Multi-View Stereo with Single-View Semantic Mesh Refinement

Andrea Romanoni Marco Ciccone Francesco Visin Matteo Matteucci
Politecnico di Milano, Italy
{andrea.romanoni, marco.ciccone, francesco.visin, matteo.matteucci}@polimi.it

Abstract

While 3D reconstruction is a well-established and widely explored research topic, semantic 3D reconstruction has only recently witnessed an increasing share of attention from the Computer Vision community. Semantic annotations allow in fact to enforce strong class-dependent priors, as planarity for ground and walls, which can be exploited to refine the reconstruction often resulting in non-trivial performance improvements. State-of-the art methods propose volumetric approaches to fuse RGB image data with semantic labels; even if successful, they do not scale well and fail to output high resolution meshes. In this paper we propose a novel method to refine both the geometry and the semantic labeling of a given mesh. We refine the mesh geometry by applying a variational method that optimizes a composite energy made of a state-of-the-art pairwise photo-metric term and a single-view term that models the semantic consistency between the labels of the 3D mesh and those of the segmented images. We also update the semantic labeling through a novel Markov Random Field (MRF) formulation that, together with the classical data and smoothness terms, takes into account class-specific priors estimated directly from the annotated mesh. This is in contrast to state-of-the-art methods that are typically based on handcrafted or learned priors. We are the first, jointly with the very recent and seminal work of [3], to propose the use of semantics inside a mesh refinement framework. Differently from [3], which adopts a more classical pairwise comparison to estimate the flow of the mesh, we apply a single-view comparison between the semantically annotated image and the current 3D mesh labels; this improves the robustness in case of noisy segmentations.

1. Introduction

Modeling a scene from a set of images has been a long-standing and deeply explored problem for the Computer Vision community. The goal is to build an accurate 3D model of the environment basing on the implicit tridimensional data contained in a set of 2D images. These methods can be useful to digitalize architectural heritage, reconstruct maps of cities or, in general, for scene understanding.

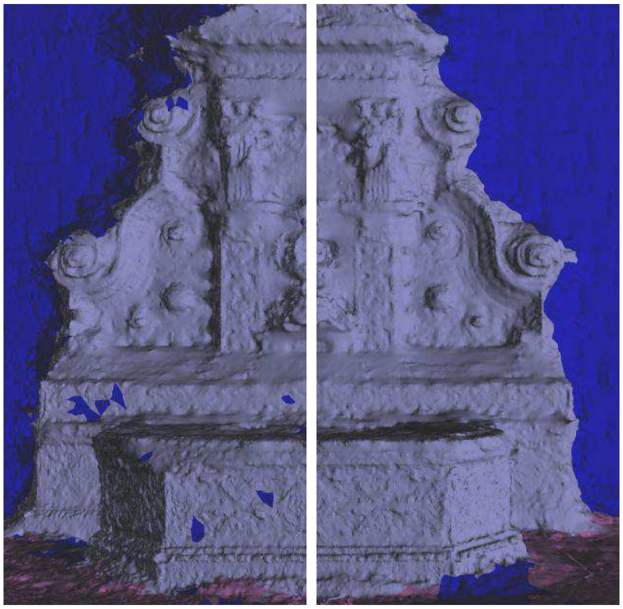


Figure 1. The reconstruction of the fountain-p11 dataset [32] without (left) and with the semantic refinement (right) proposed in this paper

Most dense 3D reconstruction algorithms consider only grayscale or color images, but thanks to the advancements in semantic image segmentation [2, 42, 6, 36], novel 3D reconstruction approaches that leverage semantic information have been proposed [30, 15, 28, 18, 4].

State-of-the-art semantic dense 3D reconstruction algorithms fuse images and semantic labels in a volumetric voxel-based representation, improving the accuracy thanks to strong class-dependent priors (learned or handcrafted), e.g., planarity of the ground or perpendicularity between ground and walls. These volumetric methods often allow to obtain impressive results, but they usually require a huge amount of memory. Only recently some effort has been put into solving this issue, for instance, through submaps [7] or multi-grid [4]. Very recently a mesh refinement guided by semantic has been proposed in [3]: the authors update the reconstruction by minimizing the reprojection error between pairs of segmented images. Differently from [3], our work compares each segmented image against the labels fused into the 3D mesh. This is much more robust to noise and errors in the images segmentations, as we show in Section 2.1 and support with an in-depth experimentation and discussion.

The paper is structured as follows: Section 2 presents an overview of the state-of-the-art of several topics involved in the proposed system; these are then discussed in detail in the context of the proposed method in Section 3. The experimental settings and results are presented in Section 4 and discussed in Section 5.

2. Related works

The method proposed in this paper crosses a variety of topics, namely classical and semantic volumetric reconstruction, photometric mesh refinement and mesh labeling. Here we review some of the most relevant works in those fields.

Volumetric 3D Reconstruction Volumetric 3D reconstruction represents the most widespread method to recover the 3D shape of an environment captured by a set of images. These methods build a set of visibility rays either from Structure from Motion, as in [20, 26], or depth maps, as in [23, 22]. After the space is partitioned, these rays are used to classify the parts as being free space or matter. The boundary between free space and matter constitutes the final 3D model. Depending on how the space is discretized, volumetric algorithms are classified as voxel-based [37, 31] or tetrahedra-based [33, 38, 20, 25]. The former trivially represent the space as a 3D grid. Despite their simplicity, these approaches often lead to remarkable results; however their scalability is quite limited due to the inefficient use of space that does not take into account the significant sparsity of the elements that should be modeled. Many attempt have been proposed to overcome this issue, e.g., by means of voxel hashing [29], but a convincing solution to the shortcomings of voxel-based methods seems to be still lacking.

On the other hand, tetrahedra-based approaches subdivide the space in tetrahedra via Delaunay triangulation: these methods build upon the points coming from Structure from motion or depth maps fusion and can adapt automatically to the different densities of the points in the space. As opposed to voxel-based methods, tetrahedra-based approaches are scalable and can be very effective in a wide variety of scenarios; however, since they tend to restrict the model to those parts of the space occupied by the points, in some cases they can make it hard to define priors on the non-visible part of the scene (e.g., walls behind cars), since they might not be modeled.

Semantic Reconstruction A recent trend in reconstruction methods has been to embed semantic information to improve the consistency and the coherence of the produced 3D model [18, 15]. Usually these methods rely on voxels representation and estimate the 3D labeled model by enriching each camera-to-point viewing ray with semantic labels; these are then typically used to replace the “matter” label of the classical method. The optimization process that leads to the final 3D reconstruction builds on class-specific priors, such as planarity for the walls or ground. Being voxel-based, these approaches lack scalability: the authors of [7] tackle this issue via submaps reconstruction and by limiting the number of labels taken into account during the reconstruction of a single submap, while [4] adopts multi-grids to avoid covering empty space with useless voxels.

Cabezas *et al*. [5] propose a semantic reconstruction algorithm that directly relies on mesh representation and fuses the data from aerial images, LiDAR and Open Street Map. Although proposing an interesting approach, such rich data is usually not available in a wide variety of applications, including the ones typical addressed in classical Computer Vision scenarios. For a more detailed overview of semantic 3D reconstruction algorithms we refer the reader to [14].

Photometric mesh refinement The approaches described so far extract the 3D model of the scene from a volumetric representation of it. In some cases these models lack details and resolution, especially due to the scalability issue mentioned before. Some works presented in the literature bootstrap from a low resolution mesh and refine it via variational methods [41, 11, 24, 39, 10]. Early approaches [41, 11] describe the surface as a continuous entity in R^3, minimize the pairwise photometric reprojection error among the cameras and finally discretize the optimized surface as a mesh. More recently, some authors [24, 39, 10] proposed a few more effective methods that compute directly the discrete gradient that minimizes the reprojection error of each vertex in the mesh. By relying on these methods Delaunoy and Pollefeys [9] proposed to couple the mesh refinement with the camera pose optimization. Li *et al.* [19] further improved the scalability of these methods by noticing that although mesh refinement algorithms usually increase the resolution of the whole mesh while minimizing the reprojection error, in some regions such as the flat ones there is no need for high vertex density. To avoid redundancy, [19] refine only the regions that produce a significant reduction of the gradient.

Mesh labeling Mesh labeling is usually modeled as a Markov Random Field (MRF) with a data term that describes the probability that a facet belongs to a certain class, and a smoothness term that penalizes frequent changes in the labeling along the mesh. Some approaches as [35] rely on handcrafted priors that define relationships among the labels basing on their 3D position and orientation with respect to the neighbors. Other methods add instead priors learned from data, such as [34, 27].

2.1. Semantic mesh refinement

The very recent work presented in [3] exhibits some similarities with what we propose in this paper. As in our case, the authors propose a refinement algorithm that extends [39] by leveraging semantic annotations. In [3] the reprojection error between pairs of views is minimized in the same fashion as [39], although instead of using just RBG images they also use pairwise masks for each label taken into account by the semantic classifier. The authors proved that this approach is effective and actually improves the photometric only refinement. However, we show that in presence of noisy or wrong classification their method lacks robustness (see Section 4.1) and we propose an alternative that does no suffer from this problem. Secondly, although also the authors of [3] update the labels of the 3D mesh with a MRF with a data term, a smoothness term and handcrafted geometric priors, we propose a simpler data term that makes the refinement much less expensive in terms of computation and a term computed from the reconstructed labeled mesh that encourages the facets with one label to have similar distribution to the input mesh facets with the same label.

3. Proposed method

The method we propose in this paper refines a labeled 3D mesh through a variational surface evolution framework: we alternate between the photo-consistent and semantic mesh refinement and the label update according to (Figure 2).

The initialization of our method is the 3D mesh estimated and labeled by the modified version of [26]. The volumetric method proposed in [26] estimates a point cloud from the images and discretizes the space through a Delaunay triangulation, initializing all the tetrahedra as *matter*, i.e., with 0 weight. It then casts all camera-to-point rays and increases the weight of the traversed tetrahedra; finally it estimates the manifold surface that contains the highest number free space tetrahedra, i.e., those with weight above a fixed threshold.

To take into account the semantic labels associated to the image pixels, and in turn to the camera-to-point rays, in our version a tetrahedron has one weight associated to the free space label and one weight associated to each new semantic label. For each ray from camera C to point P associated to label l, we increase the free space weight of the tetrahedra between C and P, as in the original case, then we increase the l weight of the tetrahedra that, following the ray direction, are just behind (below a fixed distance) the point P, similarly to [28]. Each tetrahedra is classified accordingly to the label with higher weight and the manifold is estimated as in the original version, but each triangle of the output mesh has now the label of the tetrahedron they belong to.

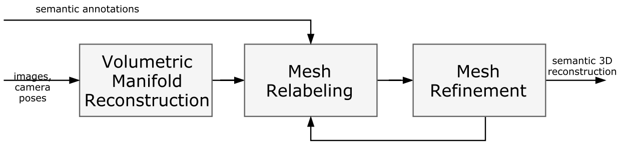


Figure 2. Architecture of the proposed system

3.1. Label smoothing and update

In two cases we need to update the labeling of the 3D mesh: 1) after the initial mesh computation and 2) after the semantic segmentation. After the volumetric reconstruction previously described, the labels of the initial mesh are prone to noise and, even if they collect more evidences for the same 3D point across the 2D set of semantically annotated images, sometimes they reflect the errors of the 2D classifier. After the refinement process the shape of the model changes, i.e., each facet change its orientation and positioning, therefore we update the labels to take into account the modifications of the value of the priors; moreover the refinement increases the resolution of the mesh at every $K = 5$ iterations, therefore some facets are subdivided and we need to label the new triangles.

We propose to model the labeling process as a Markov Random Field (MRF) using a simpler data term with respect to [3]: rather than collecting the likelihood of all the labels for all the images and for each facets, we sample these likelihoods at the vertices locations. While the geometric term in [3] considers reasonably handcrafted relationships among the facets and the corresponding labeling, the term we propose estimates the distribution of the normals of the facets belonging to the same class directly from the shape of the current scene.

Given $\mathcal{F}$ the set of facets and $\mathcal{L}$ the set of the labels, we aim at assigning a label $l \in \mathcal{L}$ to each facet $f \in \mathcal{F}$, such that we maximize the probability:

$$P_{\text{label}} = \prod_{l \in \mathcal{L}, f \in \mathcal{F}} \left(P^{lf}_{\text{data}} \cdot P^{lf}_{\text{norm}} \cdot P^{lf}_{\text{smooth}} \right) \tag{1}$$

The unary term P^{lf}_{data} describes the evidences of the label l for the facet f. In principle we need to take into account the whole area of the facet projected in each image. However, since our refinement process increases significantly the resolution of the mesh, we simplify the computation of this term only by considering the labels of the pixels where the vertices of the facet are projected and are visible. Given the 2D binary masks M_{2D}^{li} of the pixels labeled as l for the point of view of camera i, we define for each vertex v_f belonging to f:

$$P_{data}^{lf} = max(\beta, {}^{\nu(v_f, l)}/3), \tag{2}$$

$$\nu(v_f, l) = \frac{\sum_i M_{2D}^{li}(\Pi_i(v_f))}{\#\text{images } v_f \text{ is visible}}, \tag{3}$$

where β is $0 < \beta < 1$ prevents $\nu(v_f, l)$ from becoming 0 (we fixed it experimentally to 0.1), and $\Pi_i(\mathbf{x})$ projects the point $\mathbf{x}$ in the image plane of camera i. We divided the term $\nu(v_f, l)$ by 3 such that $0 < P_{data}^{lf} < 1$.

The unary term P_{norm}^{lf} represents the distribution associated to the class the facet belongs to. Instead of designing a geometric prior by hand as in [3] or learning it as in [16], we define a method to relate the normals of the facets belonging to the same class to the scene we are reconstructing. For each class associated to a label l we estimate the mean normal $\mathbf{m}_l$ and the angle variance a_l with respect to $\mathbf{m}_l$ of all the facet labeled as l. Then we define:

$$P_{\text{norm}}^{lf} = \mu e^{-\frac{\angle(\mathbf{n}_f, \mathbf{m}_l)^2}{2*(a_l)^2}}. \tag{4}$$

where μ weights the importance of P_{norm}^{lf} with respect to the other priors (we fixed $\mu = 1.5$)

Finally, we define the binary smoothness term $P_{\text{smooth}}^{f_1 f_2}$ between two adjacent facets f_1 and f_2:

$$P_{\text{smooth}}^{f_1 f_2} = \begin{cases} 0.2, & \text{if } \mathcal{L}(f_1) \neq \mathcal{L}(f_2) \\ 0.8, & \text{if } \mathcal{L}(f_1) = \mathcal{L}(f_2) \end{cases} \tag{5}$$

where $\mathcal{L}(f)$ represents the label of facet f; this term penalizes changes in the labeling of f_1 and f_2, to avoid spurious and noisy labels.

3.2. Semantic Mesh Refinement

The output of the previous steps is a mesh close to the actual surface of the scene, but it often lacks details. The most successful method to improve the accuracy of such mesh was proposed by [39]. The idea is to minimize the energy:

$$E = E_{\text{photo}} + E_{\text{smooth}}, \tag{6}$$

where E_{photo} is the data term related to the image photo-consistency, and E_{smooth} is a smoothness prior.

Given a triangular mesh S, with x and $\overrightarrow{n}$ a point and the corresponding normal on this mesh, two images I and J, and $err_{I,J}(x)$ a function that decreases if the similarity between the patch around the projection of x in J and I increases, then:

$$E_{\text{photo}} = \sum_{i,j} \int_{\Omega_{i,j}^{\text{S}}} err_{I,I_{ij}^S}(x_i) \mathrm{d}x_i, \tag{7}$$

where I_{ij}^S is the reprojection of the image from the j-th camera in the image I through the mesh S and $\Omega_{i,j}$ represents the domain of the mesh where the projection is defined. The authors in [39] minimize Eq. (11) through gradient descent by moving each vertex $X_i \in \mathbb{R}^3$ of the mesh according to the gradient:

$$\begin{aligned} \frac{\mathrm{d}E(S)}{\mathrm{d}X_i} &= \int_S \phi_i(x) \nabla E_{\text{photo}}(x) \mathrm{d}x, = \\ &- \sum_{i,j} \int_{\Omega_{i,j}^{\text{S}}} \phi_i(x) f_{ij}(x_i) / (\overrightarrow{n}^T \mathbf{d}_i) \overrightarrow{n} \mathrm{d}x_i, \end{aligned} \tag{8}$$

$$f_{ij}(x_i) = \partial_2 err_{I,I_{ij}^S}(x_i) D I_j(x_j) D \Pi_j(x) \mathbf{d}_i, \tag{9}$$

where $\phi_i(x)$ represents the barycentric coordinates if x is in the triangle containing X_i, otherwise $\phi_i(x) = 0$; Π_j is the j-th camera projection, the vector $\mathbf{d}_i$ goes from camera i to point x, the operator D represents the derivative and $\partial_2 err_{I,I_{ij}^S}(x_i)$ is the derivative of the similarity measure $err_{ij}(x)$ with respect to the second image.

In addition to the photo-consistent term, they minimize the energy E_{smooth} by means of the Laplace-Beltrami operator approximated with the umbrella operator [40], which moves each vertex in the mean position of its neighbors.

The method presented thus far considers only RGB information and a smoothness prior. To leverage the semantic labels estimated in the 2D images and on the 3D mesh, we define an energy function:

$$E = E_{\text{photo}} + E_{\text{sem}} + E_{\text{smooth}}, \tag{10}$$

where we minimize E_{photo} and E_{smooth} as in [39], and in the term E_{sem} we exploit the semantic information.

While RGB images contain relatively small noise and, to a certain extent, capture the same color for each point of the scene, when we deal with semantic masks the misclassification strongly depends on the perspective of the images and therefore these masks are not completely consistent among each other. For instance, if we have a mask J with a misclassified region r_m, even if the current 3D model of the scene is perfectly recovered, the reprojection of J (and in turn of r_m) through the surface on the i-th will unlikely match the misclassification in the mask estimated for camera i. We assume that the labels that come from image segmentation are noisier and more prone to error than the labels of the 3D mesh, which are estimated from the whole set of image segmentation and corrected with the MRF. For these reason, differently from the pairwise photometric term E_{photo}, we propose a single-view refinement method that compares the semantic mask I with the rendering of the labeled mesh on camera i (Figure 3). By doing so, our refinement affects the borders between the classes in the 3D model and we discard all the wrong classification of the single image segmentation.

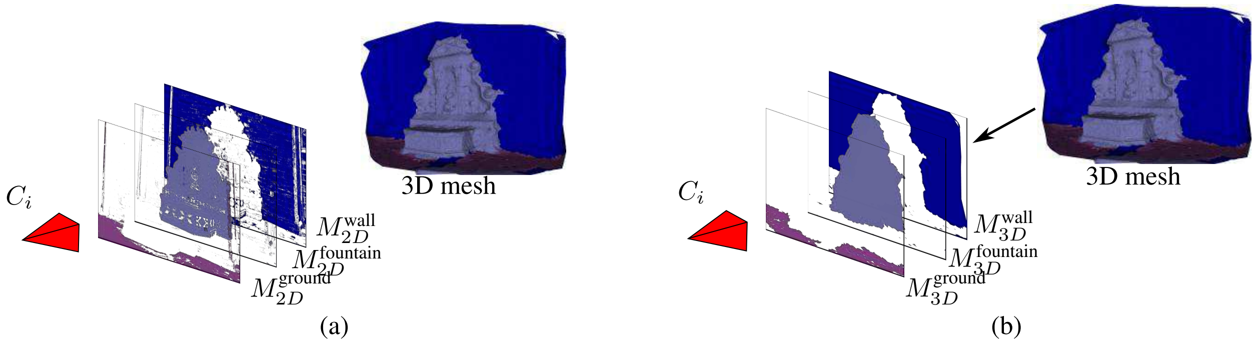


Figure 3. Masks involved in the semantic mesh refinement: for each class, we compare the masks on the left, generated from the 3D model, to the masks on the right, that come from the 2D image classification

For each camera i and for each semantic label l we have a semantic mask M^{li}_{2D} defined as $M^{li}_{2D} = 1$ where the label is equal to l, and 0 otherwise (in Figure 3(a) the binary masks are depicted in the color to discriminate the classes). For the same camera i we also project the visible part of the current 3D mesh classified as l, to form the semantic mask projection M^{li}_{3D} (see Figure 3(a)). Given these two masks, for all the cameras i we define:

$$E^{l}_{\text{sem}} = \sum_i \int_I err^{\text{sem}}_{M^{li}_{2D},M^{li}_{3D}}(x_i)\mathrm{d}x_i, \tag{11}$$

that we minimize descending the discrete gradient defined over the whole image plane I of the i-th camera:

$$\begin{aligned}\frac{\mathrm{d}E_{\text{sem}}(S)}{\mathrm{d}X_i} &= \int_S \phi_i(x)\nabla E_{\text{sem}}(x)\mathrm{d}x = \\ &= -\sum_{i,j}\int_I \phi_i(x) f_i(x_i)/(\overrightarrow{n}^T\mathbf{d}_i)\overrightarrow{n}\mathrm{d}x_i,\end{aligned} \tag{12}$$

$$f_i(x_i) = \partial_2 err^{\text{sem}}_{M^{li}_{2D},M^{li}_{3D}}(x_i) DI_i(x_i) D\Pi_i(x)\mathbf{d}_i. \tag{13}$$

Differently from Equation (8), here we use only a single camera i to compute the gradient.

While the typical error measure adopted for Equation (8) is the Zero mean Normalized Cross Correlation (ZNCC), here we adopt a modified version of Sum of Squared Differences (SSD); indeed the semantic masks are binary, therefore no illumination normalization and correction is needed. The standard SSD gives the same relevance to the two images. Here, instead, we have two semantic masks generated in two deeply different ways: by 2D image segmentation and by labeled mesh rendering. As stated in Section 2.1, the mesh labeling is usually more robust and less noisy than the image segmentation. To neglect these errors that would induce spurious contributions to the mesh refinement flow, we define the following measure in a window W:

$$err^{\text{sem}}_{M^{li}_{2D},M^{li}_{3D}} = \chi \sum_{(\mathbf{x})}^{W} \left(M^{li}_{2D}(\mathbf{x}) - M^{li}_{3D}(\mathbf{x})\right)^2, \tag{14}$$

Table 1. Resolutions and output statistic for each dataset we used.

	num. cameras	image resolution	num. facets
fountain-p11	11	3072x2048	1.9M
KITTI 95	512	1242x375	2.6M
DTU 15	49	1600x1200	0.6M

where $\chi = 1$ if the window W defined over M^{li}_{3D} contains at least one pixel belonging to the class mask and one pixel outside the class mask, and $\chi = 0$ otherwise. This neglects the flow induced by the image segmentation in correspondence of mesh regions with homogeneous labeling.

As in [39] we apply a coarse to fine approach that increases the resolution of the mesh after a fixed number of iterations. The reasons are twofold: it increase too low-resolution region for the input mesh and it prevents the refinement to get stuck in local minima and therefore improve the accuracy of the final reconstruction.

On one hand the refinement process changes the shape of the mesh, while on the other hand the coarse to fine approach enhances its resolution. In both cases the labeling estimated before the refinement could become no more consistent with the refined mesh; for this reason we re-apply the mesh-based labeling presented in Section 3.1 every time we increase the resolution of the mesh.

4. Experiments

To test the effectiveness of the proposed approach we reconstructed three different sequences depicting various scenarios: fountain-p11 from the dataset presented in [32], sequence 95 of the KITTI dataset [12] and a sequence of the DTU dataset [1]. Table 1 summarizes image resolution, number of frames and number of reconstructed facets for each dataset. We run the experiments on a laptop with a Intel(R) Core(TM) i7-6700HQ CPU at 2.60GHz, 16GB of RAM and a GeForce GTX 960M.

One of the inputs of the proposed algorithm is the semantic segmentation of the images. For the fountain-p11

Table 2. Reconstruction Accuracy measured with Mean Absolute Error and expressed in mm.

	[26]	[39]	[3]	Proposed
fountain-p11	12.7	9.2	8.6	**8.5**
KITTI 95	46.7	32.8	**32.7**	**32.7**
DTU 15	2.64	2.47	2.57	**2.40**

and the DTU sequence we manually annotated a few images and we trained a Multiboost classifier [2] on them; since the KITTI sequence is more challenging, we used ReSeg [36], a Recurrent Neural Network based model trained on the Cityscapes dataset [8]. The points adopted in our modified semantic version of [26] are a combination of Structure from Motion points [21], semi-global stereo matching [17] and plane sweeping [13].

We evaluate the accuracy of the 3D reconstruction with the method described by [32]: we consider the most significant image or images and, from the same point of view, we compute and compare the depth map generated with the ground truth and the reconstructed 3D model. For the fountain-p11 and DTU dataset we choose one image that captures the whole scene, and for the KITTI sequence we computed the depth maps from five images spread along the path. In Table 2 we illustrate the reconstruction errors (expressed in mm) of our method compared with the modified [26], the refinement in [39] and the joint semantic and photometric refinement presented in [3], applied to our labeled mesh: for all the three datasets our method improve the reconstruction error. This proves that the semantic information coupled with the photo-metric term, improves the convergence of the refinement algorithm.

To evaluate the quality of semantic labeling we project the labeled image into the same cameras we adopted to compute the depth map, and we compare them against manually annotate images. We compare against the 3D methods [26], [39] and [3] and the 2D semantic segmentation from [2] and [36], inputs of our algorithm. We show the results in Table 3: we listed several classical metric adopted in classification problems: accuracy, recall, F-score and precision. Except for the recall of the KITTI dataset, our algorithm achieves the best performances in all the datasets for each metric. This proves that the relabeling we adopted is effective and it especially regularize the labels where the noise affects the input semantic segmentation. In the KITTI dataset, where the initial image segmentations contains less noise with respect to the other dataset, the results of our refinement and [3] are very close.

4.1. Comparison with two view semantic mesh refinement

The method we presented in this paper refines the mesh accordingly to both the photometric and semantic information, in a similar yet quite different way to the very recent work appeared in [3]. For each label l defined in the image classifier, both methods compare two masks containing the pixels classified with the label l, and modify the shape of the mesh to minimize the reprojection errors of the second mask though the mesh into the first mask.

While in [3] both the first (Figure 6(a)) and the second masks (Figure 6(b)) are the outputs of the 2D image classifier, in this paper we propose a single-view method that compares the masks from camera i (Figure 6(a)) with the mask rendered from the 3D labeled mesh to the same point of view of camera i (bottom of Figure 6(d)).

To verify that, as stated in Section 2.1, our method is robust to the noise and errors that often affect the image segmentations, we implemented the method [3]. We applied it to the façade masks obtained from Figure 6(a) and Figure 6(b); Figure 6(c) shows the mask of non-zero gradients. On the other hand, in Figure 6(e), we compute the gradients with our method by comparing the masks from Figure 6(a) and the rendered mesh in Figure 6(d).

Figure 6(e) shows that the method in [3] cumulate the noise from Figure 6(a) and Figure 6(b); all the contributions outside the neighborhood of the real class borders are the consequences of misclassification in the two compared masks, therefore they evolve the mesh incoherently. These errors cumulate cross all the pairwise comparison since the classification errors are different for each view and the pairwise contributions corresponding to their location in general are not mutually compensated along the sequence. Even if the smoothing term of the refinement diminish these errors, they affect the final reconstruction. As a further proof, in Table 2 and in Table 3 our approach overcome the one in [3] especially in the DTU dataset, where the segmented images are very noisy.

Instead, our method computes a cleaner gradient flow (Figure 6(e)) thanks to the comparison with the mask rendered from the labeled mesh, that, after the MRF labeling, is robust to noise and errors.

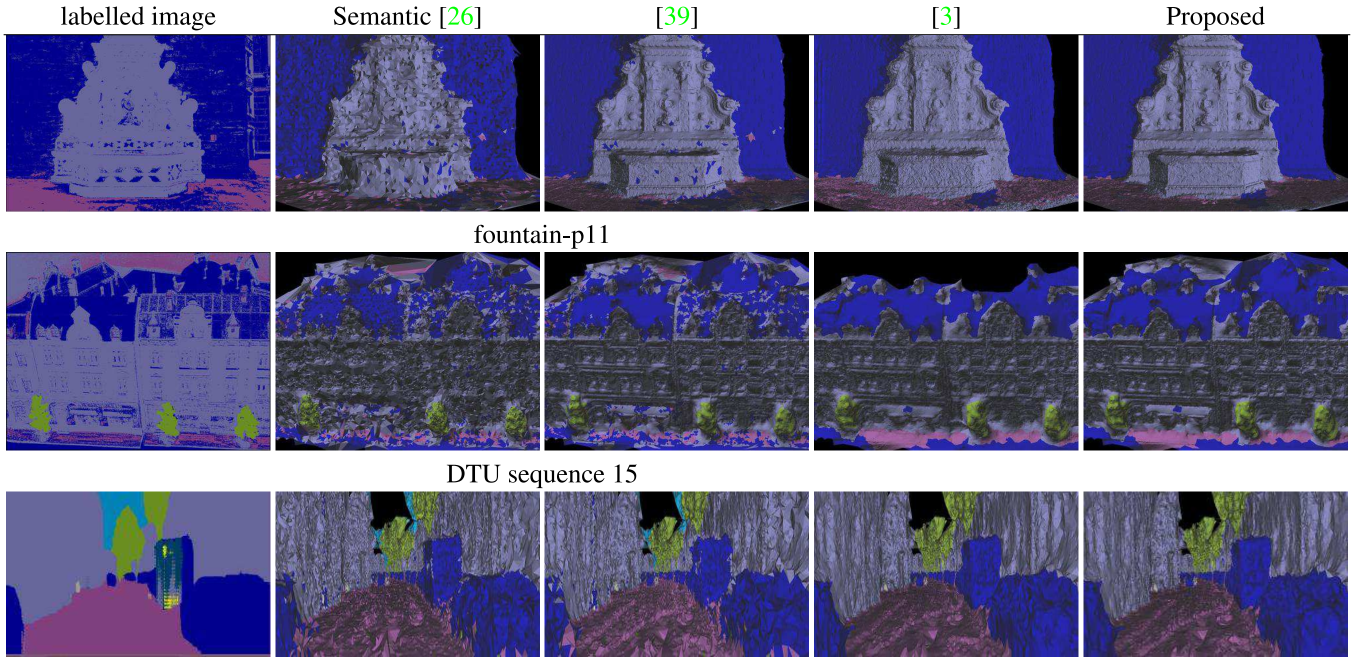
5. Conclusions and Future works

In this paper we presented a novel method to refine a semantically annotated mesh through single-view variational energy minimization coupled with the photo-metric term. We also propose to update the labels as the shape of the reconstruction is modified, in particular our contribution in this case is a MRF formulation that takes into account class-specific normal prior that is estimated from the existing annotated mesh instead of the handcrafted or learned priors proposed in the literature.

The refinement algorithm proposed in this paper could be further extended by adding geometric priors or we could investigate how it can enforce the convergence in challenging dataset, e.g., when the texture is almost flat. We also plan to evaluate how the accuracy of the initial mesh could

Table 3. Segmentation statistics.

		accuracy	recall	F-score	precision
Fountain	Multiboost [2]	0.9144	0.8495	0.8462	0.8594
	Semantic [26]	0.9425	0.8318	0.8592	0.9145
	[39]	0.9400	0.8256	0.8533	0.9095
	[3]	0.9532	0.8679	0.8923	0.9295
	Proposed	**0.9571**	**0.8755**	**0.9003**	**0.9385**
DTU	Multiboost [2]	0.9043	0.7230	0.6991	0.6837
	Semantic [26]	0.9204	0.6753	0.6837	0.7241
	[39]	0.9226	0.6617	0.6782	0.7311
	[3]	0.9551	0.7843	0.7920	0.8242
	Proposed	**0.9561**	**0.7935**	**0.8000**	**0.8329**
KITTI 95	ReSeg [36]	0.9700	0.9117	0.9092	0.9140
	Semantic [26]	0.9668	0.9093	0.8968	0.8906
	[39]	0.9672	0.9107	0.8984	0.8922
	[3]	**0.9709**	**0.9246**	0.9107	0.9084
	Proposed	**0.9709**	0.9241	**0.9109**	**0.9089**



labelled image

Semantic [26]

[39]

[3]

Proposed

fountain-p11

DTU sequence 15

KITTI sequence 95

Figure 4. Results on fountain-p11, DTU and KITTI datasets

affect the final reconstruction with or without the semantic refinement term.

Acknowledgments

This work has been supported by the “Interaction between Driver Road Infrastructure Vehicle and Environment (I.DRIVE)” Inter-department Laboratory from Politecnico di Milano, and the “Cloud4Drones” project founded by EIT Digital. We thank Nvidia who has kindly supported our research through the Hardware Grant Program.

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. 5

[2] 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. 1, 6, 7

[3] M. Blaha, M. Rothermel, M. R. Oswald, T. Sattler, A. Richard, J. D. Wegner, M. Pollefeys, and K. Schindler. Semantically informed multiview surface refinement. *Inter-*

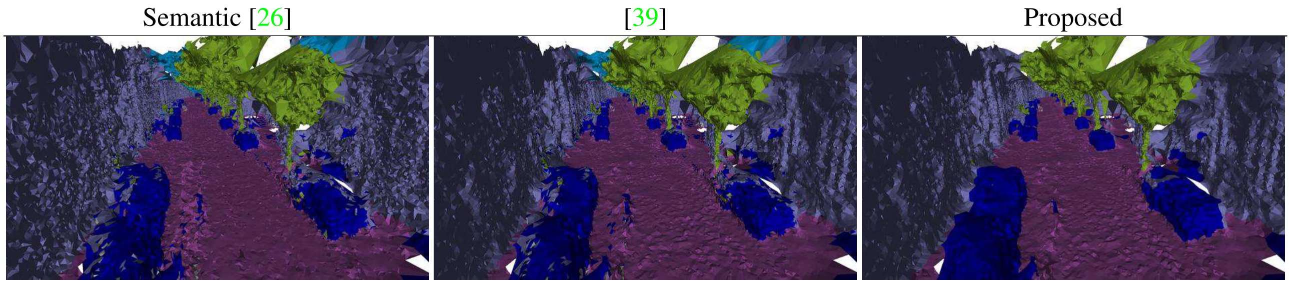
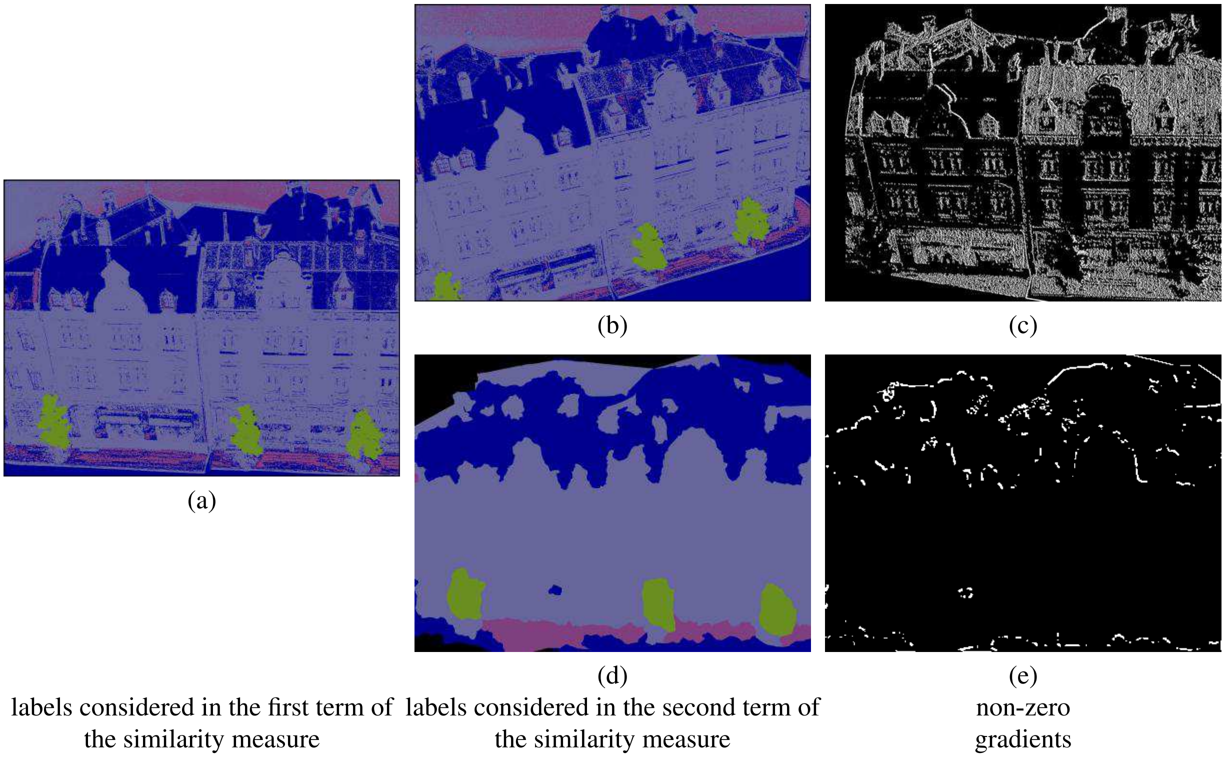


Figure 5. A wide view of the KITTI reconstruction



(a)

(b)

(c)

(d)

(e)

labels considered in the first term of the similarity measure

labels considered in the second term of the similarity measure

non-zero gradients

Figure 6. Comparison of the gradients computed by the method presented in [3] (top) and the method we presented in this paper (bottom). In the first and second columns we show the two terms compared by the similarity measure; in the third column the resulting gradients. Notice that our method uses a single point of view.

national Journal of Computer Vision*, 2017. 1, 2, 3, 4, 6, 7, 8

[4] 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*, pages 3176–3184, 2016. 1, 2

[5] R. Cabezas, J. Straub, and J. W. Fisher. Semantically-aware aerial reconstruction from multi-modal data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2156–2164, 2015. 2

[6] 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. 1

[7] I. Cherabier, C. Häne, M. R. Oswald, and M. Pollefeys. Multi-label semantic 3d reconstruction using voxel blocks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 601–610. IEEE, 2016. 1, 2

[8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 6

[9] A. Delaunoy and M. Pollefeys. Photometric bundle adjustment for dense multi-view 3d modeling. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1486–1493. IEEE, 2014. 2

[10] A. Delaunoy, E. Prados, P. G. I. Piracés, J.-P. Pons, and

P. Sturm. Minimizing the multi-view stereo reprojection error for triangular surface meshes. In *BMVC 2008-British Machine Vision Conference*, pages 1–10. BMVA, 2008. 2

[11] P. Gargallo, E. Prados, and P. Sturm. Minimizing the reprojection error in surface reconstruction from images. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007. 2

[12] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 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 *3D Vision (3DV), 2014 2nd International Conference on*, volume 1, pages 57–64. IEEE, 2014. 6

[14] C. Häne and M. Pollefeys. An overview of recent progress in volumetric semantic 3d reconstruction. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 3294–3307. IEEE, 2016. 2

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

[16] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3d scene reconstruction and class segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 97–104. IEEE, 2013. 4

[17] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2008. 6

[18] 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

[19] S. Li, S. Y. Siu, T. Fang, and L. Quan. Efficient multiview surface refinement with adaptive resolution control. In *European Conference on Computer Vision*, pages 349–364. Springer, 2016. 2

[20] V. Litvinov and M. Lhuillier. Incremental solid modeling from sparse structure-from-motion data with improved visual artifacts removal. In *International Conference on Pattern Recognition (ICPR)*, 2014. 2

[21] P. Moulon, P. Monasse, R. Marlet, and Others. Openmvg. an open multiple view geometry library. `https://github.com/openMVG/openMVG`. 6

[22] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. Dtam: Dense tracking and mapping in real-time. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2320–2327. IEEE, 2011. 2

[23] M. Pollefeys, D. Nistér, J.-M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S.-J. Kim, P. Merrell, et al. Detailed real-time urban 3d reconstruction from video. *International Journal of Computer Vision*, 78(2-3):143–167, 2008. 2

[24] J.-P. Pons, R. Keriven, and O. Faugeras. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. *International Journal of Computer Vision*, 72(2):179–193, 2007. 2

[25] A. Romanoni, A. Delaunoy, M. Pollefeys, and M. Matteucci. Automatic 3d reconstruction of manifold meshes via delaunay triangulation and mesh sweeping. In *Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016. 2

[26] A. Romanoni and M. Matteucci. Incremental reconstruction of urban environments by edge-points delaunay triangulation. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 4473–4479. IEEE, 2015. 2, 3, 6, 7, 8

[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. 3

[28] N. Savinov, L. Ladicky, C. Hane, 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, 3

[29] T. Schops, T. Sattler, C. Hane, and M. Pollefeys. 3d modeling on the go: Interactive 3d reconstruction of large-scale scenes on mobile devices. In *3D Vision (3DV), 2015 International Conference on*, pages 291–299. IEEE, 2015. 2

[30] S. Sengupta, E. Greveson, A. Shahrokni, and P. H. Torr. Urban 3d semantic modelling using stereo vision. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 580–585. IEEE, 2013. 1

[31] F. Steinbrucker, J. Sturm, and D. Cremers. Volumetric 3d mapping in real-time on a cpu. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 2021–2028. IEEE, 2014. 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. 1, 5, 6

[33] E. Tola, C. Strecha, and P. Fua. Efficient large-scale multiview stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 23(5):903–920, 2012. 2

[34] 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. 3

[35] Y. Verdie, F. Lafarge, and P. Alliez. Lod generation for urban scenes. Technical report, Association for Computing Machinery, 2015. 3

[36] 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. 1, 6, 7

[37] G. Vogiatzis, P. H. Torr, and R. Cipolla. Multi-view stereo via volumetric graph-cuts. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 391–398. IEEE, 2005. 2

[38] H. H. Vu. *Large-scale and high-quality multi-view stereo*. PhD thesis, Paris Est, 2011. 2

[39] H. H. Vu, P. Labatut, J.-P. Pons, and R. Keriven. High accuracy and visibility-consistent dense multiview stereo. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(5):889–901, 2012. 2, 3, 4, 5, 6, 7, 8

[40] M. Wardetzky, S. Mathur, F. Kälberer, and E. Grinspun. Discrete laplace operators: no free lunch. In *Symposium on Geometry processing*, pages 33–37, 2007. 4

[41] A. Yezzi and S. Soatto. Stereoscopic segmentation. *International Journal of Computer Vision*, 53(1):31–43, 2003. 2

[42] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015. 1