\in s} \phi_v p^1_{s,v} \text{ and } \nabla\phi_s = \sum_{v\in s} \phi_v J_v, \tag{5}$$

with coefficients $\phi_v \in \mathbb{R}$. $J_v \in \mathbb{R}^d$ denotes a vector that is normal to the face f_v opposite node v, has length $\frac{|f_v|}{|s|d}$, and points towards the simplex centre. $|f_v|$ denotes the area of the face f_v, and $|s|$ the volume of the simplex s (*cf*. Fig. 2 and supplementary material).

3.4 Discretisation

To apply our Lagrange basis to (2) we first make use of the divergence theorem:

$$\int_\Omega \langle x^i(z), \nabla\cdot\lambda^i(z)\rangle \mathrm{d}z = \int_\Omega \langle \nabla x^i(z), \lambda^i(z)\rangle \mathrm{d}z - \underbrace{\int_{\partial\Omega} \langle \nu(z), \lambda^i(z)\rangle \mathrm{d}z}_{=0}. \tag{6}$$

The latter summand vanishes by our choice of λ. Our approach for a discretisation in the Lagrange basis is to choose the labeling function $x^i \in P^1(M)$. This implies that our dual space consists of constant vector-fields per simplex: $\lambda^i_s \in \mathbb{R}^d$. To fulfill the constraint set in (2) we have to verify that, per simplex, the λ^i_s lie in the respective Wulff-shape. The simplex constraints on the x^i have to be modeled per vertex. According to (4), the labeling functions are convex combinations of their values at the vertices and thus stay within the simplex.

We also have to convert the continuous data costs ρ^i into a cost per vertex ρ^i_v, which can be achieved by convolving the continuous cost with the respective basis function: $\rho^i_v := \int_\Omega \sum_{s\in\mathcal{N}(v)} \phi_s(x)\rho^i(x)\mathrm{d}x$. In practice, the integral can be computed by sampling ρ. Integrating the right hand side in (6) over the simplex s leads to a weighting with its volume $|s|$ and the energy (2) in the discrete setting becomes:

$$\min_{x^i} \max_{\lambda^i} \sum_{v,i} \rho^i_v x^i_v + \sum_{s,i} |s| \langle \nabla x^i, \lambda^i_s\rangle \text{ s.t. } (\lambda^i_s - \lambda^j_s) \in W^{ij}\ \forall i<j,\ s\in S, \sum_{i=1}^m x^i_v = 1,\ x^i_v \geq 0\ \forall v\in V. \tag{7}$$

3.5 Non-metric extension

To start with, we note that a non-metric model does not exist in the continuous case [27] and our extension works only *after* the discretisation into the FEM basis. Please refer to the supplementary material for an in-depth discussion. Note that our label set of semantic classes does not have a natural order (in contrast to, *e.g.*, stereo depth or denoised brightness); and also the direction-dependent regulariser is unordered and does not induce a metric cost. To allow for non-metric regularisation we transform the constraint set ($\lambda^i - \lambda^j \in W^{ij}$), by introducing auxiliary variables z^{ij} and Lagrange multipliers y^{ij}, and use Fenchel-Duality:

$$\max_{\lambda^i_s, z^{ij}_s} \min_{y^{ij}_s} \sum_{i<j} \langle (\lambda^i_s - \lambda^j_s) - z^{ij}, y^{ij}_s \rangle - \delta_{W^{ij}}(z^{ij}_s) = \max_{\lambda^i_s} \min_{y^{ij}_s} \sum_{i<j} -\langle (\lambda^i_s - \lambda^j_s), y^{ij}_s \rangle + ||y^{ij}_s||_{W^{ij}}. \tag{8}$$

The dual functions of the indicator functions for the convex sets W^{ij} are 1-homogeneous, of the form $||\cdot||_{W^{ij}} := \sup_{w\in W^{ij}} w^\mathsf{T}\cdot$. Recall that our label costs are not metric: $\forall i<j<k$:

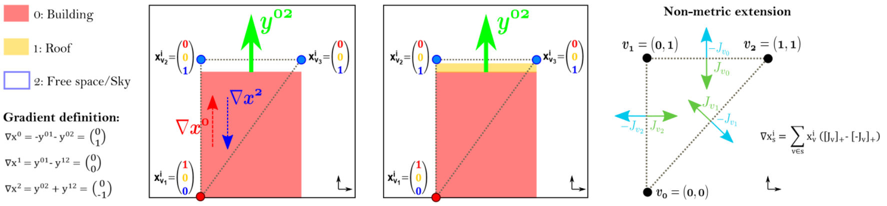

Figure 3: *Left:* Without our non-metric extension, optimisation w.r.t. (7) can lower transition costs by inserting another label (here 1 between 0 and 2). *Right:* A solution is to split the gradients of the indicator functions and use direction-dependent variables x^{ij}.

$|y^{ij}|_{W^{ij}} + |y^{jk}|_{W^{jk}} \geq |y^{ik}|_{W^{ik}}$, does not hold. It was shown [48] that a regulariser of the form (8) transforms any non-metric cost to the metric case. Figure 3 shows an example. Here, an expensive transition between labels 0 and 2 will be replaced by two cheaper transitions 0–1 and 1–2. To prevent this, we replace the y^{ij} with direction dependent variables x^{ij}: We rearrange (8) and combine the first summand with the regulariser from (7) to arrive at the following equations (for now ignoring $|s|$):

$$\sum_s \sum_i \langle \nabla x^i, \lambda^i_s \rangle - \sum_i \langle \lambda^i_s, \sum_{j \neq i} \left(y^{ij}_s [i < j] - y^{ji}_s [i > j] \right) \rangle,$$

with $[\cdot]$ denoting Iverson brackets. Let $x^{ij} := [y^{ij}]_+$ and $x^{ji} := [-y^{ji}]_+$, where $[\cdot]_+ := \max(0, \cdot)$. Expanding the gradient (5) we get, per simplex s,

$$\sum_i \sum_{v \in s} \lambda^i_s x^i_v ([J_v]_+ - [-J_v]_+) - \langle \lambda^i_s, \sum_{j : i \neq j} (x^{ij} - x^{ji}) \rangle, \tag{9}$$

which we analyse further to achieve non-metric costs. It was observed in [48] that the $x^{ij} \in \mathbb{R}^d$ can be interpreted as encoding the "label mass" that transitions from label i to label j in a specific direction. Positivity constraints (by definition) on the x^{ij} avoid the transport of negative label mass. To anchor transport of mass on the actual mass of label i present at a vertex, we introduce the variables x^{ii} for the mass that remains at label i, and split the above constraints into two separate sets with the help of additional dual variables θ:

$$\lambda^i_s \Big(\sum_{v \in s} x^i_v [J_v]_+ - \sum_j x^{ij}_s \Big) + \theta^i_s \Big(\sum_{v \in s} x^i_v [-J_v]_+ - \sum_j x^{ji}_s \Big) + \sum_{i,j} \delta_{\geq 0}(x^{ij}_s). \tag{10}$$

Note that this construction is only possible because our elements (simplices) are of strictly positive volume, in contrast to zero sets in Ω w.r.t. the Lebesgue measure. Finally, we can write down our discrete energy in the Lagrange basis defined on the simplex mesh M:

$$\begin{aligned} \min_{x^i, x^{ij}} \max_{\lambda^i, \theta^i} & \sum_{v \in V} \sum_i \rho^i_v x^i_v + \delta_\Delta(x^i_v) + \sum_{i<j} \sum_{s \in S} |s| \| x^{ij}_s - x^{ji}_s \|_{W^{ij}} + \\ & \sum_s \sum_i \theta^i_s \Big(\sum_{v \in s} x^i_v [-J_v]_+ - \sum_j x^{ji}_s \Big) + \sum_s \sum_i \lambda^i_s \Big(\sum_{v \in s} x^i_v [J_v]_+ - \sum_j x^{ij}_s \Big) + \sum_{i,j} \delta_{\geq 0}(x^{ij}_s), \end{aligned} \tag{11}$$

where we have moved the weighting with $|s|$ from the constraint set to the regulariser, and denote by $\delta_\Delta(\cdot)$ the indicator function of the unit simplex.

4 Semantic Reconstruction Model

A prime application scenario for our FEM multi-label energy model (11) is 3D semantic reconstruction. In particular, we focus on an urban scenario and let our labeling functions

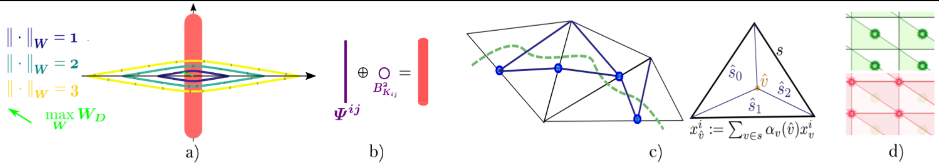

Figure 4: *(a):* Wulff-shape (*red*) with isolines. *(b):* Minkowski sum of two Wulff-shapes. *(c):* Simplices are split after inserting new vertices (blue) close to the surface (green). *Right:* Initialisation of vertices after refinement. *(d):* Finite differences on a regular grid ([15]) only cover constraints in *green* areas, the P1 basis covers all of the domain Ω.

encode *freespace* ($i = 1$), *building wall*, *roof*, *vegetation* or *ground*. Objects that are not explicitly modeled are collected in an extra *clutter* class. We define the data cost ρ at a 3D-point $x \in \Omega$ as in [15]: project x into the camera views $c \in \mathcal{C}$, and retrieve the corresponding depth $\hat{d}_c(x)$ and class likelihoods $\sigma^i_c(x)$ from the image space. The σ^i are obtained from a MultiBoost classifier. For the depth we look at the difference between the actual distance $d_c(x)$ to the camera and the observed depth: $d(x,c) := d_c(x) - \hat{d}_c(x)$. For the *freespace* label we always set the cost to 0, for $i \neq 1$ we define:

$$\rho^i(x) := \sum_{c \in \mathcal{C}} \sigma^i_c(x)[(k-1)\varepsilon \leq d(x,c) \leq k\varepsilon] + \beta[|d(x,c)| \leq k\varepsilon]\operatorname{sign}(d(x,c)). \qquad (12)$$

This model assumes independence of the per-pixel observations, and exponentially distributed inlier noise in the depthmaps, bounded by a parameter $k\varepsilon$ (k=3 in practice). It is essentially a continuous version of [15], see that paper for details. The parameter ε sets a lower bound for the minimal height of the simplices in the tesselation, and thus defines the target resolution. The discretisation of the data cost involves a convolution with the respective basis functions, which can be approximated via sampling. Please refer to the supplementary material for details. The Wulff-shapes W^{ij} in (11) are given as the Minkowski sum of the L_2-Ball, $B^2_{\kappa^{ij}} := \{x \in \mathbb{R}^d \,|\, \|x\|_2 \leq \kappa^{ij}\}$ and an anisotropic shape Ψ^{ij}: $W^{ij} := \Psi^{ij} \oplus B^2_{\kappa^{ij}}$. In the isotropic part, κ^{ij} contains the neighbourhood statistics of the classes. The anisotropic part Ψ^{ij} models the likelihood of a transition between classes i and j in a certain direction. Fig. 4 (a,b) shows an example. For our case we prefer flat, horizontal surfaces at the following label transitions: *ground-freespace*, *ground-building*, *building-roof*, *ground-vegetation* and *roof-freespace*. A second prior prefers vertical boundaries for the transitions *building-freespace* and *building-vegetation*. More details on the exact form can be found in [15].

The energy (11) is already in primal-dual form, such that we can apply the minimisation scheme of [6], with pre-conditioning [31]. That numerical scheme requires us to project onto shapes that are Minkowski sums of convex sets. In our case, the sets are simple and the projection onto each shape can be performed in closed form. We employ a Dykstra-like projection scheme [4], which avoids storing additional variables and proves remarkably efficient, see supplementary material. We also project the labeling functions x^i directly onto the unit simplex [44]. In order to extract the transition surface, we employ a variant of marching tetrahedra (triangles) [38], using the isolevel at 0.5 for each label.

We conclude with two interesting remarks. First, note that a tesselation with a regular grid [15] can be seen as a simplified version of our discretisation in the P^1 Lagrange basis. In Fig. 4d we consider the 2D case of the regular grid used in [15]. Here, variables are defined at voxel level. In its dual graph, the vertex set consists of the corners of the primal grid cubes, leading to shifted indicator variables. Per vertex the data term is mainly influenced from the cost in its Voronoi area. Similarly, [15] evaluates the data cost at grid centers, approximately

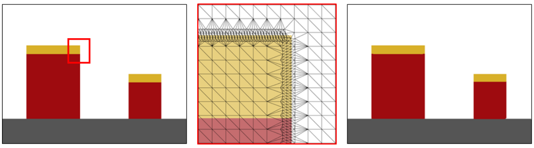

Figure 5: *Left:* Synthetic 2D scene, colors indicate ground (*gray*), building (*red*) and roof (*yellow*). *Middle:* Control mesh. *Right:* Example reconstruction.

overall acc. [%]	Tetra	Octree	MB
Scene 1	84.0	83.9	82.5
Scene 2	92.5	92.8	89

Table 1: Quantitative comparison with octree model [3] and MultiBoost input data.

corresponding to integration within the respective Voronoi-area of grid cell. Furthermore, taking finite differences in this regular case corresponds to verifying constraints for only one of the two triangles (Fig. 4d). The supplementary material includes a more formal analysis.

Second, our formulation is adaptive, in the sense of [3]: Hierarchical refinement of the tesselation can only decrease our energy. Hence, our scheme is applicable when refining the model on-the-fly. We again must defer a formal proof to the supplementary material and give an intuitive, visual explanation in Fig. 4c . Assume that $(x^*, \lambda^*, \theta^*)$ is an optimal solution for a certain triplet $M = \{F, V, S\}$. Then a refined tesselation $\hat{M} = \{\hat{F}, \hat{V}, \hat{S}\}$ can be found by introducing additional vertices, *i.e.* $V \subset \hat{V}$ (ideally on the label transition surfaces). To define a new set of simplices, we demand that no faces are flipped, $\forall \hat{s} \in \hat{S}, \exists s \in S$ with $\hat{s} \cap s = \hat{s}$. Then one can find a new variable set and data cost $\hat{\rho}$ with the same energy: We initialise the new variables from the continuous solution at the respective location, and find new $\rho_{\hat{v}}$ by integration. Subsequent minimisation in the refined mesh can only decrease the energy. The argument works in both ways: Vertices that have the same solution as their adjacent neighbors can be removed without changing the energy. For now we stick to this simple scheme, future work might explore more sophisticated ideas, *e.g.* along the lines of [14].

5 Evaluation

Before we present results on challenging real 3D data we evaluate our method in 2D on a synthetic dataset. All results are obtained with a multi-core implementation, on a 12-core, 3.5 GHz machine with 64GB RAM. For clarity, we only present the Lagrange discretisation. We refer to the supplementary material for an evaluation of the Raviart-Thomas discretisation.

Input Data. We create a synthetic 2D scene composed of 4 labels: *free space, building, ground* and *roof*, surrounded by 17 virtual cameras. To replace depth maps and class-likelihood images, we extract 2D points on the boundary "surface" and assign ground truth label costs to each point. For the evaluation in 3D, we use three real-world aerial mapping data sets. Our method requires two types of input data: depthmaps and pixel-wise class probabilities (*cf*. Sec. 4). Moreover, we build a *control mesh M* around the initially predicted surface, to facilitate our FEM discretisation. Ideally, the *control mesh* enwraps the true surface, using a finer meshing close to it. We densely evaluate the data cost at the vertices of a regular *data cost grid* and let each *control vertex* accumulate the cost of its nearest neighbours in that grid, to approximate an integration over its Voronoi cell.

2D Lagrange results. Fig. 5 illustrates the result we obtain in a *perfect* setting. The original 2D image serves as ground truth for our quantitative evaluation. In this baseline setting, our method achieves 99.8% of *overall accuracy* and 99.7% of *average accuracy*, confirming the soundness of our Lagrange discretisation. In order to evaluate how our model behaves in a more realistic setting, we conduct a series of experiments where we incrementally add different types of perturbations. Our algorithm is tested against: *(i)* noise in the initial 2D point cloud, respectively depth maps, *(ii)* wrong class probabilities and *(iii)* ambiguous class

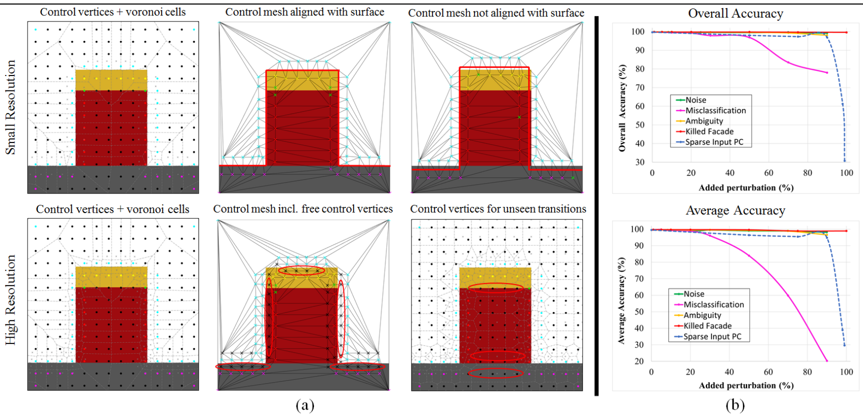

Figure 6: (a) Illustration of the control mesh foundation. Dots represent values of the datacost grid and crosses the *control vertices*. Voronoi cells of the *control vertices* are depicted with dashed grey lines and the *control mesh* with a solid black line. Colors indicate ground (*purple*), building (*red*), roof (*yellow*), free space (*cyan*) and no datacost (*black*). (b) Quantitative evaluation of Lagrange FEM method w.r.t. different degradations of the input data.

probabilities of random subsets, *(iv)* missing data, *e.g.* deleting part of a facade to simulate unobserved areas, *(v)* sparsity of the initial point cloud. Fig. 6b illustrates the influence of defective inputs. Under reasonable assumptions on the magnitude of the investigated perturbations, we do not observe a significant loss in accuracy. The reconstruction quality starts to decrease if more than half of the input data is misclassified or if the input point cloud is excessively sparse, meaning that >50% of the input is wrong or nonexistent. Average accuracy is naturally more sensitive, due to the larger relative error in small classes.

Influence of the control mesh. Recall from (12) that the data cost of a *control vertex* $v \in V$ is approximately equal to an integral of ρ over its respective Voronoi area (*cf*. Fig. 6a, left). Therefore, but also because of the sign change in (12), vertices close to the surface receive small cost values and are mainly steered by the regulariser, *i.e.* these vertices realise a smooth surface. On the other hand, vertices that integrate only over areas with positive or negative sign determine the inside/outside decision, but are more or less agnostic about the exact location of the surface. We conclude that a sufficient amount of *control vertices* should lie within the band $[\hat{d} - 3\varepsilon; \hat{d} + 3\varepsilon]$ defined by the truncation of the cost function around the observed depth $\hat{d}$ (*cf*. (12) and [15]). Ideally *control vertices* are equally distributed along each line-of-sight in front and behind the putative match (*cf*. Fig. 6a, middle column). Undersampling within the near-band can lead to smooth, but inaccurate results (*cf*. Fig. 6a, top right). Unobserved transitions, *e.g. building-ground* or *roof-building*, can also lead to problems if the affected simplices are too large. To mitigate the effect, we add a few vertices (*e.g.*, a sparse regular grid) on top of the *control mesh* (*cf*. Fig. 6a, bottom row). Finally, oversampling each line-of-sight in order to increase the resolution of the *control mesh* is not recommended, the right spacing is determined by the noise level and ε and k, chosen in (12).

To conclude, it is an important advantage of the FEM framework that additional vertices can be inserted as required, without changing the energy. In future work we will use this flexibility to develop smarter *control meshes*, possibly as a function of the *local* noise level.

3D Lagrange results. To test our algorithm on real world data, we focus on a dataset from the city of Enschede. Complementary results for other datasets are shown in the supplementary material. As baseline we use [3], the current state-of-the-art in large-scale semantic

3D reconstruction. Due to the lack of 3D ground truth, we follow their evaluation protocol and back-project subsets of the 3D reconstruction to image space, where it is compared to a manual segmentation. As can be seen in Fig. 7 and Tab. 1, the two results are similar in terms of quantitative correctness. We note that measuring labeling accuracy in the 2D projection does not consider the geometric quality of the reconstruction within regions of a single label.

Figs. 1 and 7 show city-modelling results obtained from (nadir and oblique) aerial images. Visually, our models are crisper and less "clay-like". Compared to axis-aligned discretisation schemes, *e.g*. [3, 15], our method appears to better represent surfaces not aligned with the coordinate axis, and exhibit reduced grid aliasing. Both effects are consistent with the main strength of the FEM framework, to adapt the size *and* the orientation of the volume elements to the data. Small tetrahedra, and vertices that coincide with accurate 3D points on surface discontinuities, favour sharp surface details and crease edges (*e.g*., substructures on roofs). Faces that follow the data points rather than arbitrary grid directions mitigate aliasing on surfaces not aligned with the coordinate axes (*e.g*., building walls). The freedom of a local *control mesh* unleashes the power of the regulariser in regions where the evidence is weak or ambiguous (*e.g*., roads, weakly textured building parts).

As already mentioned, our FEM framework can be readily combined with on-the-fly adaptive computation, as used in the baseline [3]. Compared to their voxel/octree model, adaptive refinement is straight-forward, due to the flexibility of the FEM framework, which allows for the introduction of arbitrary new vertices. As a preliminary proof-of-concept, we have tested the naive refinement scheme described in Sec. 4. We execute three refinement steps, where we repeatedly reconstruct the scene and subsequently refine simplices that contain surface transitions, while lowering ε by half. Compared to computing everything at the final resolution, this already yields substantial savings of 89–97% in memory and 82–93% in computation time, without any loss in accuracy. Targetting $\varepsilon \geq \frac{1}{\sqrt{3}}$ (measured w.r.t. a bounding box of 256 units), the runtimes for the tested scenes are 1h04m–1h47m and memory consumption is 573–764 MB, on a single machine.

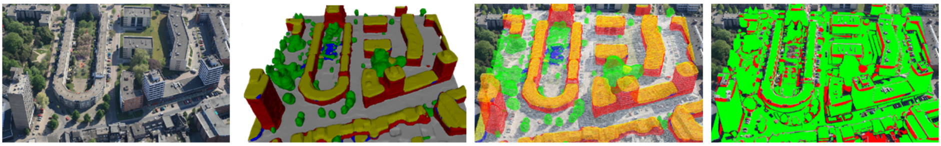

Figure 7: Quantitative evaluation of Scene 1 from Enschede. *Left:* One of the input images. *Middle left:* Semantic 3D model. *Middle right:* Back-projected labels overlayed on the image. *Right:* Error map, misclassified pixels are marked in red.

6 Conclusion

We have proposed a novel framework for the discretisation of multi-label problems, and have shown that, in the context of semantic 3D reconstruction, the increased flexibility of our scheme allows one to better adapt the discretisation to the data at hand. Our basic idea is generic and not limited to semantic 3D reconstruction or the specific class of regularisers. We would like to explore other applications where it may be useful to abandon grid discretisations and move to a decomposition into simplices.

Acknowledgements: Audrey Richard was supported by SNF grant 200021_157101. Christoph Vogel and Thomas Pock acknowledge support from the ERC starting grant 640156, 'HOMOVIS'.

References

[1] Yingze Bao, Manmohan Chandraker, Yuanqing Lin, and Silvio Savarese. Dense object reconstruction using semantic priors. In *CVPR*, 2013.

[2] Sören Bartels. Total variation minimization with finite elements: Convergence and iterative solution. *SIAM*, 50(3):1162–1180, 2012.

[3] Maros Blaha, Christoph Vogel, Audrey Richard, Jan Dirk Wegner, Thomas Pock, and Konrad Schindler. Large-scale semantic 3d reconstruction: An adaptive multiresolution model for multi-class volumetric labeling. In *CVPR*, 2016.

[4] J. P. Boyle and R. L. Dykstra. A method for finding projections onto the intersection of convex sets in Hilbert spaces. *Lecture Notes in Statistics*, 1986.

[5] Franco Brezzi and Michel Fortin. *Mixed and Hybrid Finite Element Methods*. Springer-Verlag New York, 1991.

[6] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *JMIV*, 40(1):120–145, 2011.

[7] Antonin Chambolle, Daniel Cremers, and Thomas Pock. A convex approach to minimal partitions. *SIAM*, 5(4):1113–1158, 2012.

[8] R. Courant. Variational methods for the solution of problems of equilibrium and vibrations. *Bull. Amer. Math. Soc.*, 49(1):1–23, 01 1943.

[9] D. Cremers and K. Kolev. Multiview stereo and silhouette consistency via convex functionals over convex domains. *PAMI*, 33(6):1161–1174, 2011.

[10] D. Cremers, T. Pock, K. Kolev, and A. Chambolle. Convex Relaxation Techniques for Segmentation, Stereo and Multiview Reconstruction. In *Advances in Markov Random Fields for Vision and Image Processing*. MIT Press, 2011.

[11] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. *SIGGRAPH*, 1996.

[12] Ricardo G. Durán. *Mixed Finite Element Methods*, pages 1–44. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[13] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multi-view stereopsis. *PAMI*, 2010.

[14] Eitan Grinspun, Petr Krysl, and Peter Schröder. CHARMS: A Simple Framework for Adaptive Simulation. *SIGGRAPH*, 2002.

[15] Christian Häne, Christopher Zach, Andrea Cohen, Roland Angst, and Marc Pollefeys. Joint 3d scene reconstruction and class segmentation. In *CVPR*, 2013.

[16] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *CVPR*, 2011.

[17] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *EUROGRAPHICS*, 2006.

[18] K. Kolev, T. Brox, and D. Cremers. Fast joint estimation of silhouettes and dense 3D geometry from multiple images. *PAMI*, 2012.

[19] Ilya Kostrikov, Esther Horbert, and Bastian Leibe. Probabilistic labeling cost for high-accuracy multi-view reconstruction. In *CVPR*, 2014.

[20] Abhijit Kundu, Yin Li, Frank Dellaert, Fuxin Li, and James M. Rehg. Joint semantic segmentation and 3d reconstruction from monocular video. In *ECCV*, 2014.

[21] Patrick Labatut, Jean-Philippe Pons, and Renaud Keriven. Efficient Multi-View Reconstruction of Large-Scale Scenes using Interest Points, Delaunay Triangulation and Graph Cuts. In *ICCV*, 2007.

[22] L'ubor Ladický, Paul Sturgess, Christopher Russell, Sunando Sengupta, Yalin Bastanlar, William Clocksin, and Philip Torr. Joint optimisation for object class segmentation and dense stereo reconstruction. In *BMVC*, 2010.

[23] Mats G. Larson and Fredrik Bengzon. *The Finite Element Method: Theory, Implementation, and Applications*. Springer Publishing Company, Incorporated, 2013. ISBN 3642332862, 9783642332869.

[24] Jan Lellmann and Christoph Schnörr. Continuous multiclass labeling approaches and algorithms. *SIIMS*, 4(4):1049–1096, 2011.

[25] Shubao Liu and David B. Cooper. Ray Markov random fields for image-based 3d modeling: Model and efficient inference. In *CVPR*, 2010.

[26] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *SIGGRAPH*, 1987.

[27] F. Maggi. *Sets of Finite Perimeter and Geometric Variational Problems: An Introduction to Geometric Measure Theory*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2012.

[28] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality*, ISMAR, pages 127–136, Washington, DC, USA, 2011. IEEE Computer Society. ISBN 978-1-4577-2183-0. doi: 10.1109/ISMAR.2011.6092378.

[29] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Trans. Graph.*, 32(6):169:1–169:11, November 2013. ISSN 0730-0301. doi: 10.1145/2508363.2508374.

[30] Claudia Nieuwenhuis, Eno Töppe, and Daniel Cremers. A survey and comparison of discrete and continuous multi-label optimization approaches for the potts model. *IJCV*, 104(3):223–240, 2013.

[31] Thomas Pock and Antonin Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *ICCV*, 2011.

[32] Thomas Pock, Daniel Cremers, Horst Bischof, and Antonin Chambolle. Global solutions of variational models with convex regularization. *SIIMS*, 3(4):1122–1145, 2010.

[33] J. Reddy. *An Introduction to the Finite Element Method.* McGraw-Hill Education, 2005. ISBN 9780072466850.

[34] Walter Ritz. Über eine neue methode zur lösung gewisser variationsprobleme der mathematischen physik. *Journal für die reine und angewandte Mathematik*, pages 1–61, 1909.

[35] Nikolay Savinov, L'ubor Ladický, Christian Häne, and Marc Pollefeys. Discrete optimization of ray potentials for semantic 3d reconstruction. In *CVPR*, 2015.

[36] F. Steinbruecker, J. Sturm, and D. Cremers. Volumetric 3d mapping in real-time on a cpu. In *ICRA*, 2014.

[37] E. Strekalovskiy and D. Cremers. Generalized ordering constraints for multilabel optimization. In *ICCV*, 2011.

[38] G. Treece. Regularised marching tetrahedra: improved iso-surface extraction. *Computers & Graphics*, 23(4):583–598, August 1999. ISSN 00978493.

[39] Ali Osman Ulusoy, Michael J. Black, and Andreas Geiger. Patches, planes and probabilities: A non-local prior for volumetric 3D reconstruction. In *CVPR*, 2016.

[40] Ali Osman Ulusoy, Michael J. Black, and Andreas Geiger. Semantic multi-view stereo: Jointly estimating objects and voxels. In *CVPR*, 2017.

[41] Vibhav Vineet, Ondrej Miksik, Morten Lidegaard, Matthias Nießner, Stuart Golodetz, Victor A. Prisacariu, Olaf Kähler, David W. Murray, Shahram Izadi, Patrick Perez, and Philip H. S. Torr. Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. In *ICRA*, 2015.

[42] George Vogiatzis, Carlos Hernández Esteban, Philip H. S. Torr, and Roberto Cipolla. Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *PAMI*, 29(12):2241–2246, 2007.

[43] H. H. Vu, P. Labatut, J. P. Pons, and R. Keriven. High accuracy and visibility-consistent dense multiview stereo. *PAMI*, 34(5):889–901, May 2012.

[44] Weiran Wang and Miguel Á. Carreira-Perpiñán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *CoRR*, abs/1309.1541, 2013.

[45] Martin Weber, Andrew Blake, and Roberto Cipolla. Sparse finite elements for geodesic contours with level-sets. In Tomás Pajdla and Jiří Matas, editors, *ECCV*, 2004.

[46] C. Zach, T. Pock, and H. Bischof. A globally optimal algorithm for robust TV-L1 range image integration. In *ICCV*, 2007.

[47] Christopher Zach. Fast and high quality fusion of depth maps. *3DV*, 2008.

[48] Christopher Zach, Christian Häne, and Marc Pollefeys. What is optimized in convex relaxations for multilabel problems: Connecting discrete and continuously inspired MAP inference. *PAMI*, 2014.