A Data-driven Prior on Facet Orientation for Semantic Mesh Labeling

Andrea Romanoni
Politecnico di Milano, Italy
andrea.romanoni@polimi.it

Matteo Matteucci
Politecnico di Milano, Italy
matteo.matteucci@polimi.it

Abstract

Mesh labeling is the key problem of classifying the facets of a 3D mesh with a label among a set of possible ones. State-of-the-art methods model mesh labeling as a Markov Random Field over the facets. These algorithms map image segmentations to the mesh by minimizing an energy function that comprises a data term, a smoothness terms, and class-specific priors. The latter favor a labeling with respect to another depending on the orientation of the facet normals. In this paper we propose a novel energy term that acts as a prior, but does not require any prior knowledge about the scene nor scene-specific relationship among classes. It bootstraps from a coarse mapping of the 2D segmentations on the mesh, and it favors the facets to be labeled according to the statistics of the mesh normals in their neighborhood. We tested our approach against five different datasets and, even if we do not inject prior knowledge, our method adapts to the data and overcomes the state-of-the-art.

Figure 1: Example of semantic mesh labeling using the method proposed in this paper on the fountain-P11 dataset

1. Introduction

In the last decades we witnessed a significant evolution in 3D modeling from images: moving from the early point-based approaches, researchers have proposed a wide variety of solutions that output a dense 3D mesh model of the environment. The most recent trend in computer vision, as well as in robotics and photogrammetry, is to add semantic labels to the 3D reconstruction, in order to estimate a more informative 3D model for robot navigation and localization or visualization and inspection.

On one side volumetric semantic 3D reconstruction [16, 19, 29] couples the 3D reconstruction with semantic labeling; this improves the quality of the 3D model recovered from the images, but the complex minimization techniques involved and the spatial complexity of the volumetric representation affect negatively the efficiency and the scalability of this approach. Recent semantic Multi-View Stereo (MVS) mesh refinement address the problem by estimating a low resolution mesh and refining it exploiting photometric and semantic information [4, 25]. As the shape and res-
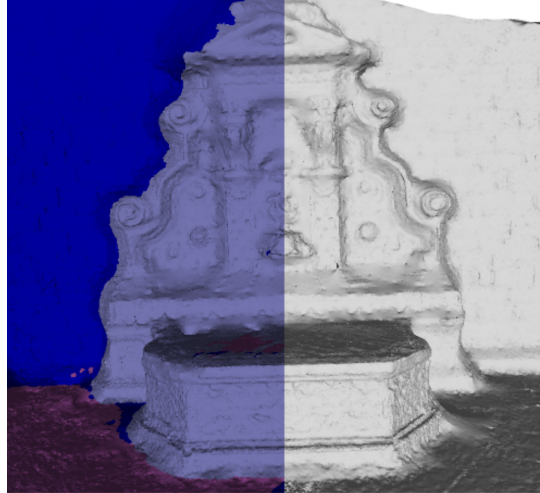
olution of the mesh change while the algorithm refines the mesh, the labels need to be updated accordingly. Therefore a time-crucial step is the semantic mesh labeling. Moreover, when a big and accurate 3D model, recovered from image-based algorithms, such as [39], is already available, recomputing the entire model through expensive volumetric semantic 3D reconstruction algorithms is inconvenient or, in some cases, impossible due to memory restrictions. In such cases mesh labeling represents the unique feasible possibility to recover a semantic reconstruction.

Semantic mesh labeling aims at classifying each facet of a 3D mesh model by exploiting available 2D images segmentation, 3D cues, or geometric priors, as in Figure 1. Successful approaches leverage 2D image segmentations, and model the labeling problem as energy minimization over a Markov Random Field (MRF). In the MRF, the nodes represent facets whereas the potentials gather semantic information from image segmentations, enforce smoothing, and represent class-specific priors. For instance, Blaha *et al.* [4] proposes a unary prior which, for each class, defines a preferred orientation with respect to the ground vector.
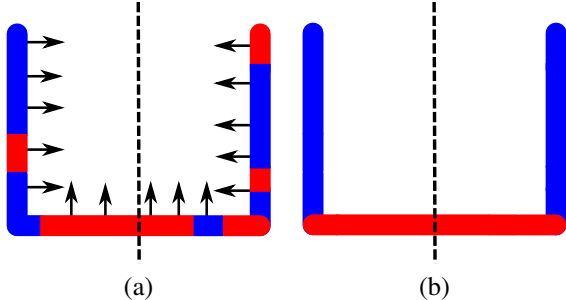
Figure 2: Toy example that shows the idea behind the term E_{norm}

In this paper we propose a novel MRF formulation that does not require any additional knowledge of the environment except the 3D model and the image segmentations. We propose to use a novel unary term in conjunction with the usual data term a smoothing term and the discontinuity term proposed in [25]. We exploit a coarse labeling of the 3D mesh to collect locally the distribution of the normals for each class, it biases the facet classification according to them; instead of defining manually the preferred orientation for each class, as in [4] or estimating it globally as in [25].

2. Related Works

Recovering a semantically annotated 3D model has been addressed by joining the labeling with the 3D reconstruction, or by decoupling the two tasks.

Joint semantic and 3D reconstruction has first been proposed by Häne *et al*. [16] and it stems from the idea that scene semantics enables the definition of class-priors to improve the quality and the robustness of a 3D reconstruction algorithm. In their work, the authors define a convex minimization problem in the continuum 3D space with a unary and a pairwise potential, then they relax it to work in the discrete voxel space. The per-voxel unary potential is estimated from 2D image segmentations and the visibility rays from each camera (data term). The pairwise term is the combination of smoothness and class-specific priors, the latter is learned from manually annotated data, and defines the probability of class changes depending on the normal transition. This work has been extended in [29, 28] to embed higher order ray potentials into the optimization of the graphical model. For a more in-depth discussion about semantic volumetric reconstruction we refer the reader to [15]. Conversely, Kundu *et al*. [19] propose to bootstrap from Structure from Motion (SfM) points and they define a Conditional Markov Random Field on a 3D voxel grid, where they estimate the unary potentials from image segmentations, and they add handcrafted class-specific priors to enforce reconstruction consistency. The method is effective for large scale environment, but the 3D labeled model has a low resolution due to the low number of SfM points.

Even if joint segmentation and reconstruction algorithms are effective, they strongly rely on volumetric reconstruction and the optimization techniques are usually expensive so that they require a difficult trade off between accuracy scalability and computational speed. More efficient representations have been proposed, *e.g*., octrees as in [31], voxel-hashing [37], submaps [8], or multi-grids [5]; however the scalability of multi-label optimization of these methods is still an open issue.

Different approaches decouple the reconstruction and the mesh labeling, so that the complexity required by the 3D reconstruction module is reduced, and the scene segmentation is delegated to a semantic labeling algorithm. Sengupta *et al*. [30] build a 3D volumetric reconstruction through the standard Truncated Signed Distance Function (TSDF), then they label the images and fuse them in the 3D model. Instead, Valentin *et al*. [35], design a CRF whose unary potentials correspond directly to 3D voxels and are trained by taking into account images color, and geometric features of the facets. The authors improve the labeling with respect to Sengupta *et al*. [30], however the algorithm is computationally expensive and, in our paper, we show that the priors learning phase is not needed.

Herman *et al*. [17] and McCormac *et al*. [21] fuse incrementally the 2D images segmentation into the reconstruction. These methods, as well as [30], benefit from the vast literature about 2D segmentation algorithms for 2D image segmentation [7, 38, 11]. However, the output of [17] and [21] are not continuous triangular meshes, but, respectively, a point cloud and a surfel model. Tung *et al*. [34] proposed an interesting method that estimates the unary potential of a CRF without any classifier, but by comparing the regions of the image computed as in [33], with a database of annotated examples. However they do not use any semantic prior to add robustness of the final outcome.

Kalogerakis *et al*. [18] segment meshes into semantics categories by learning different 3D features of the mesh such as shape curvature. Verdie *et al*. [36] take as input an existing 3D mesh of a city reconstructed from aerial images. They subdivide it into superfacets, which are subareas of the 3D mesh, and they collect geometric attributes, *i.e*., elevation, planarity and horizontality. Then they model the problem with a Markov Random Field with unary class-likelihood and pairwise class-transition terms defined manually. Even if the results are remarkable, the handcrafted priors do not generalize the method in non urban areas. For this reason Rouhani *et al*. [27] and Martinovic *et al*. [20] extended [36] with a classifier trained on both geometrical features and color information from the mesh texture. The method improves the results of [36] and it avoids using handcrafted priors; however, due to the features taken into

account, its scope is still limited to city models from aerial images. In [10] the authors propose a series of classifiers that labels a point cloud but even in this case the algorithm is domain-specific (restricted to façades) and its application is restricted to building facades. The main drawback of such methods that rely on just 3D cues to segment the mesh, is that they are not able to exploit the available deep learning methods such as [7, 38, 11], that in the last years have shown impressive performances in image segmentation. Riemenschneider *et al.* [24] propose to classify a mesh according to 2D image segmentation, however their method requires a learning stage to estimate which surface orientation has to be labeled from which viewing direction by relying on 3D geometric features.

Blaha *et al.* [4] and Romanoni *et al.* [25] propose two different mesh labeling methods, which are the closest to our proposal, and which rely on 2D image segmentation. Both [4] and [25] are two MVS algorithms that extend the photometric refinement of [39]. They bootstrap from a 3D mesh and 2D image segmentations and they jointly refine the former through both semantics and photometric data. Mesh labeling plays a substantial role here; since the model changes as new refinement iterations are applied, the mesh needs to be relabeled at each iteration. They define a Markov Random Field (MRF) over the mesh facets with unary potentials induced by the 2D image segmentations and pairwise smoothing terms for adjacent facets. Blaha *et al.* [4] add a unary geometric term that encodes handcrafted class-specific priors; Romanoni *et al.* [25] estimate per-class normal distributions from a coarse mesh labeling and they add a unary term that, if the facets labeled as c show a strong preference for a direction, other facets with similar direction are biased to belong to the same class c. The relevant aspect of these two methods and the method proposed in this paper is that they do not need to learn unary potentials over class labels.

3. Proposed Method

The proposed algorithm models the labeling as an energy minimization problem over a Markov Random Field. The inputs are: the 3D triangular mesh model of the scene under analysis, and, for each image that captures the scene, the corresponding camera calibration and the output of a semantic classifier in the form of a likelihood distribution for each class c.

Given a label assignment $\mathbf{l}$, which associates a label l_f to each facet $f \in \mathcal{F}$, we define the following energy:

$$E(\mathbf{l}) = \sum_{f \in \mathcal{F}} E_{\text{data}}(l_f) + \mu_1 \sum_{f \in \mathcal{F}} E_{\text{norm}}(l_f) +$$
$$\mu_2 \sum_{f,h \in \mathcal{F}} E_{\text{disc}}(l_f, l_h) + \mu_3 \sum_{f,h \in \mathcal{F}} E_{\text{smooth}}(l_f, l_h) \quad (1)$$
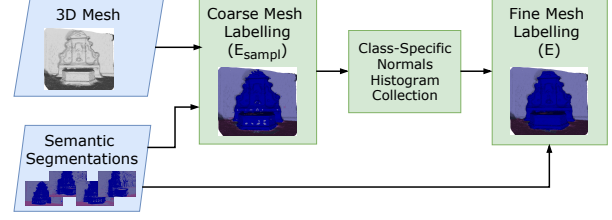

Figure 3: Pipeline of the proposed labeling algorithm

where E_{data} and E_{smooth} are respectively the data unary term proposed in [4] and a smoothness term adapted from [4]. The pairwise term E_{disc}, we named discontinuity prior, was proposed in [25] and biases the class changes between adjacent facets where significant variations between normal orientations occur.

The main contribution of this paper is the unary term E_{norm} that replaces the priors defined manually as in [4] or by means of learning algorithms as in [16] or [27]. Assuming a coarse labeling of the mesh is available, through the term E_{norm} we label each facet by comparing its normal direction with the normal directions of the facets in its neighborhood. For instance, in Figure 2(a) we have a toy 3D model with a coarse labeling: the normals directed to the right located on the left and, conversely, the normals directed to the left located on the right likely belongs to blue facets; both on the left and on the right sides the normals directed upwards likely belong to red facets. The term E_{norm} aims at collecting such likelihoods and extend them locally, so to obtain the result in Figure 2(b). To compute the coarse labeling we first apply a simplified version of the MRF that minimizes the energy:

$$E_{\text{simpl}}(\mathbf{l}) = \sum_f E_{\text{data}}(l_f) + \sum_{f,h} E_{\text{smooth}}(l_f, l_h) \quad (2)$$

Figure 3 summarizes the pipeline of our algorithm.

3.1. Data and Smoothness Terms

The term E_{data} collects the evidences estimated by the image classifier among the different views. Since we know the likelihoods $I_i(l, x)$ for label l in the pixel x of the i-th image, we define:

$$E_{\text{data}}(l_f) = -log \left(\sum_i \int_{\Omega_i} I_i(l_f, x) dx \right) \quad (3)$$

where Ω_i is the area of the i-th image where the 3D model is projected.

We define the smoothness term between two adjacent facets as:

$$E_{\text{smooth}}(l_f, l_h) = \begin{cases} 1.0, & \text{if } l_f \neq l_h \\ 0.0, & \text{if } l_f = l_h \end{cases} \quad (4)$$

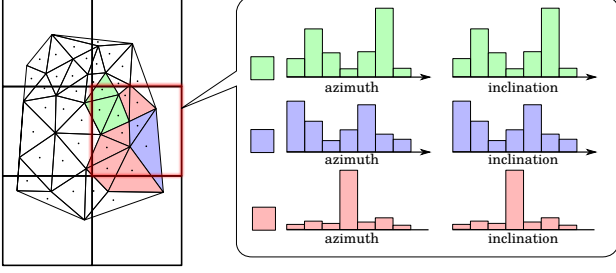to penalize frequent class changes among nearby facets.

Figure 4: Per-cell distributions of azimuth and inclination among the normal of the facts

Table 1: Resolutions and output statistic for each dataset for both the high and low resolution models

	num. cameras	image resolution	num. facets
fountain-P11	11	3072x2048	1.9M
castle-P30	30	3072x2048	5M
Southbuilding	14	3072x2304	3M
KITTI 95	548	1242x375	1.2M
Dagstuhl	67	1600x1200	4.7M

3.2. Discontinuity Prior

Given two adjacent facet f_1 and f_2 with the corresponding normals $\mathbf{n}_{f_1}$ and $\mathbf{n}_{f_2}$, we define:

$$
E_{\text{disc}}(l_f, l_h) = \begin{cases} e^{-\frac{(\angle(\mathbf{n}_{f_1}, \mathbf{n}_{f_2}))^2}{2*(\frac{\pi}{2})^2}}, & \text{if } l_f \neq l_h \\ 0.0, & \text{if } l_f = l_h \end{cases} \quad (5)
$$

This term represents the idea that two adjacent objects or two adjacent parts of the scene that belong to two different classes show discontinuities along the the mutual boundary. For instance, in the region between a car and the street the facets incident to the wheel-street boundary have significantly different orientations. On the other hand is likely that adjacent facets with the same labels have similar orientation. Only in the case of sharp edges the assumption does not hold, however even if we favor a label change in those areas, the data and smoothness terms provide enough evidence to avoid misclassification.

3.3. Normal Likelihood

The previous terms are able to map the label from 2D segmentation to the 3D mesh. However, in some cases the smoothing term is not able to tackle the noise affecting the segmentations. In such cases the common approach is to define class specific priors as in [4] that favor one class with respect to others depending on the surface orientation. These priors are effective and they are able to compensate noise and classification errors; however, they require both
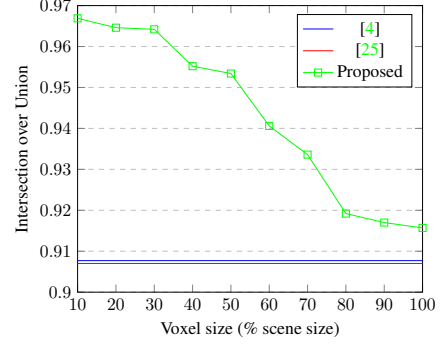

Figure 5: IoU with varying voxel size

the knowledge of the gravity vector and to define and tune, for each class, which orientations are more likely.

We replace the prior term, such that we no longer require prior knowledge about the environment, but we still favor a labeling depending on facets orientation. The result is a more flexible and general algorithm. We follow and extend the intuition reported in [25], that is, given a 3D mesh with a coarse labeling, we can collect, for each label, the distribution of the normals and use them to estimate the likelihood of a facet f with a normal n_f to belong to a class with label l, therefore we turn the prior term into a likelihood term.

For each class, the authors in [25] *et al*. collect the mean normal and its variance: if the facets belonging to a class of label l, *e.g.*, the ground, show a strong tendency towards a single orientation, *i.e.*, the normals have a unimodal distribution with low variance, the method favors the other facets with similar orientation to be labeled as l. This "global" approach fails in two cases. One case is exemplified as follows: if the algorithm biases all the facets oriented upwards to become ground, also the roofs of the cars, or some roofs of buildings are biased to be ground. The second case happens if the normals (of the facets) belonging to a class have a multimodal distribution, such as the walls in a street; in this scenario the single mean and variance computed in [25] are not expressive enough.

To face these two issues, we propose to locally estimate the distributions of the normals for each class/label, and to represent and use directly these distributions as a likelihood term in the energy function instead of defining it from the mean and the variance.

First we define a 3D lattice that covers the whole 3D mesh, its cells are cubes of size d_i, *e.g.*, the left part of Figure 4 exemplifies it in 2D. For each cell c_k and each class l we turn the normal of the facets whose centroid is inside of c_i into unit spherical coordinates. Then, we define two histograms $hist_{k,l}^{\text{azim}}$ and $hist_{k,l}^{\text{incl}}$ to collect respectively the distribution of the normal azimuths and inclinations.

When we evaluate a facet f with the centroid inside the

Table 2: Segmentation results w.r.t. the state of the art and the baseline method without the proposed term E_{norm}

dataset	method	average accuracy	average recall	average F-score	average precision	overall accuracy	overall recall	overall F-score	overall precision	IoU
fountain-P11	Blaha *et al.* [4]	0.9499	0.8468	0.9291	0.8776	0.9478	0.9132	0.9157	0.9110	0.8827
	Romanoni *et al.* [25]	0.9476	0.8447	0.9291	0.8759	0.9449	0.9099	0.9115	0.9070	0.8793
	Baseline (w/o E_{norm})	0.9481	0.8389	0.9244	0.8701	0.9464	0.9118	0.9124	0.9082	0.8800
	Proposed	**0.9548**	**0.8536**	**0.9384**	**0.8851**	**0.9540**	**0.9228**	**0.9248**	**0.9202**	**0.8900**
castle-P30	Blaha *et al.* [4]	0.9623	0.8401	0.7859	0.8114	0.9399	0.9757	0.8985	0.9352	0.9070
	Romanoni *et al.* [25]	0.9628	**0.8986**	0.8819	0.8886	0.9282	**0.9798**	0.8993	0.9366	0.9077
	Baseline (w/o E_{norm})	0.9630	0.8666	0.7955	0.8264	0.9410	0.9754	0.8995	0.9351	0.9079
	Proposed	**0.9816**	0.8695	**0.9423**	**0.9008**	**0.9721**	0.9774	**0.9613**	**0.9686**	**0.9646**
Southbuilding	Blaha *et al.* [4]	0.9561	**0.7717**	**0.8658**	**0.8129**	**0.9527**	0.9040	0.9428	0.9225	0.8920
	Romanoni *et al.* [25]	0.9507	0.7644	0.8419	0.8000	0.9444	0.8935	0.9328	0.9124	0.8839
	Baseline (w/o E_{norm})	0.9097	0.6581	0.7333	0.6884	0.8763	0.8564	0.8398	0.8461	0.7984
	Proposed	**0.9572**	0.7701	0.8651	0.8128	0.9523	**0.9073**	**0.9448**	**0.9249**	**0.8936**
KITTI 95	Blaha *et al.* [4]	**0.9648**	0.8601	0.8541	0.8588	**0.9645**	0.9155	0.8409	0.8744	**0.8942**
	Romanoni *et al.* [25]	0.9491	0.8006	0.7832	0.8232	0.9519	0.9004	0.8050	0.8445	0.8470
	Baseline (w/o E_{norm})	**0.9647**	0.8599	0.8534	0.8579	**0.9642**	0.9158	0.8393	0.8736	0.8938
	Proposed	**0.9648**	**0.8761**	**0.8682**	**0.8724**	**0.9642**	**0.9290**	**0.8509**	**0.8860**	**0.8942**
Dagstuhl	Blaha *et al.* [4]	0.9813	**0.9291**	0.9431	0.9581	0.9776	**0.9258**	0.9400	0.9327	0.9536
	Romanoni *et al.* [25]	0.9586	0.8327	0.8868	0.9519	0.9443	0.7936	0.9329	0.8561	0.8972
	Baseline (w/o E_{norm})	0.9592	0.8130	0.8856	**0.9769**	0.9427	0.7740	**0.9701**	0.8600	0.8839
	Proposed	**0.9821**	0.9241	**0.9472**	0.9717	**0.9787**	0.9192	0.9601	**0.9391**	**0.9561**

cell c_k, we define the likelihood:

$$P_k(f|l) \sim e^{\mathrm{hist}^{\mathrm{azim}}_{k,l}(\mathrm{az}(\mathbf{n}_f)) * \mathrm{hist}^{\mathrm{incl}}_{k,l}(\mathrm{inc}(\mathbf{n}_f))} \qquad (6)$$

where $\mathrm{az}(\cdot)$ and $\mathrm{inc}(\cdot)$ convert a vector into its azimuth and inclination. Therefore the normal likelihood term is:

$$E_{\mathrm{norm}} = -log(\mathrm{hist}^{\mathrm{azim}}_{k,l}(\mathrm{az}(\mathbf{n}_f)) * \mathrm{hist}^{\mathrm{incl}}_{k,l}(\mathrm{inc}(\mathbf{n}_f))) \quad (7)$$

Normal Likelihood vs. Smoothing The term E_{norm} let that, locally, facets with similar orientation takes similar labels. Even if this behavior may seems similar to a classical contrast-sensitive smoothing term, its meaning, implementation and result is different. The two-step approach (coarse labeling with histogram computation + fine labeling) let us to define E_{norm} as a simple unary term for each facet, instead of a pairwise term, as the usual smoothing term. This means that during the optimization procedure E_{norm} does not change its value for each facet f, similarly to what happens with prior terms differently from the behavior of a smoothness term that spreads the label with current higher support. In principle, our method could be imitated by a classic smoothing term that adds a pairwise term connecting every facet to all the facets inside the big voxel of the grid. However in this case the dimension of the MRF would quickly become intractable, *e.g.*, in the fountain-P11 dataset we have around 12000 facets for each voxel, therefore this approach would require $\frac{12000!}{2!(2-12000)!} \approx 72 \cdot 10^6$ pairwise terms for each facet.

4. Experiments

We tested our labeling method against five datasets: fountain-P11, castle-P30 [32], Southbuilding [16], KITTI 95 [12] and Dagstuhl. In Table 1 we illustrate the statistics of the datasets and the mesh we estimated. Even though it is related to the problem we faced in this paper, we did not used the Varcity dataset [24], indeed the classes are all spread along the building facade, and in this case a class-specific normal prior does not have any influence on the labeling stage. To extract the input mesh we calibrate the cameras with Structure from Motion (SfM) [22], and we estimate the depth maps with the plane sweeping library [13, 14]. We fuse SfM points and the depth maps by means of the volumetric reconstruction algorithm proposed in [26], and we refined the model through photometric refinement [39]. For fountain-P11, castle-P30, Southbuilding and Dagstuhl datasets, we extract the 2D segmentation with the per-pixel likelihood score through the Multiboost classifier [3]: we manually labeled few images and we classify all the datasets images into *ground*, *wall*, *vegetation* and *other*. For the KITTI sequence we used a pre-trained version of [38]. We modeled the MRF through the openGM library [2] and we adopted the alpha expansion algorithm [6] to infer the labeling.

We tested our method against the labeling algorithms in [4] and [25]. Even if [4] and [25] iterate between mesh labeling and mesh refinement, and both steps indirectly influence each other, a single labeling step is a single block of the pipeline, and works independently from the refinement. Since our method focuses on labeling, we compared it against the mesh labeling of [4] and [25] isolated from the refinement step.

As a first step, we tuned the three parameters μ_1, μ_2 and μ_3 of the energy function (1) on the dataset fountain-P11. We assigned different values to each parameter while keep-

2D Segmentation	GT	Blaha *et al.* [4]	Romanoni *et al.* [25]	Proposed

fountain-P11

castle-P30

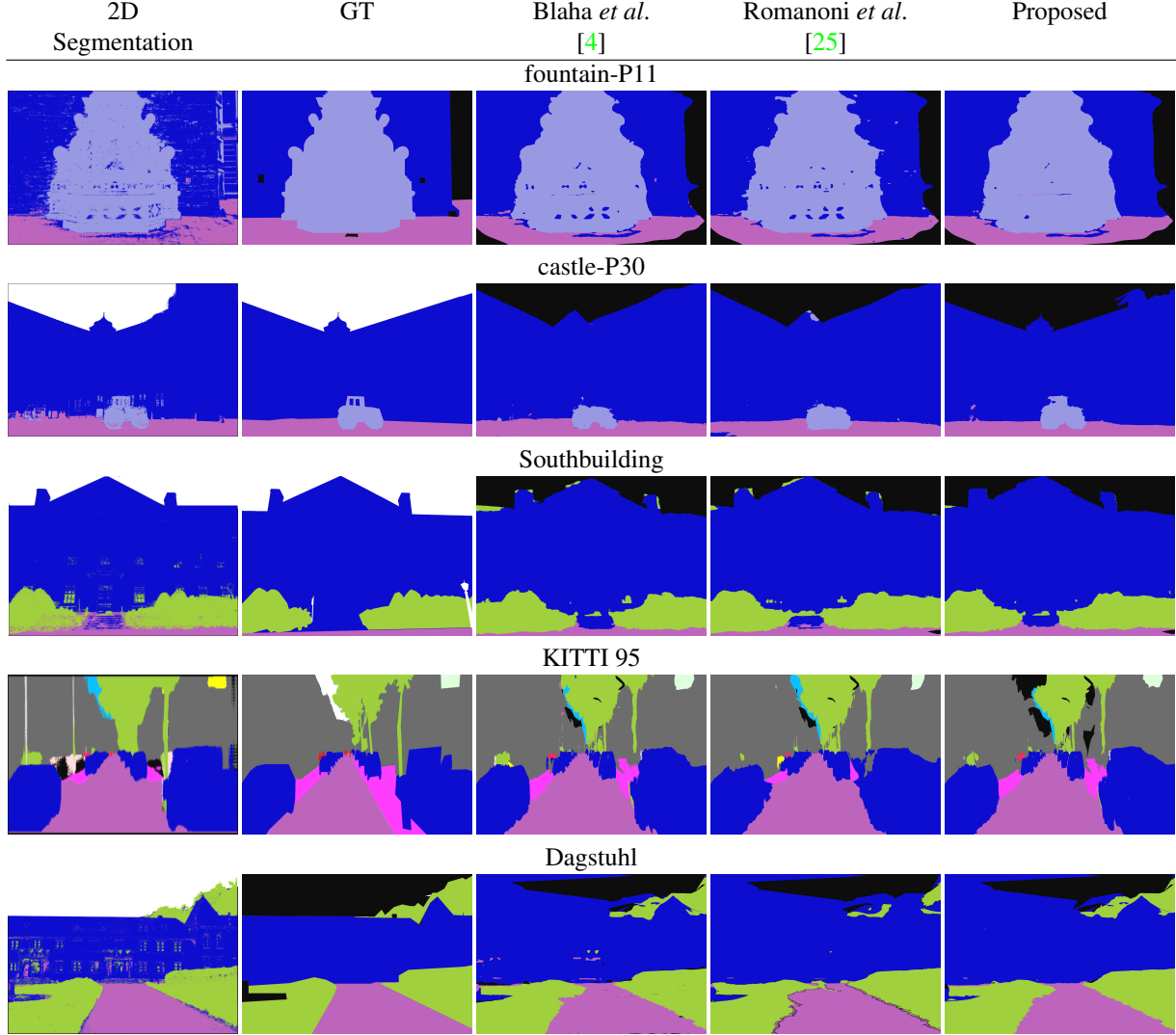Southbuilding

KITTI 95

Dagstuhl

Figure 6: Results on fountain-P11, castle-P30, Southbuilding, KITTI 95 and Dagstuhl

Table 3: Segmentation results in the failure case

dataset	method	average accuracy	average recall	average F-score	average precision	overall accuracy	overall recall	overall F-score	overall precision	IoU
	[3]	0.9518	**0.7236**	0.8579	**0.7743**	**0.9569**	**0.8569**	0.8645	**0.8461**	**0.8462**
DTU15	[17]	0.9517	**0.7238**	0.8261	0.7494	0.9552	0.8125	0.8715	0.8161	**0.8460**
	Proposed	**0.9547**	0.6442	**0.8948**	0.6739	0.9444	0.7203	**0.8989**	0.7074	0.8552

ing the other two fixed. We choose $\mu_1 = 0.2$, $\mu_2 = 0.2$ and $\mu_3 = 1.0$, and we use these same values with the other datasets. To have a fair comparison we followed the same procedure to tune [4] ($\mu_1 = 0.4$, $\mu_2 = 1.0$) and [25] ($\mu_1 = 0.2$, $\mu_2 = 0.4$ and $\mu_3 = 0.4$). For [4] we also need to manually estimate the gravity vector.

While tuning the size of the cubes in the 3D lattice adopted to estimate the local distributions, we find that the parameter d_i is quite robust to variations, it however de-

pends on the dataset dimension and the labeling granularity and shows best results when it is not too big with respect to the scene, otherwise the results tends to be similar to the global method proposed in [25]. We choose $1.0\,\text{m}$, $7\,\text{m}$, $3.2\,\text{m}$, $4\,\text{m}$ and $4\,\text{m}$ respectively for the fountain-P11, castle-P30, Southbuilding, KITTI 95 and Dagstuhl datasets. In Figure 5 we illustrate the intersection over union for the castle-P30 dataset with varying values of d_i in comparison with the results for [4] and [25].

	Blaha *et al.* [4]	Romanoni *et al.* [25]	Proposed

fountain-P11

castle-P30
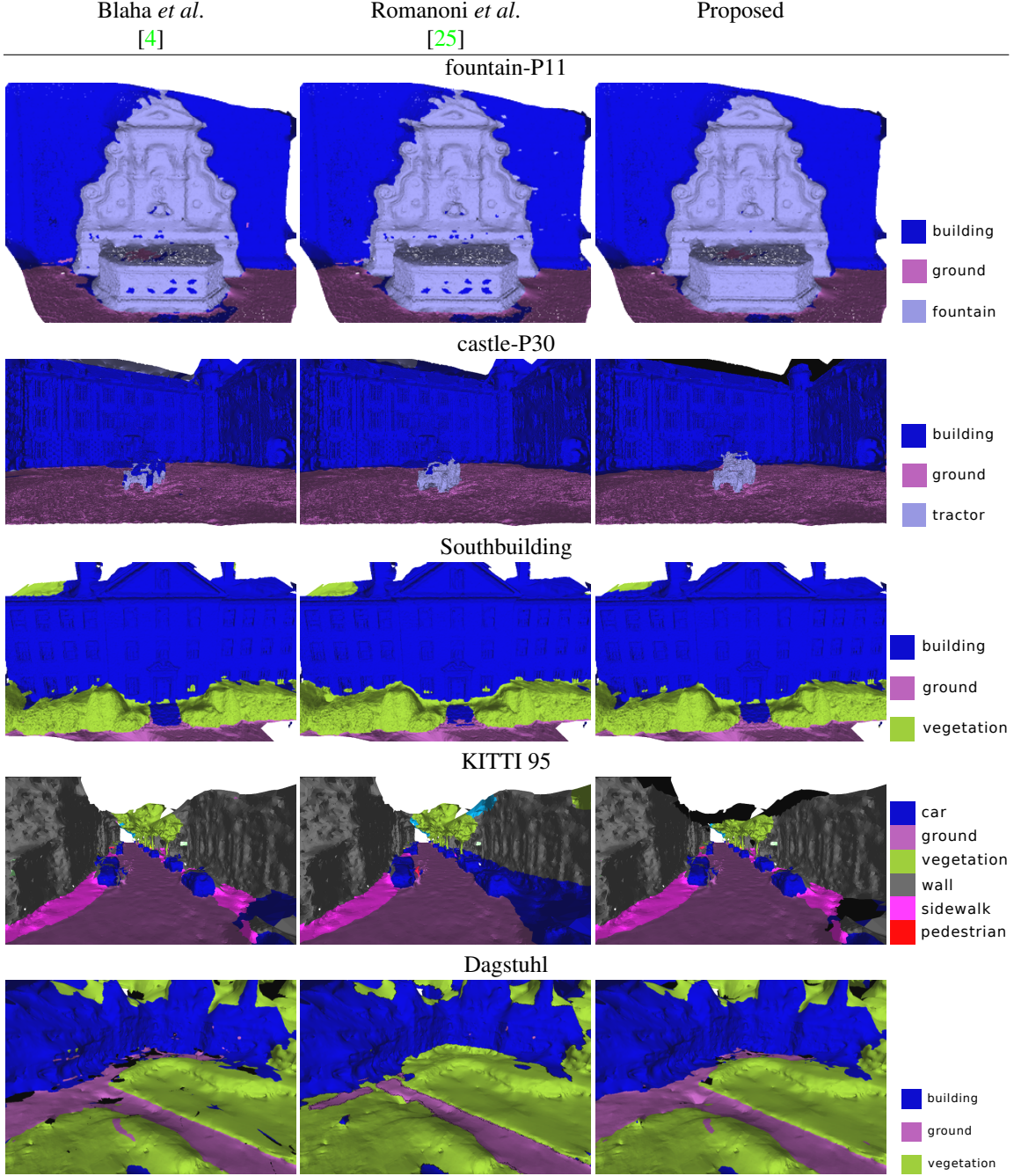
Southbuilding

KITTI 95

Dagstuhl

Figure 7: Labeled meshes

In Table 2 we show the results of the proposed method against the methods in [4] and [25] and we evaluate the ablation of the proposed term E_{norm} (Baseline). We report different metrics: the average accuracy, recall, F-score and precision are computed as the average metric over all the pixel, while the overall measures are computed as the average over the classes. To compute those metrics we compare the mesh rendering against 11, 15, 12, 5 and 4 different ref-

erence ground truth view respectively for the fountain-P11, the castle-P30, Southbuilding, KITTI 95 and Dagstuhl. In Figure 6 we illustrate the segmentation results for a reference view for the three datasets and in Figure 7 the labeled mesh. Even if the image segmentations coming from Multiboost are very noisy (first column of Figure 6), the three algorithm managed to well compensate most of the misclassification. Even if the method we propose does not require
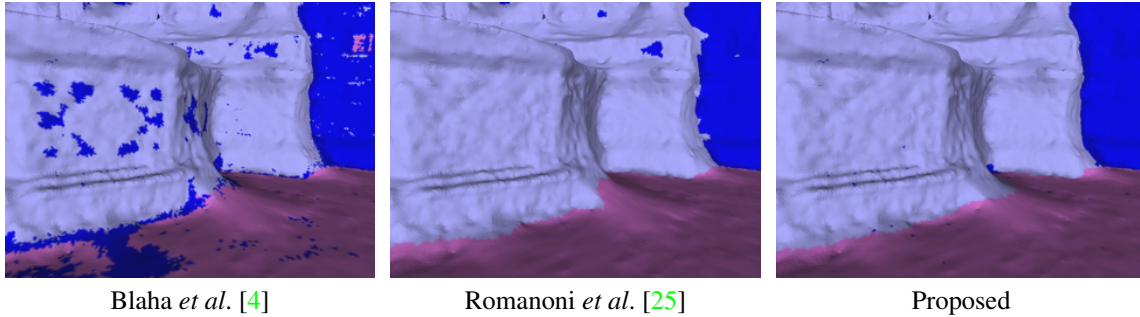
| Blaha *et al*. [4] | Romanoni *et al*. [25] | Proposed |

Figure 8: Detail of the fountain-P11 dataset

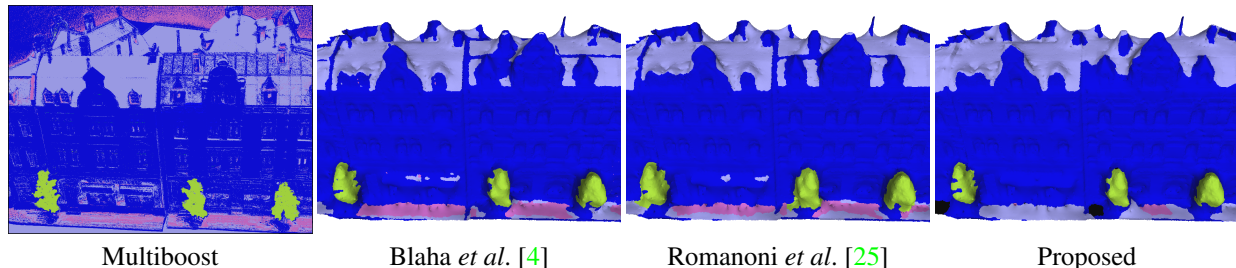| Multiboost | Blaha *et al*. [4] | Romanoni *et al*. [25] | Proposed |

Figure 9: Failure case with the DTU15 dataset

to add any prior information, its performance is better than the other methods in most of the metrics for each of the five datasets. One of the reason is that, since the meshes have high resolution, the per-class distributions of the normals are quite complex, *e.g.*, the ground is not simply flat, but it contains crispier details, therefore the simple prior of [4] and the P_{norm} term of [25] are often not sufficiently expressive. Instead, our algorithm explicitly models these distributions through the histograms; therefore the prior-like normal term adapts itself to the scenario. Another reason of the improvement is evident in the fountain-P11 dataset. Here both [4], through the handcrafted prior, and [25], through its global prior, favor all the facets perpendicular with respect to the ground to be labeled as wall: this avoids to fill most of the 2D misclassification that happen on the fountain surface. Our approach completely overcome this issue since it collects the statistics locally, and not just because it acts as a strong smoothing term, indeed the small road sign in the KITTI 95 dataset is preserved (see Figure 7). Finally, in addition to the processing shared with [4] and [25], our method needs to compute the histogram; however the operation takes only 1.6 to 4.6 seconds.

In Table 3 we report a failure case that happened with the sequence 15 of the DTU dataset [1]. Here the Multiboost algorithm fails to provide a reliable segmentation of the ground (Figure 9). Therefore our method has not enough support to understand that facets directed upwards likely belongs to the ground. The ground is better segmented in [4] because they enforce directly this concept by the class-specific prior and even in [25] because they collect the

normal statistics globally, therefore with a wider support.

5. Conclusion and Future Works

In this paper we proposed a novel method to label 3D models with semantic labels. In particular we defined a Markov Random Field over the mesh, as in the state-of-the-art, endowed with a new term that biases the labeling depending on the facet orientation and acts as a class-specific geometric prior, but, instead of learning it or fixing it manually, we estimate it directly from a coarse labeling of the mesh, collecting locally the distribution of the normals. Experiments on different datasets showed that the method is not only able to run without a learning phase or prior knowledge about the geometry of the classes, but it also improves the accuracy of the labeling.

As a future extension, we will investigate the possibility of running the method incrementally and on embedded multicores, coupling it with a real-time 3D reconstruction method [23] also accounting the energy constraints imposed by the computing platform [40, 41]. We also plan to test the method with a wider dataset such as ScanNet ([9]), by replacing the Multiboost algorithm, that failed to estimate a reasonable segmentation of ScanNet scenes, with a more sophisticated image segmentation algorithm based on neural networks.

Acknowledgements

References

[1] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, pages 1–16, 2016. 8

[2] B. Andres, T. Beier, and J. Kappes. OpenGM: A C++ library for discrete graphical models. *CoRR*, abs/1206.0111, 2012. 5

[3] D. Benbouzid, R. Busa-Fekete, N. Casagrande, F.-D. Collin, and B. Kégl. Multiboost: a multi-purpose boosting package. *Journal of Machine Learning Research*, 13(Mar):549–553, 2012. 5

[4] M. Blaha, M. Rothermel, M. R. Oswald, T. Sattler, A. Richard, J. D. Wegner, M. Pollefeys, and K. Schindler. Semantically informed multiview surface refinement. *International Journal of Computer Vision*, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[5] M. Blaha, C. Vogel, A. Richard, J. D. Wegner, T. Pock, and K. Schindler. Large-scale semantic 3D reconstruction: an adaptive multi-resolution model for multi-class volumetric labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3176–3184, 2016. 2

[6] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001. 5

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016. 2, 3

[8] I. Cherabier, C. Häne, M. R. Oswald, and M. Pollefeys. Multi-label semantic 3D reconstruction using voxel blocks. In *International Conference on 3D Vision (3DV)*, pages 601–610. IEEE, 2016. 2

[9] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3D reconstructions of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 2017. 8

[10] R. Gadde, V. Jampani, R. Marlet, and P. Gehler. Efficient 2D and 3D facade segmentation using auto-context. *IEEE transactions on pattern analysis and machine intelligence*, 2017. 3

[11] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*, 2017. 2, 3

[12] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361. IEEE, 2012. 5

[13] C. Häne, L. Heng, G. H. Lee, A. Sizov, and M. Pollefeys. Real-time direct dense matching on fisheye images using plane-sweeping stereo. In *International Conference on 3D Vision (3DV)*, volume 1, pages 57–64. IEEE, 2014. 5

[14] C. Häne, L. Heng, G. H. Lee, A. Sizov, and M. Pollefeys. Real-time direct dense matching on fisheye images using plane-sweeping stereo. In *International Conference on 3D Vision (3DV)*, volume 1, pages 57–64. IEEE, 2014. 5

[15] C. Häne and M. Pollefeys. An overview of recent progress in volumetric semantic 3D reconstruction. In *International Conference on Pattern Recognition (ICPR)*, pages 3294–3307. IEEE, 2016. 2

[16] C. Häne, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3D scene reconstruction and class segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 97–104. IEEE, 2013. 1, 2, 3, 5

[17] A. Hermans, G. Floros, and B. Leibe. Dense 3D semantic mapping of indoor scenes from RGB-D images. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2631–2638. IEEE, 2014. 2

[18] E. Kalogerakis, A. Hertzmann, and K. Singh. Learning 3D mesh segmentation and labeling. In *ACM Transactions on Graphics (TOG)*, volume 29, page 102. ACM, 2010. 2

[19] A. Kundu, Y. Li, F. Dellaert, F. Li, and J. M. Rehg. Joint semantic segmentation and 3D reconstruction from monocular video. In *European Conference on Computer Vision*, pages 703–718. Springer, 2014. 1, 2

[20] A. Martinovic, J. Knopp, H. Riemenschneider, and L. Van Gool. 3D all the way: Semantic segmentation of urban scenes from start to end in 3D. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4456–4465, 2015. 2

[21] J. McCormac, A. Handa, A. Davison, and S. Leutenegger. Semanticfusion: Dense 3D semantic mapping with convolutional neural networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4628–4635. IEEE, 2017. 2

[22] P. Moulon, P. Monasse, R. Marlet, and Others. Openmvg. an open multiple view geometry library. https://github.com/openMVG/openMVG. 5

[23] E. Piazza, A. Romanoni, and M. Matteucci. Real-time cpu-based large-scale three-dimensional mesh reconstruction. *IEEE Robotics and Automation Letters*, 3(3):1584–1591, 2018. 8

[24] H. Riemenschneider, A. Bódis-Szomorú, J. Weissenberg, and L. Van Gool. Learning where to classify in multi-view semantic segmentation. In *European Conference on Computer Vision*, pages 516–532. Springer, 2014. 3, 5

[25] A. Romanoni, M. Ciccone, F. Visin, and M. Matteucci. Multi-view stereo with single-view semantic mesh refinement. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 706–715, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[26] A. Romanoni and M. Matteucci. Incremental reconstruction of urban environments by edge-points delaunay triangulation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4473–4479. IEEE, 2015. 5

[27] M. Rouhani, F. Lafarge, and P. Alliez. Semantic segmentation of 3D textured meshes for urban scene analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 123:124–139, 2017. 2, 3

[28] N. Savinov, C. Häne, L. Ladicky, and M. Pollefeys. Semantic 3D reconstruction with continuous regularization and ray

potentials using a visibility consistency constraint. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5460–5469, 2016. 2

[29] N. Savinov, L. Ladicky, C. Häne, and M. Pollefeys. Discrete optimization of ray potentials for semantic 3D reconstruction. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 5511–5518. IEEE, 2015. 1, 2

[30] S. Sengupta, E. Greveson, A. Shahrokni, and P. H. Torr. Urban 3D semantic modelling using stereo vision. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 580–585. IEEE, 2013. 2

[31] S. Sengupta and P. Sturgess. Semantic octree: Unifying recognition, reconstruction and representation via an octree constrained higher order mrf. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 1874–1879. IEEE, 2015. 2

[32] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 5

[33] F. Tung and J. J. Little. Collageparsing: Nonparametric scene parsing by adaptive overlapping windows. In *European Conference on Computer Vision*, pages 511–525. Springer, 2014. 2

[34] F. Tung and J. J. Little. MF3D: Model-free 3D semantic scene parsing. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4596–4603. IEEE, 2017. 2

[35] J. P. Valentin, S. Sengupta, J. Warrell, A. Shahrokni, and P. H. Torr. Mesh based semantic modelling for indoor and outdoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2067–2074, 2013. 2

[36] Y. Verdie, F. Lafarge, and P. Alliez. LOD generation for urban scenes. *ACM Transactions on Graphics (TOG)*, 34(3):30, 2015. 2

[37] V. Vineet, O. Miksik, M. Lidegaard, M. Nießner, S. Golodetz, V. A. Prisacariu, O. Kähler, D. W. Murray, S. Izadi, P. Pérez, et al. Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 75–82. IEEE, 2015. 2

[38] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, and A. Courville. ReSeg: A recurrent neural network-based model for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 41–48, 2016. 2, 3, 5

[39] H. H. Vu, P. Labatut, J.-P. Pons, and R. Keriven. High accuracy and visibility-consistent dense multiview stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):889–901, 2012. 1, 3, 5

[40] D. Zoni, L. Colombo, and W. Fornaciari. Darkcache: Energy-performance optimization of tiled multi-cores by adaptively power-gating llc banks. *ACM Trans. Archit. Code Optim.*, 15(2):21:1–21:26, May 2018. 8

[41] D. Zoni, L. Cremona, A. Cilardo, M. Gargliardi, and W. Fornaciari. Powertap: All-digital power meter modeling for run-time power monitoring. *Microprocessors and Microsystems*, 2018. 8