Semantic Segmentation of 3D Textured Meshes for Urban Scene Analysis

Mohammad Rouhani, Florent Lafarge, Pierre Alliez

Semantic Segmentation of 3D Textured Meshes for Urban Scene Analysis

Mohammad Rouhani, Florent Lafarge & Pierre Alliez

Inria Sophia Antipolis - Méditerranée

Abstract

Classifying 3D measurement data has become a core problem in photogrammetry and 3D computer vision, since the rise of modern multiview geometry techniques, combined with affordable range sensors. We introduce a Markov Random Field-based approach for segmenting textured meshes generated via multi-view stereo into urban classes of interest. The input mesh is first partitioned into small clusters, referred to as superfacets, from which geometric and photometric features are computed. A random forest is then trained to predict the class of each superfacet as well as its similarity with the neighboring superfacets. Similarity is used to assign the weights of the Markov Random Field pairwise-potential and accounts for contextual information between the classes. The experimental results illustrate the efficacy and accuracy of the proposed framework.

Keywords: Semantic Segmentation, Textured Meshes, Urban Scene, Markov Random Fields, Random Forest, Joint Label Prediction, Contextual Information.

1. Introduction

The recent advances on Multi-View Stereo (MVS) imagery (Seitz et al., 2006; Vu et al., 2009) make it possible to generate in routine dense meshes from airborne images acquired on large-scale urban scenes. Several commercial solutions such as ContextCapture (Acute3D/Bentley) and Pix4Dmapper (Pix4D) generate meshes with greater geometric accuracy and completeness than the common digital surface models. Contrary to LIDAR scans, such dense meshes represent 2-manifold surfaces, and do not require the interpolation of sampled points. As depicted by Figure 1, these meshes exhibits

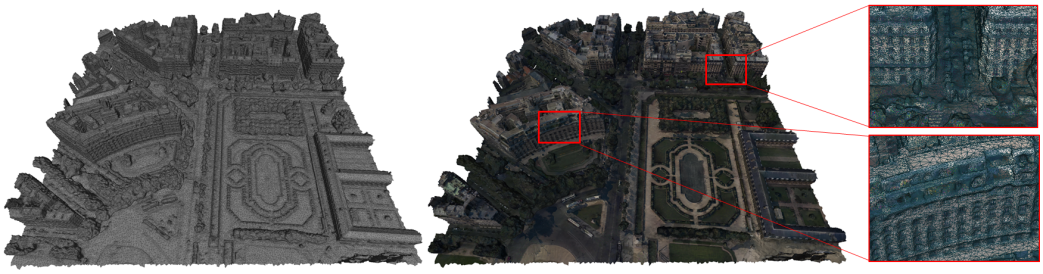

Figure 1: Textured meshes produced from MVS imagery. Our input is a textured mesh combining geometry (left) and radiometry (right).

geometric details both on roofs and facades as MVS systems can altogether deal with oblique and vertical airborne images.

Dense meshes from multiview stereo imagery are relevant 3D representations for visualization-based applications such as navigation or augmented reality. They are however too raw for applications that require additional structure and semantic information to interpret the represented scenes. There is a dire need to recover the nature of the objects composing the scenes.

We propose a dense classification algorithm that infer the class of urban objects in such dense meshes. Departing from previous work, our approach leverages both radiometric and geometric information in a supervised manner. In addition, we learn the contextual knowledge required to recover additional coherence between the urban classes of interest.

2. Related Work

We review below the most recent and related works on semantic segmentation, with focus on methods dealing with meshes of urban scenes.

2.1. Classification

Many different classification approaches have been used in photogrammetry and computer vision in order to partition images or point clouds, and to identify the nature of each area. Classification approaches differ mainly in the level of supervision: Supervised algorithms require a training set to learn how to correctly classify data, whereas unsupervised approaches require tuning the model parameters. Classification approaches also differ in the use of spatial dependencies and contextual information. In Markov Random Fields (MRFs) and Conditional Random Fields (CRFs) for instance, a datum is

not classified independently from the rest of the data: the classification decision relies upon non-local information that accounts for spatial consistency between neighboring areas.

Supervised learning. Among the many classification approaches, Texton Boost and Texton Forest can be directly applied to the input data without requiring any type of feature descriptor (Shotton et al., 2008). In addition, these methods are extremely fast as they avoid computing filter-bank responses. A multi-feature variant of TextonBoost Sengupta et al. (2013) is used for labeling each image of the stereo pair and fuse them into a scene. Both the image and geometric features have been used to train a JointBoost classifier (Valentin et al., 2013) for segmenting RGB-D images. Xiao and Quan (2009) use a series of one-vs-all AdaBoost classifiers to perform multiview semantic segmentation. In contrast, Kalogerakis et al. (2010) learn a label compatibility function between the neighboring segments of an input mesh through training a JointBoost classifier on the pairwise geometric features. Randomized decision forests (or Random Forests) are popular parametric classifiers for the segmentation and regression tasks (Zhang et al., 2010a). They are relevant for real-time applications as they generate label predictions very efficiently through performing few simple tests on the query data. Kähler and Reid (2013) employ random forests for segmenting RGB-D images; the feature vectors describe photometric and geometric information for every segment and pair of segments.

MRFs and CRFs. MRF or CRF formulations usually rely on an energy minimization problem. The energy is commonly composed of two terms: a data term that measures the coherence of each datum with respect to a label, and a pairwise potential that favors label smoothness. A supervised classifier can be used as prediction function to model the unary data term of MRFs and CRFs models. The contextual information provides relevant clues for improving the results of semantic segmentation. Co-occurrence statistics are modeled in (Ladicky et al., 2013) through a global potential function recording which pairs of classes are likely to occur in the same image. Galleguillos et al. (2008) model the co-occurrence and relative locations through the following pairwise term: Four different types of relative locations are considered (*above, bellow, inside and around*) to capture the spatial context through frequency matrices that record the likelihood of two labels to appear in a relative position. Myeong et al. (2012) propose a pairwise cost function

based on a similarity graph that encodes the relationship between two regions through context links. A context learning algorithm estimates the strength of a context link in the query image.

Yao et al. (2012) introduces auxiliary variables to consider different terms altogether, such as segmentation energy, object-reasoning as well as scene and class presence potentials; the relation between these potentials are formulated in a general CRF model. Global and local contexts have been modeled in (Mottaghi et al., 2014) to improve both semantic segmentation and object detection. The primer considers the presence or absence of a class in the scene, while the latter refers to the classes appearing in the vicinity of the object.

Formulating the segmentation problem as MRFs makes it possible to leverage many efficient inference algorithms to find the optimal labeling. Simulated annealing (Kähler and Reid, 2013), range/swap-based approaches (Liu et al., 2015), or mean-field approximation (Krähenbühl and Koltun, 2011) are relevant inference approaches when the configuration space is large. Inferring on global co-occurrences is commonly formulated as integer programming, and solved via linear relaxation (Ladicky et al., 2013). Yao et al. (2012) relies upon a message-passing algorithm to solve a holistic CRF that includes contextual terms such as scene and class-presence potentials.

2.2. Mesh Segmentation

Mesh segmentation is still a research challenge in geometry processing, robotic, and computer vision (Shamir, 2008; Chen et al., 2009; Theologou et al., 2015). In its simplest form mesh segmentation may be seen as an unsupervised clustering problem based on specific geometric criteria (Shlafman et al., 2002). Region growing (Page et al., 2003) and spectral analysis (Zhang et al., 2010b) are other instances of such deterministic approaches for mesh segmentation. The probabilistic approaches such as MRFs or CRFs provide an efficient means to enforce spatial consistency (Lafarge et al., 2010). Designing the energy terms for such approaches ranges from totally unsupervised (Verdie et al., 2015) to supervised (Van Kaick et al., 2011), through semi-supervised (Lv et al., 2012).

A key aspect of the mesh segmentation approaches is the design of feature vectors encoding geometric information such as normals, curvatures or planarity. For textured meshes, additional features based on photometric information such as colors or texture histograms, provide relevant clues to design the energy terms of MRFs and CRFs. Verdie et al. (2015) design three

geometric attributes to define the unary term of a MRF. In our approach we instead adopt a supervised approach to learn an effective classifier.

In MVS contexts, some methods first perform classification directly from the images before mapping it to the output 3D model (He and Upcroft, 2013). In (Sengupta et al., 2013), each image is labeled by a supervised CRF before mapping and fusing the sets of labels with the mesh. Lafarge et al. (2013) propose a hybrid approach in which the output model is progressively refined while detecting regular urban entities. Kundu et al. (2014) proposes a joint CRF model defined in 3D (volumetric) space and infers altogether the semantic label and voxel occupancy. Savinov et al. (2015) and Blaha et al. (2016) define the unary cost function along rays by considering both the semantic label and the depth of the first occupied voxel along the ray. Referred to as semantic 3D reconstruction by Haene et al. (2013), such approaches are typically memory intensive and require complex inference approaches. The incremental approach proposed by Vineet et al. (2015) operates in near real time and delivers a rough reconstruction with a street-based semantics. Xiao and Quan (2009) propose a larger MRF that includes all the views, and models all connections between the associated areas. The smoothness term between two views is defined either based on the color similarity or using the number of common feature tracks between the two associated images. These image-based methods are compute-intensive and insufficiently exploit the geometric properties of the observed scene.

Literature on the segmentation of textured meshes is marginal. Chauve et al. (2009) proposed an interactive approach based on transductive segmentation. In this approach the classifier is learned from basic photometric and geometric features derived from a set of strokes sketched on the input textured mesh.

2.3. Urban Modeling

Urban modeling is a notoriously hard challenge in photogrammetry and computer vision (Musialski et al., 2013), and extracting semantics from raw input data is a major step of this challenge. In an urban context, the classes of interest mainly correspond to the stationary objects such as buildings, roads and trees. We refer the reader to two comprehensive surveys (Rottensteiner et al., 2012; Schindler, 2012). From images (Volpi and Ferrari, 2015; "Montoya-Zegarra" et al., 2015) or from 3D point clouds (Lai et al., 2014; Niemeyer et al., 2014), the non-local approaches that exploit spatial and contextual dependencies of urban scenes have shown highly effective.

Some approaches as (Cabezas et al., 2015) exploit multiple sources of data for reconstructing urban scenes with more semantic information. In real-world scenarios, the availability of these different types of data is however mostly unlikely.

Literature on classification of urban meshes is more marginal. Verdie et al. (2015) segment 3D urban scenes into different semantic parts such as *ground*, *facade*, *roof* and *vegetation*. In addition, the scene is represented with different levels of details derived from the classification. In this approach several geometric attributes based on elevation, planarity and verticality are used to design a labeling cost function. In contrast, Martinovic et al. (2015) exploits a supervised classifier trained to define a labeling cost function. Some weak architecture rules are then enforced as a post-processing step by iterative quadratic programming.

3. Positioning and Contributions

MVS meshes are a recent type of input data in city modeling, and the semantic segmentation of these meshes has been overlooked. Most of the few existing approaches such as (Verdie et al., 2015), suffer from several issues, including (i) frequent labeling errors between some urban classes, in particular between roof and vegetation, (ii) trial-and-error processes required to tune the parameters, and (iii) ad-hoc post-processing operations required to exploit the local context.

Departing from the work of Verdie et al. (2015), we propose a series of technical improvements to alleviate some of the aforementioned issues. Although the use of (i) region decomposition for reducing the algorithmic complexity, and (ii) a general MRF formulation for imposing spatial consistency is kept, we propose three important contributions:

- **Joint Photometric and geometric analysis.** Both geometric and photometric features are used for local analysis and description of the input meshes. In addition, some geometric features are extracted at different scales to increase robustness to the diversity of urban landscapes. Contrary to Verdie et al. (2015) that exploit some mono-scale geometric features only, this richer set of features allows urban classes to be better discriminated, in particular parts of trees and pieces of ground from hilly areas mislabeling as roof.

- **Supervision.** A supervised approach based on random forests is used for building a local classifier in reasonable computational times. Contrary to the unsupervised approach of Verdie et al. (2015), our approach has a limited number of parameters, and in particular, does not require tuning weighting coefficients between the different features. This design choice brings both more flexibility to the classification procedure and more robustness to data variety.

- **Joint labeling.** Departing from common labeling approaches in which a label corresponds to one of the classes of interest, our label space also takes into account the class transition between neighboring areas. This provides us with a means to account for contextual knowledge in an intuitive and effective way. In particular, the weights of the pairwise potential are directly learned from a training set to favor certain configurations of adjacent classes. For instance, roof is more likely to be adjacent to facade than to ground or vegetation.

As illustrated by Figure 2, the input textured mesh is first partitioned into small segments, referred to as superfacets. A set of geometric and photometric features relevant for discriminating urban objects are then computed from each superfacet. The urban class of each superfacet is then predicted using a random forests classifier within a MRF formulation.

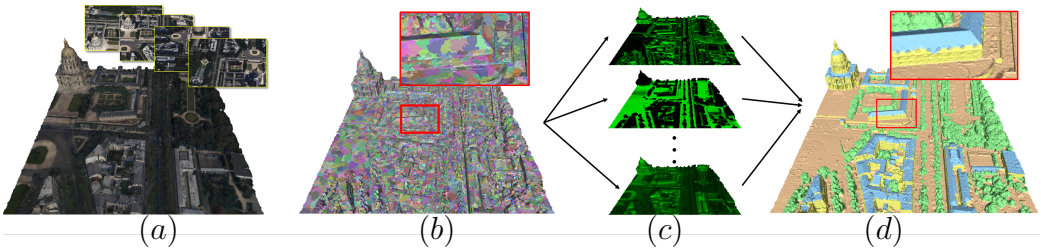

Figure 2: Overview. The input textured mesh generated from MVS images (a) is first partitioned into superfacets. (b) A set of geometric and photometric features are then extracted from the superfacets, (c) and combined into a MRF-based formulation to classify them at a non-local scale. (d) The class *roof* is depicted in blue, *ground* in brown, *tree* in green, and *facade* in yellow. This color code is used throughout the paper.

4. Proposed Approach

We now detail the three main steps of our approach, as illustrated by Figure 2. The third step, i.e., MRF-based labeling, constitutes our main contribution.

4.1. Superfacet Partitioning

As the input meshes are very dense, typically 12M facets representing $1km^2$ of a urban scene, labeling each facet of such meshes is very compute-intensive. To lower the computing times and improve scalability, we first group facets into small clusters referred to as *superfacets*. Such a clustering step is analogous to over-segmenting images into superpixels, a popular approach for image analysis.

Clustering is performed through a region growing approach based on two similarity measures: (i) a covariance-based shape operator for measuring the geometric similarities between the triangular facets (Cohen-Steiner et al., 2004), and (ii) the $L1$-distance measured in color space for quantifying the photometric similarity. The former favors the grouping of coplanar facets whereas the latter helps preserving image discontinuities of the texture map. A region grows until one of the two similarity measures exceeds user-specified thresholds (for both geometry and photometry). In addition, the maximum area of a superfacet can be determined by an optional user-specified parameter, as in superpixel approaches. Figure 2-(b) illustrates the superfacets obtained through clustering.

Two superfacets are said adjacent if their borders share at least one edge from the input triangular mesh.

4.2. Feature Extraction

For each superfacet of the input mesh we compute a set of geometric and photometric features. These features, which correspond to local statistics over the geometry of the superfacets and their color distributions, are later used by the classification step to discriminate between the urban objects of interest.

Geometric Features. These features are mostly similar to the ones proposed by Verdie et al. (2015). We compute three features relating to elevation, planarity and horizontality:

- **Elevation** a_e**.** This feature discriminates the ground from higher objects such as roofs. It measures the height of the superfacet with respect to the ground:

$$a_e = \sqrt{\frac{z - z_{min}}{z_{max} - z_{min}}}, \tag{1}$$

where z denotes the mean Z-coordinate of the superfacet, and z_{min} (resp. z_{max}) denotes the minimal (resp. maximal) Z-coordinate of the superfacets contained in a local window on the XY plane. The size of the local window is chosen so as to trade robustness to highly variable terrain height, for accuracy of the estimation for large buildings. More specifically, the size of the local window should be larger than the largest building in the scene. We consider multiple elevation features, each being associated to a window size. The idea is to let the subsequent classifier select the window size in accordance to the urban landscapes observed in the training set. We consider three sizes by default: 10 meters, 20 meters and 40 meters.

- **Planarity** a_p**.** We measure the planarity of a superfacet through computing the variation of the covariance matrix of all facets of the superfacet:

$$a_p = 1 - \frac{3\lambda_0}{\lambda_0 + \lambda_1 + \lambda_2}, \tag{2}$$

where $\lambda_0 < \lambda_1 < \lambda_2$ denote the eigenvalues of the covariance matrix. For near-planar superfacets this feature is close to 1 as λ_0 is much smaller than the two other eigenvalues.

- **Verticality** a_v**.** This feature is computed as the cosine of the angle between the normal $\mathbf{n}$ of the superfacet and the vertical axis $\mathbf{n}_z$:

$$a_v = |\mathbf{n}.\mathbf{n}_z|. \tag{3}$$

The feature value is low for the vertical components that mainly correspond to facades.

In ideal settings the aforementioned geometric features are sufficient to discriminate ground, facade, roof and trees. On real-world datasets however,

these three features are insufficient to resolve ambiguous cases, and often confuse complex roofs and trees with regular shapes, as discussed by Verdie et al. (2015). This motivates our proposal for using additional radiometric features to improve the classification performances.

Photometric Features. The photometric information contained in the textured meshes provide a complementary clue for superfacet classification. For each superfacet we analyze its color distribution, measured in HSV color space. We choose the HVS instead of RGB color space as it is more effective for discriminating objects with different reflectivity properties. We then derive three photometric features for each superfacet: (1) the average color, (2) the standard deviation of its color distribution (see Fig 3-bottom), and (3) its color distribution. For the third feature we discretize the distribution by clustering the HSV color values of the whole texture map into a color palette. Our experiments show that such features are more discriminant than pure geometric features ones for classifying the superfacets (see Table 1).

For each superfacet we then construct a feature vector by concatenating all geometric and photometric features. We denote by $\mathbf{f}_i$ the feature vector of superfacet i.

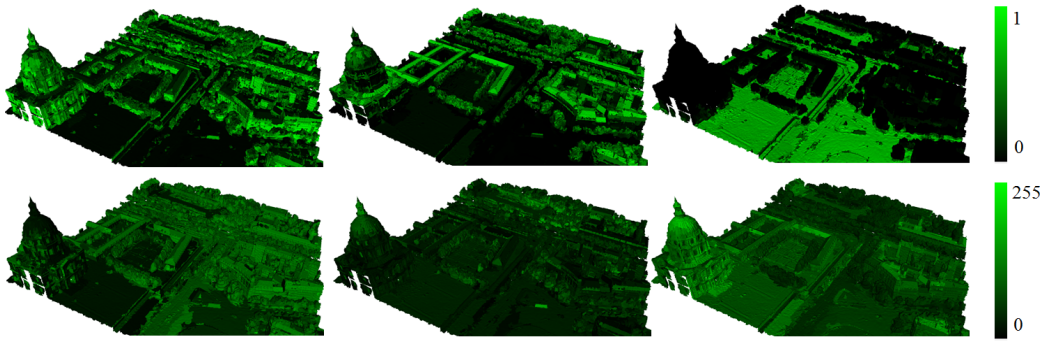

Figure 3: Geometric and photometric features. We compute the verticality, planarity and elevation are geometric features (*top*). We complement them by photometric features measured in the HVS color space (bottom).

4.3. Labeling

Label space. In previous work the label space is often limited to one label per area, associated to one of the classes of interest. We enlarge the label space

so as to include the class transition between neighboring areas, in addition to the class of each area.

Given the set of classes of interest $\mathcal{L} = \{1, 2, ..., N\}$, we consider a random variable $l_i \in \mathcal{L}$ that associates a class to superfacet i. In our experiments N is set to 4 or 5, covering the classes of *ground*, *trees*, *facades*, *roofs*, and optionally *superstructures*. Our joint labeling system extends $\mathcal{L}$ into a larger set of classes $\hat{\mathcal{L}} = \{1, 2, ..., 2N\}$ so that

$$\forall \hat{l}_{ij} \in \hat{\mathcal{L}}, \quad \hat{l}_{ij} = l_i + N.\mathbb{1}_{\{l_i \neq l_j\}}, \tag{4}$$

where j denotes the index of a superfacet adjacent to superfacet i, and $\mathbb{1}_{\{.\}}$ is the characteristic function. The joint label $\hat{l}_{ij}$ records the class of superfacet i, and whether superfacet j has the same class. While previous approaches such as (Kähler and Reid, 2013)preformed joint labeling with N^2 labels, they are very compute-intensive in both training and testing stages.

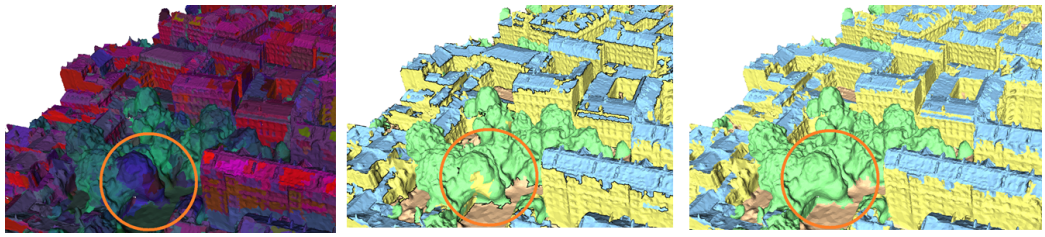

Figure 4: Joint labeling. In addition to the most probable class for each superfacet, our classifier records when a class transition is most likely to occur (see black lines, middle) given an input mesh (left, color code: HSV channels). Such an information is exploited in the MRF formulation to disambiguate complex cases, such as clusters mislabeled as facade on the vertical parts of some trees (right).

Classifier training. For training a classifier over our joint labeling approach, we concatenate feature vectors $\mathbf{f}_i$ and $\mathbf{f}_j$ into a single vector $\hat{\mathbf{f}}_{ij}$ for all $\hat{l}_{ij} \in \hat{\mathcal{L}}$. For machine learning we use randomized decision trees. They enable efficient prediction since each feature vector undergoes only a few branching nodes before reaching the leaves of the tree. In our implementation we use by default 100 decision trees with 25 as maximum depth. Figure 4 illustrates the label prediction obtained by the classifier.

MRF formulation. The trained random forest is able to predict the joint labels locally. However, the quality of label prediction may become spatially

incoherent in presence of geometric defects and texture noise as illustrated by Figure 5. We thus apply a common MRF formulation for smoothing the labels via the following energy minimization:

$$E(l) = \sum_{i \in S} \psi_i(l_i) + \gamma \sum_{(i,j) \in \mathcal{E}} \varphi_{ij}(l_i, l_j), \tag{5}$$

where S denotes the set of superfacets of the input mesh, and $\mathcal{E}$ denotes the set of pairs of neighboring superfacets. The unary data term ψ_i measures the coherence of the class l_i of superfacet i whereas the pairwise potential φ_{ij} encourages similar classes for neighboring superfacets with similar features. The regularization parameter γ balances the importance of the latter with respect to the former. Figure 6 shows the impact of this parameter on different cases. Both data term and pairwise potential are designed using the aforementioned joint labeling approach.

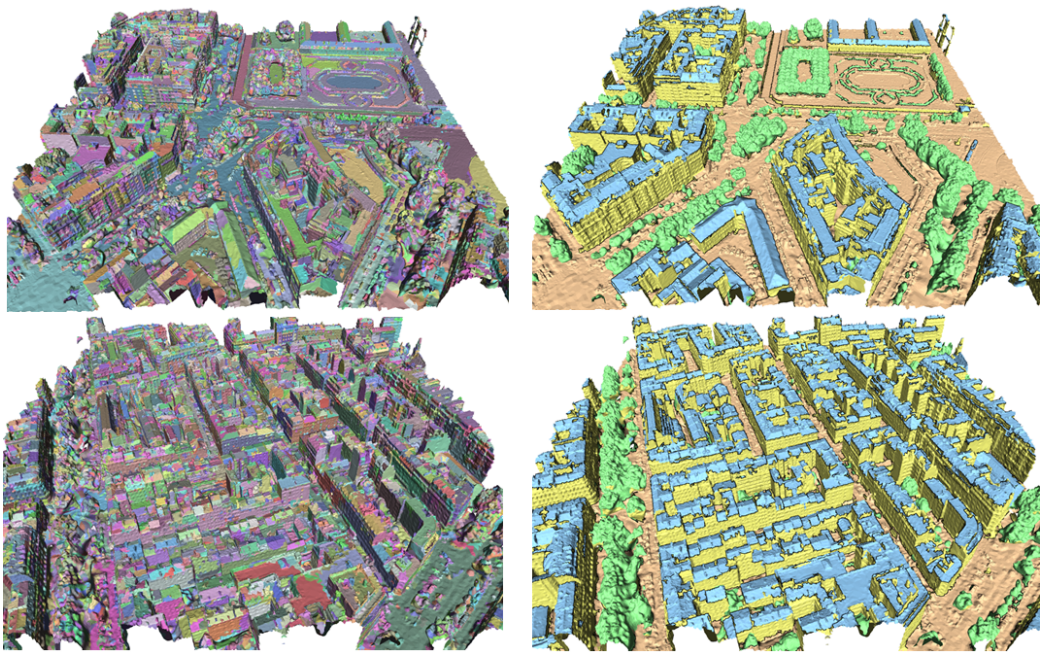

Figure 5: Local label prediction. The trained classifier alone produces decent results (right) given a superfacet partitioning (left). The subsequent use of the MRF formulation is often sufficient to resolve the small mislabeled areas on roofs and trees. The black lines underline the locations where the classifier advocates for a class transition.

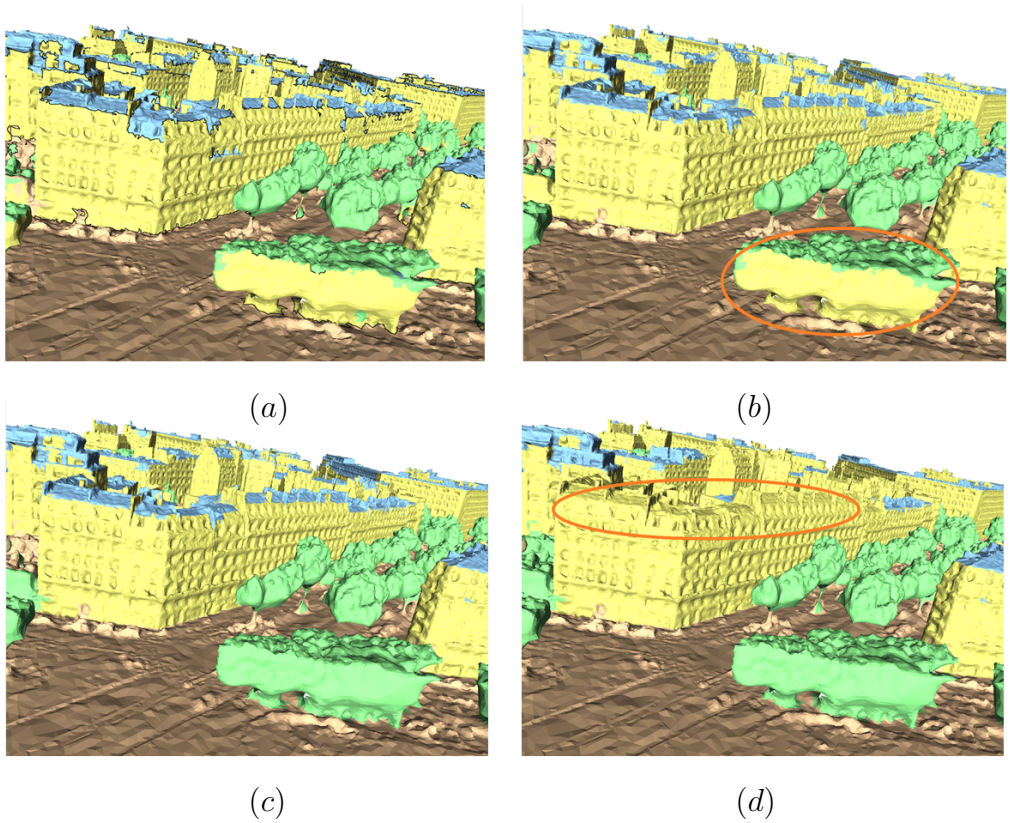

Figure 6: Impact of the regularization parameter γ. Without regularization ($\gamma = 0$), the classification result is quite noisy (a). Increasing γ to 0.2 (b) and 0.5 (c) progressively improves the result. However, a too large value such as 1 (d) may translate into over-smoothing in which some roofs are merging with facades.

Data term. The unary data term of superfacet i is computed from the output probabilities of the trained decision trees:

$$\psi_i(l_i) = -\frac{1}{|T|} \sum_{t \in T} \log(P_t(l_i)), \tag{6}$$

where T denotes the set of decision trees, and P_t denotes the prediction probability of the class l_i for the decision tree t defined as

$$P_t(l_i) = \frac{1}{|\mathcal{N}^i|} \sum_{j \in \mathcal{N}^i} \left(p_t(\hat{l}_{ij} = l_i | \hat{\mathbf{f}}_{ij}) + p_t(\hat{l}_{ij} = l_i + N | \hat{\mathbf{f}}_{ij}) \right), \tag{7}$$

where $\mathcal{N}^i$ denotes the set of neighboring superfacets of superfacet i, and p_t denotes the output probability function of the decision tree t. Intuitively, we average the prediction of the class of superfacet i over all its neighboring superfacets, without distinguishing whether the joint label must yield similar classes ($\hat{l}_{ij} = l_i$) or different classes ($\hat{l}_{ij} = l_i + N$).

Pairwise potential. This term introduces both spatial consistency between superfacets and contextual information between classes in the decision process. We express the pairwise potential term as

$$\varphi_{ij}(l_i, l_j) = \frac{1}{|T|} \sum_{t \in T} \omega_{ij} \cdot A_t(l_i, l_j), \tag{8}$$

where ω_{ij} denotes the length of the common frontier between superfacets i and j, and A_t denotes a function equating to 0 if $l_i = l_j$, and to a positive penalty otherwise. More specifically, A_t is expressed through our joint labeling approach as

$$A_t(l_i, l_j) = \begin{cases} 0 & \text{if } l_i = l_j \\ p_t(\hat{l}_{ij} = l_i + N | \hat{\mathbf{f}}_{ij}) & \text{otherwise} \end{cases} \tag{9}$$

This term acts as a generalized Potts model in which the penalty weight is defined through the output probability function of random forests.

Inference. For inference through energy minimization we use an α-expansion approach (Boykov et al., 2001). Although the input meshes are very dense, the size of our MRF formulation is moderate when applied to superfacets.

5. Experimental Results

We evaluated our approach on textured meshes generated by the ContextCapture software (Acute3D) applied to multiview aerial images shot on urban landscapes (cities of Paris and Toulouse, France). Such textured meshes correspond to a planimetric area of approximately $300m \times 300m$, and contain around 1M triangular facets and two texture maps of size 8192×8192 pixels. Four classes of interest are considered in our experiments: roof, facade, vegetation and ground.

We used a unique 3D mesh of $50K$ superfacets for training our classifier, except for the experiments conducted in Figure 15. This mesh corresponds

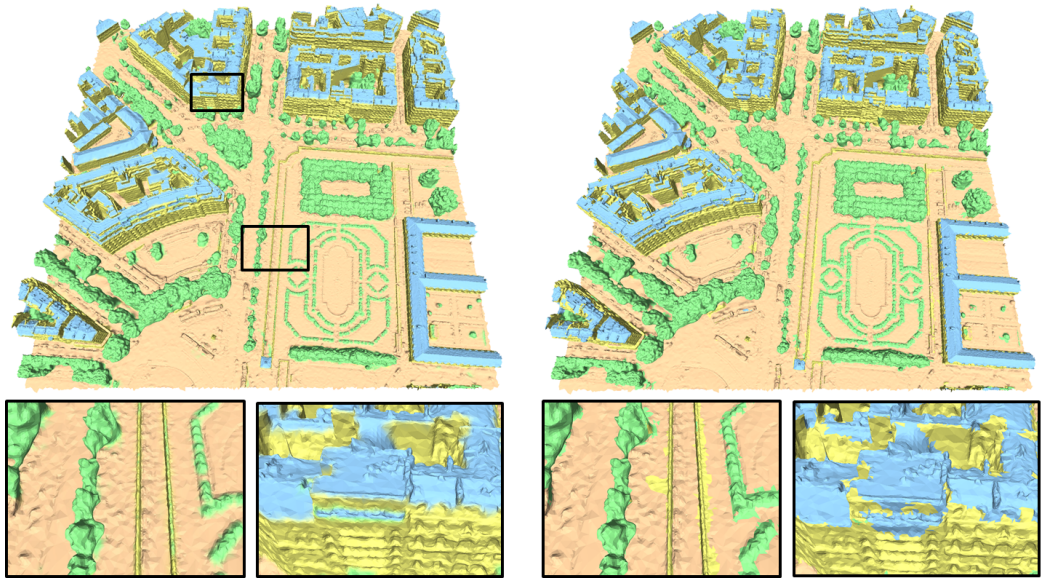

Figure 7: Training set. The training set used for our experiments corresponds to a $300m \times 300m$ tile whose superfacets have been manually labeled by an expert (left). The trained classifier is globally consistent with the training set (right). Typical learning errors are shown in the close-ups.

to a $300m \times 300m$ tile whose superfacets have been manually labeled by an expert. As illustrated in Figure 7, this tile represents a part of Paris with a few building blocks, some isolated buildings and vegetation along roads. In particular, only 10% of the large scale scenes presented in Figures 17, 18 and 19 overlap with the training set.

5.1. Parameters

Our algorithm takes as input several model parameters. To extract the superfacets, two thresholds are required to define the growing condition: the geometric threshold imposes a maximal angle between the normal of the region and the facet candidate, whereas as the photometric threshold guarantees a maximal L1-distance between the colors of the region and the candidate. In all experiments shown the geometric (respectively photometric) threshold has been set to 20 deg (respectively 30 over 256). We limit the size of superfacets to $100m^2$ in order to avoid the propagation of potential classification errors, as illustrated by Figure 8. A single textured mesh ($1M$ facets) is used to train a random forest of 100 decision trees. The recursive

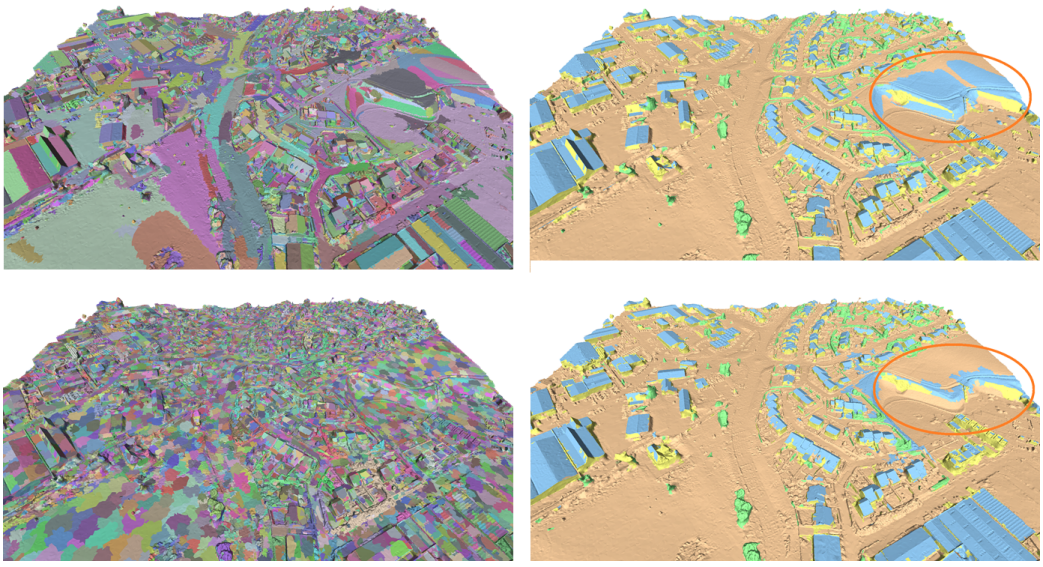

Figure 8: Size of superfacets. The use of large superfacets (top) enables fast processing. Conversely, small superfacets (bottom) often yield fewer mislabeling errors as the label propagation process operates at a more local scale. Notice the mislabeling on the hilly area.

splits in the testing nodes are stopped when the maximum depth (set by default to 25) is reached or when the number of samples is smaller than 20. Figure 9 shows the impact of the size of the training set on the label prediction accuracy during the testing stage. Parameter γ is the only tradeoff parameter of the energy. By default we set γ to 0.5 except for Figure 6.

We consider four main classes of interest: *Ground, Roof, Facades, Vegetation*. In addition, and in order to demonstrate the flexibility of our algorithm, Figure 10 showcases an experiment where an additional class is used to detect roof superstructures such as chimneys or dormer-windows. Note that despite a very rough manual labeling for the training stage, the achieved labeling results are promising. These results can be improved by defining a more complete and accurate training set with more samples of chimneys.

5.2. Robustness to Urban Landscapes

We evaluate our approach on different types of urban scenes with a wide range of geometric and photometric features. Figures 11 - 13 depict classification results for industrial, residential and dense urban landscapes, respec-

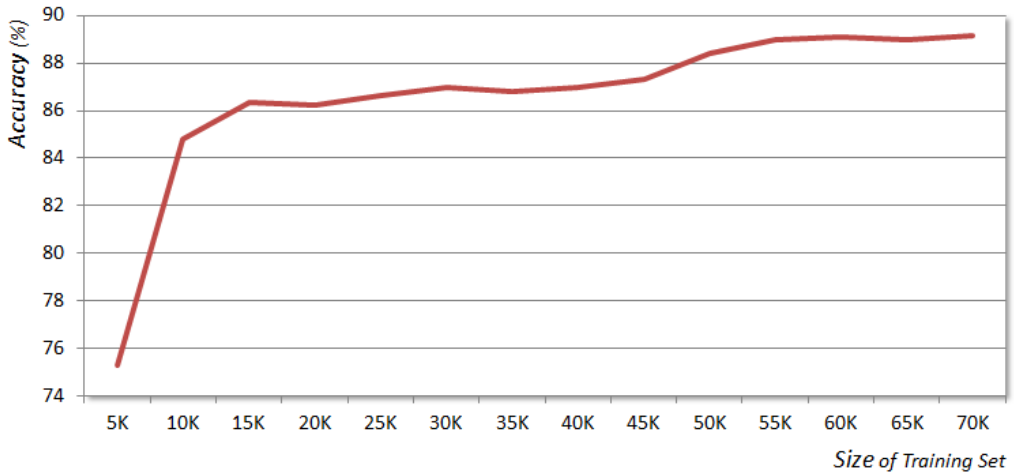

Figure 9: Classifier accuracy vs. size of training set. Using at least 10K superfacets in the training set allows the classifier to yield correct predictions for more than 85% of samples. Beyond 10K the gain obtained by increasing the number of samples is negligible.

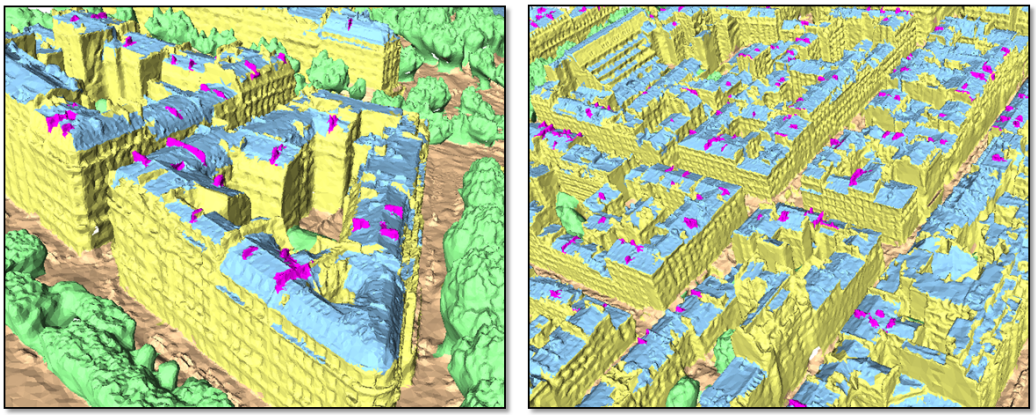

Figure 10: Adding a new class. Our formulation is flexible enough to define new classes of interest. When adding the new class *superstructure* that includes chimneys and dormer-windows, our algorithm can detect it reasonably well (see purple clusters) without adding any specific features.

tively. Our approach is sufficiently versatile to correctly classify these three types of urban landscapes.

Our multi-scale approach for computing the elevation features is particularly relevant to classify hilly areas, as illustrated by Figure 14. However,

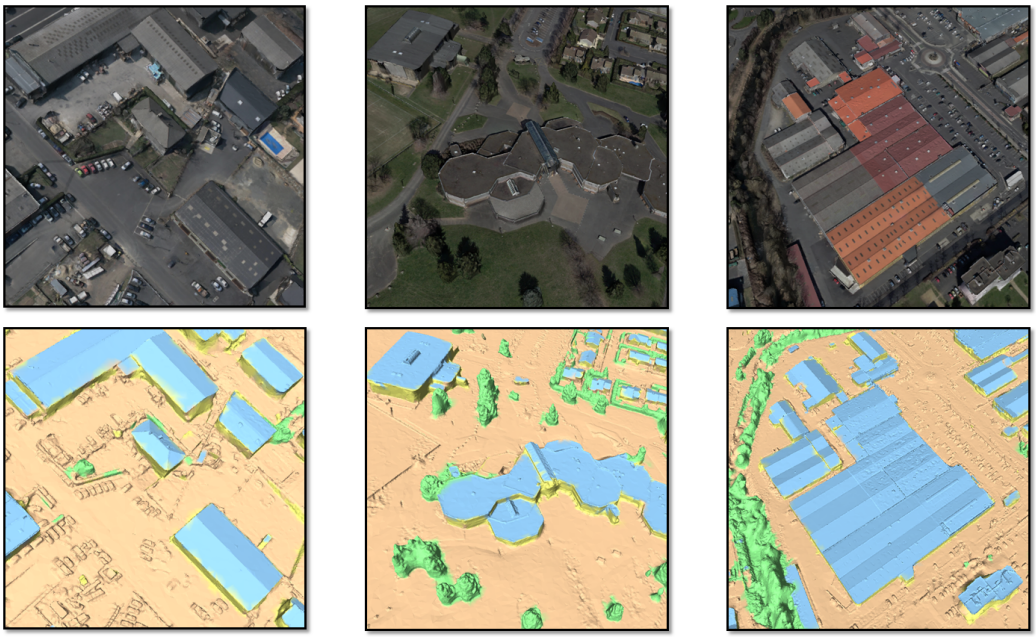

Figure 11: Classification of industrial areas. Very large industrial buildings are challenging to classify due to confusions between, e.g., roof and ground. Relying upon multi-scale elevation features yields satisfactory results on such buildings.

extremely sharp hills challenges our approach that fails to correctly discriminate ground from roof and facade. Such cases that often correspond to quarries are quite marginal in practice.

We trained some classifiers on different urban areas in order to measure the impact of urban diversity on the learning phase. As illustrated in Figure 15, the most accurate results are obtained when the classifier is trained and tested on a similar urban landscape. Note that using an industrial landscape as training set gives the most generic classifier. This choice must be considered when the observed urban landscapes in the data are unknown.

5.3. Feature Importance

Five geometric features (elevation at 3 scales, planarity and verticality) and three photometric features (color histogram, HSV average intensity and variance) are considered to describe the input surfaces locally. These features, grouped into a single feature vector of 36 components for describing each superfacet, do not exhibit the same discriminative power, referred to as importance, when training the classifier. Table 1 highlights the important

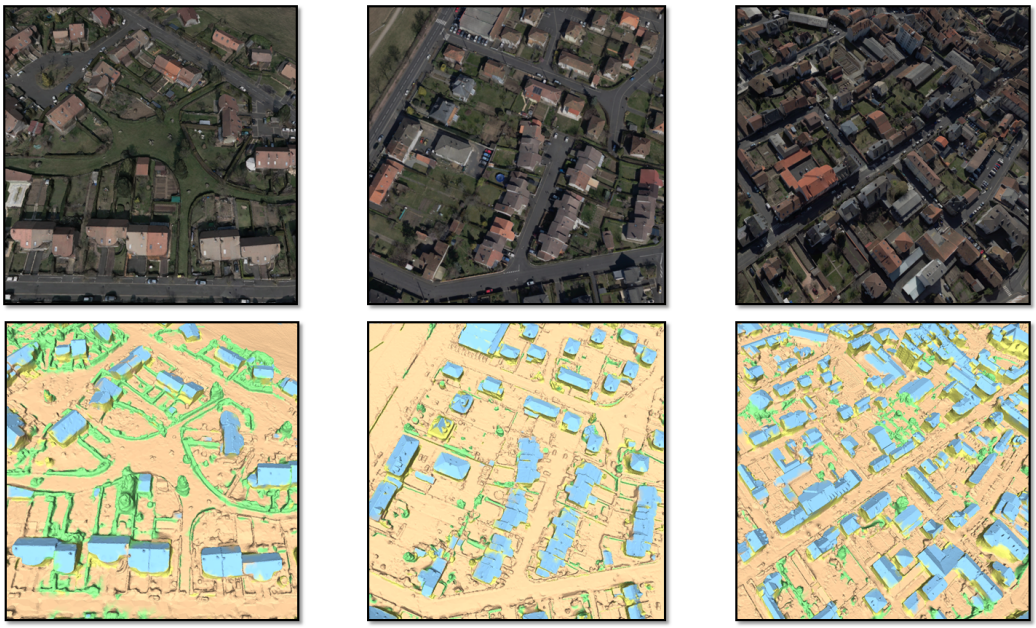

Figure 12: Classification of residential areas. Despite the small size of buildings and the bad geometry of their facade, our algorithm correctly distinguishes the classes of interest, even when trees are merging with roof and facade components.

features by showing their frequencies during the training phase. The color histogram is the most relevant feature with an occurrence above 50%. The photometric features are on average twice more important than the geometric features. Figure 16 illustrates the impact of a joint use of geometric and photometric features.

5.4. Comparisons

We compare our method on large-scale urban scenes with the unsupervised approach of Verdie et al. (2015) where the MRF data term relies upon only geometric attributes. We also perform experiment with a variant of our model that does not use the joint labeling system, but a standard system with N classes. In this model, no contextual information is taken into account. We provide qualitative comparisons in Figures 17, 18 and 19, where the input meshes (3M facets, 1.5M vertices), cover a third of one kilometer square of urban scenes. Our supervised approach, combined with a joint labeling system and both geometric and photometric attributes, outperforms the two competing approaches. More specifically, our method discriminates more

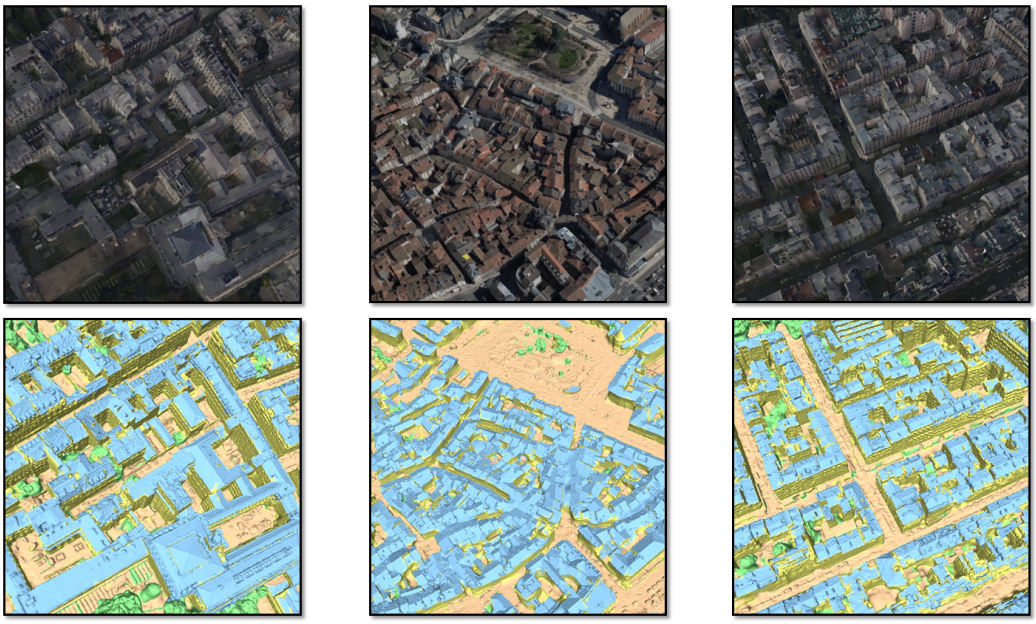

Figure 13: Classification of dense urban areas. In such urban landscapes, the main challenge is to distinguish trees from roofs and to identify ground in presence of narrow streets (see middle example). Our results are mostly correct except a few confusions for trees in inner courtyards.

Table 1: Feature importance during the random forest training. The joint feature vector $\hat{\mathbf{f}}_{ij}$ contains 72 channels for describing a superfacet i and one of its neighbors j. "dim." and "frq." show the dimension and the frequency of features, respectively.

Features	Superfacet i		Neighboring superfacet j		Pair of superfacets	
	dim.	frq.	dim.	frq.	dim.	frq.
Average HSV	3	3.72%	3	3.17%	6	6.89%
Variance HSV	3	3.85%	3	2.86%	6	6.71%
Color Histogram	25	27.36%	25	23.05%	50	50.41%
Geometric Features	5	19.38%	5	16.61%	10	35.99%
Total	36	54.31%	36	45.69%	72	100%

accurately the trees, complex roof structures, as well as the small common urban objects such as fences and cars.

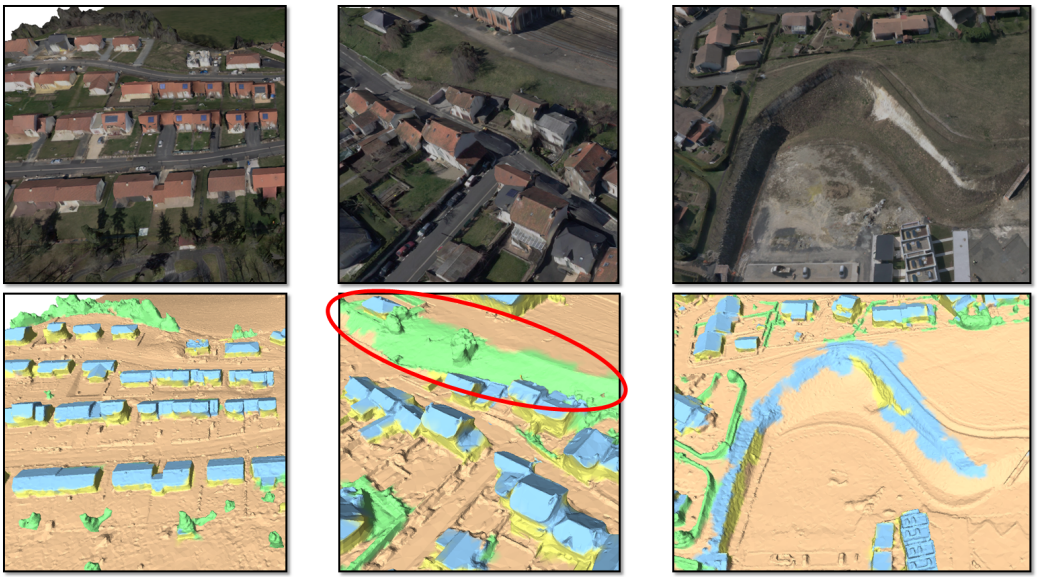

Figure 14: Classification of hilly areas. Using the elevation feature computed at different scales our classifier correctly detects sloppy ground in most cases (see left and middle). In presence of very abrupt relief changes, however, it tends to misclassify ground into facade or roof (right).

5.5. Performance

Running times of our algorithm are given for different sizes of input mesh in Table 2. Training the classifier is the most time-consuming operation. It takes close to 3 minutes, but has to be done only once. Superfacet partitioning, feature extraction and MRF-labeling require each a few seconds for a 1Mfacet input mesh. Memory peak reached from a standard $300m \times 300m$ tile is typically $3.2Gb$.

Table 2: Running times. Timings (sec) are given for a Intel Core i7 processor clocked at 2GHz.

	Superfacet partitioning	Feature extraction	Labeling (testing)	Classifier (training)
training set (1Mfacets, Fig 7)	10	16	28	175
large scene (3.4Mfacets, Fig 18)	34	45	85	-

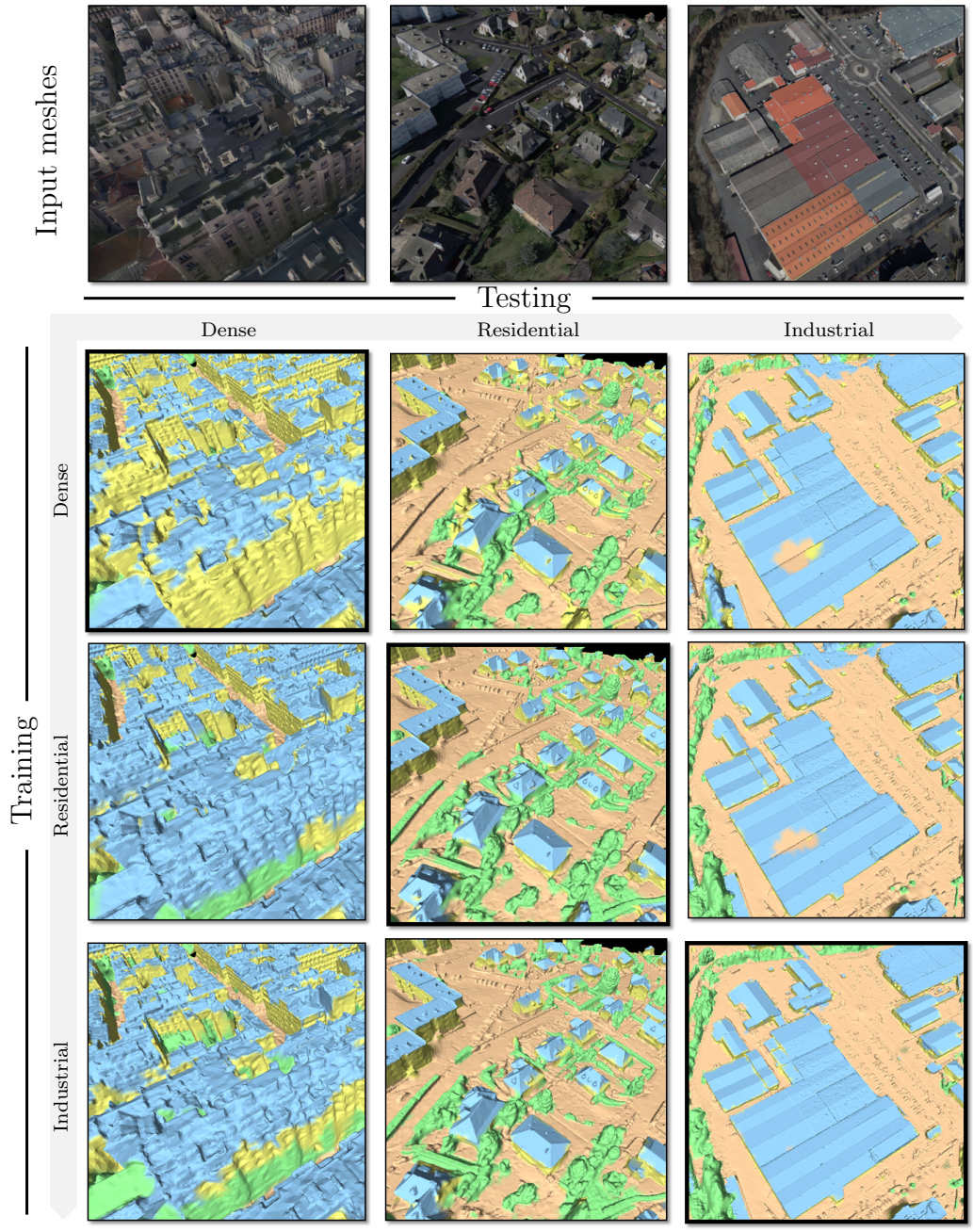

Figure 15: Impact of diversity of urban landscapes on the learning process. As expected training and testing the classifier on similar urban landscape yields the best classification results (see diagonal). As dense and residential areas have distinctive urban signatures, the classifiers learned from these landscapes is insufficiently general. Industrial landscapes, with a more diverse signature, enable the associated classifier to also perform well on other landscapes (see bottom row).

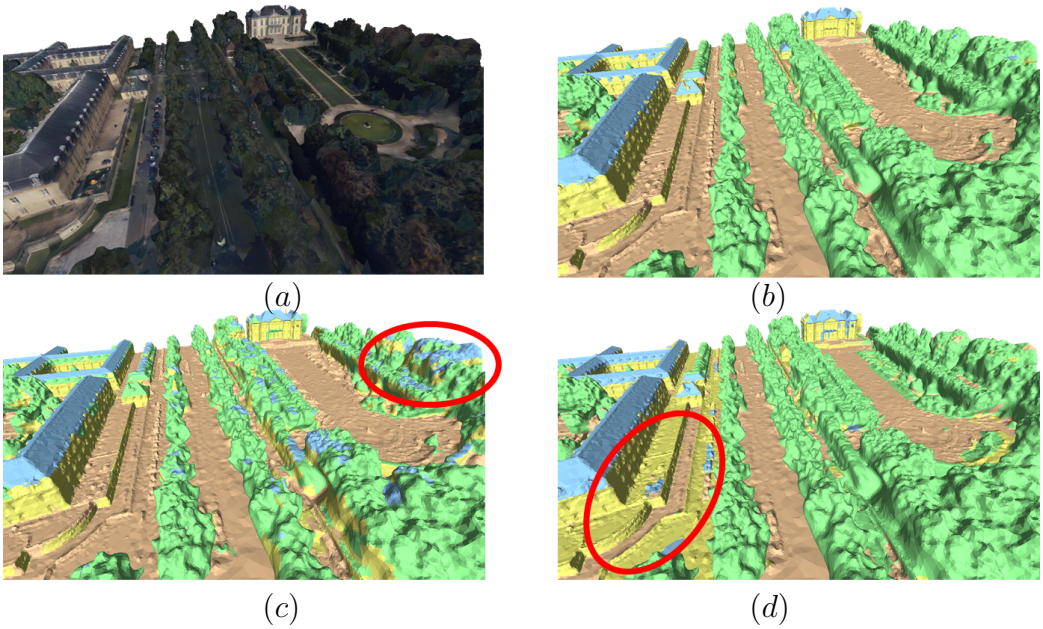

Figure 16: Joint use of geometric and photometric features. Relying upon geometry only often leads to confusions between vegetation and roof (c), whereas the use of photometry alone does not allow the ground and facades to be discriminated correctly (d). Combining geometry and photometry improves the classification by a significant margin, in particular by correcting the aforementioned confusion errors (b).

5.6. Limitations

The cases that challenge our algorithm are typically hilly areas for which the relief is particularly sharp and sloppy. Our algorithm often mislabels vertical components of quarries into facades. Also, the classification accuracy decreases with low quality input meshes, typically when occluded parts in the MVS images locally generate over-smooth surfaces that tend to be classified as trees.

6. Conclusions

We proposed a supervised approach for classifying urban landscapes provided in the form of textured meshes. Our approach relies upon two main technical contributions: a joint labeling system that leverages contextual knowledge in an intuitive and effective way, and a joint use of photometric and geometric attributes to locally describe the input meshes. We demonstrated these contributions combined with a supervised strategy provide an

Figure 17: Large scene classification. The results of Verdie et al. (2015) yield frequent confusions on vegetation with groups of trees mislabeled as facade and roof (second row). Our approach (first and fourth row) alleviates these errors. By restricting our model to a standard labeling system, ie with N classes of interest, no contextual information is exploited anymore (third row). Some mislabeled patches typically appear on top of trees as tree superfacets adjacent to roof or facade ones are not penalized anymore.

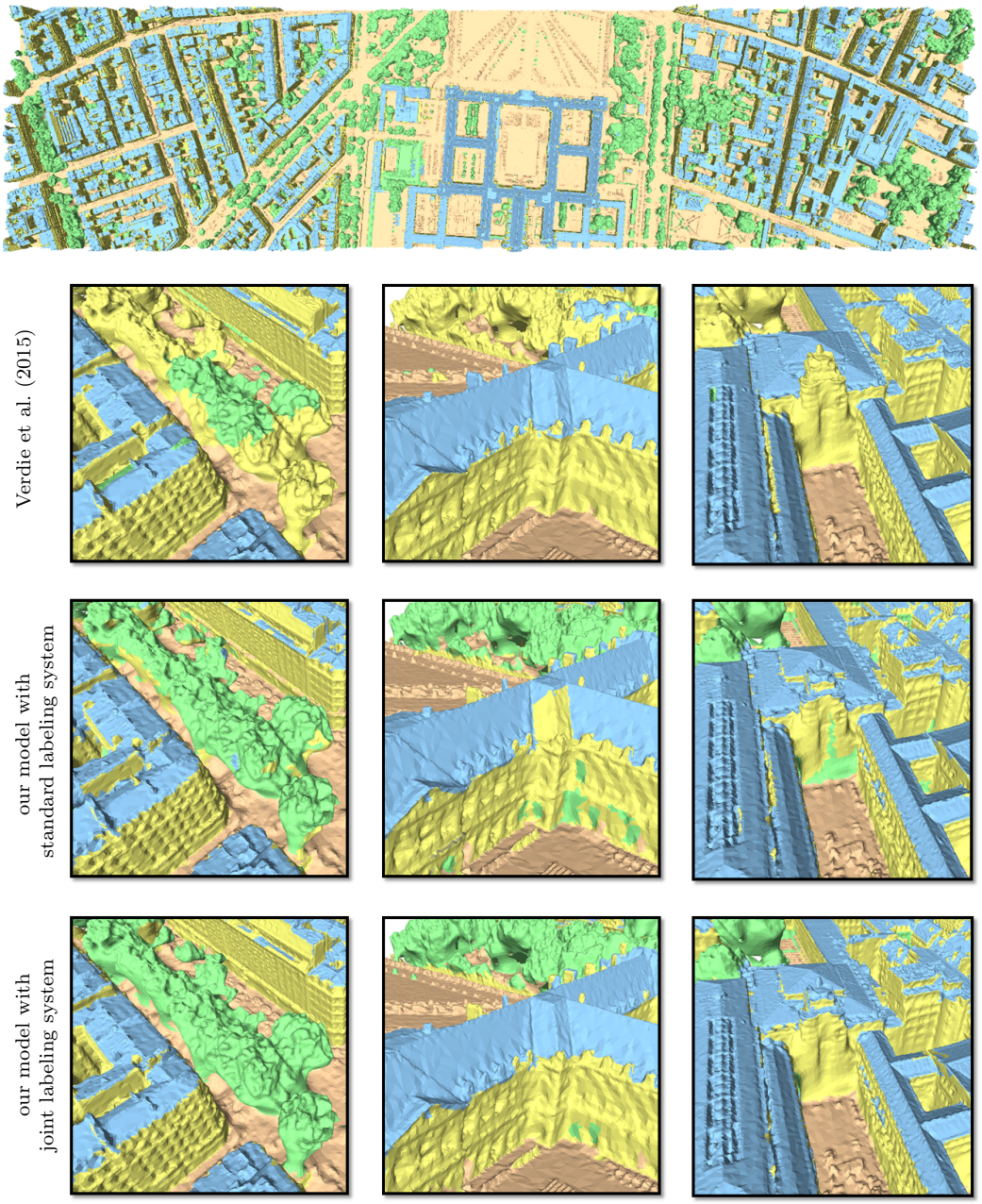

Figure 18: Large scene classification. For some complex buildings, none of the three methods achieve to correctly classify the roof superstructures as a few chimneys are typically labeled as facade. For such cases, our contextual information is not helpful as adjacent facade and roof superfacets are considered as plausible.

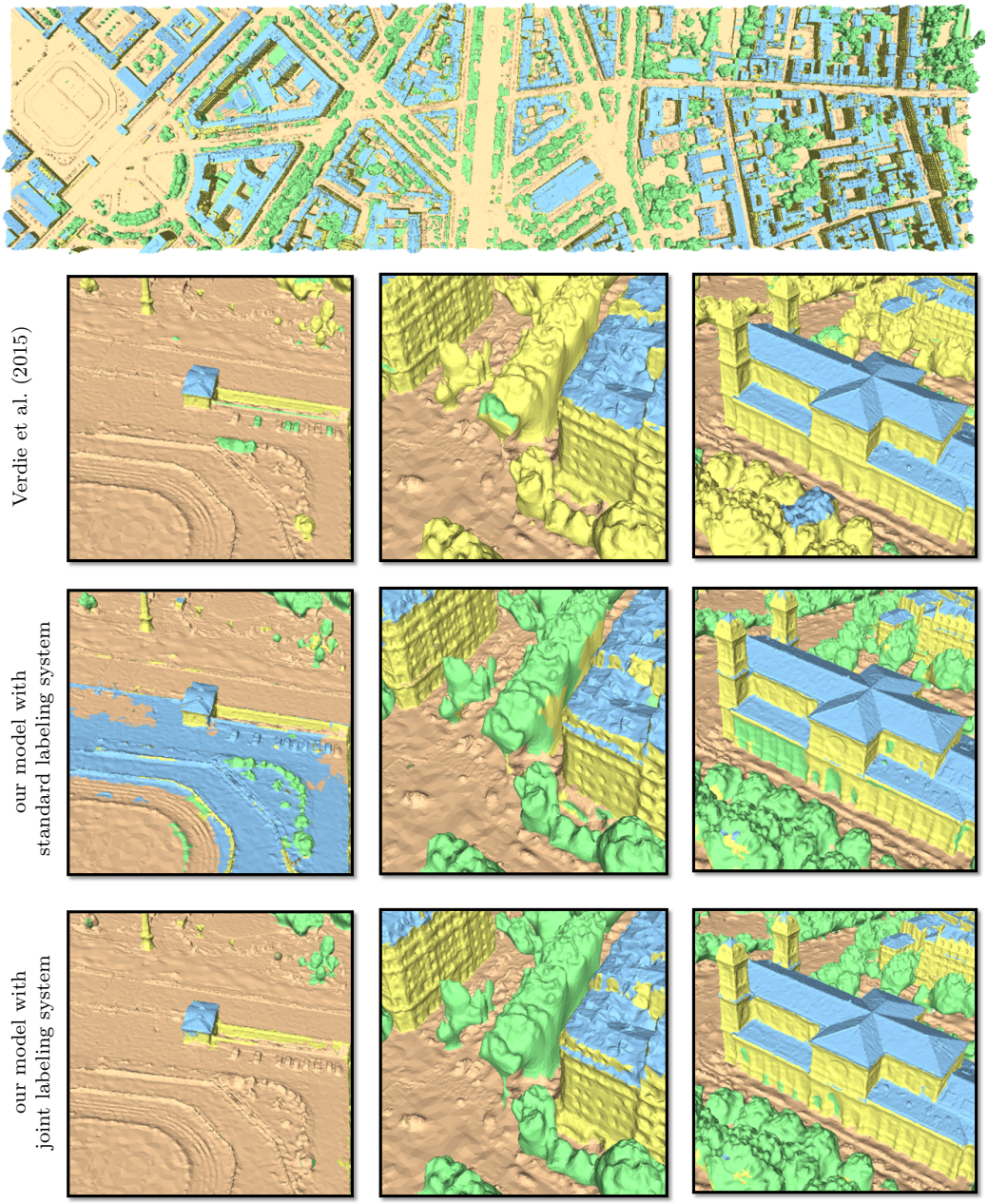

Figure 19: Large scene classification. Small common urban objects such as low fences and cars often create mislabeling errors using either (Verdie et al., 2015) or our model with a standard labeling system. The insertion of contextual knowledge with the joint labeling system proves to be more flexible by correctly classifying these elements (see left column).

efficient solution that outclasses the approach proposed by Verdie et al. (2015) in terms of correctness, ease-to-use, and robustness to data diversity. Globally speaking, such a two step strategy where mesh is first reconstructed before being partitioned into semantic classes remains more efficient that semantic reconstruction methods. Recovery 3D geometry and semantics simultaneously is methodologically more elegant but offers serious challenges in terms of algorithmic complexity.

In future work we would like to investigate on hierarchical classification of urban scenes where for instance the class roof would have some children classes as roof section or dormer-windows, and a parent class as building. we also plan to leverage the function of urban objects such as buildings to improve the accuracy of results. An open question is how to use such additional knowledge in a supervised classification approach.

Acknowledgments

We wish to thank Bentley/Acute3D and Interatlas for providing datasets.This work was supported by the European Research Council (ERC Starting Grant Robust Geometry Processing, Grant agreement 257474).

References

Blaha, M., Vogel, C., Richard, A., Wegner, J. znd Pock, T., Schindler, K., 2016. Large-scale semantic 3d reconstruction: an adaptive multi-resolution model for multi-class volumetric labeling, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) 23, 1222–1239.

Cabezas, R., Straub, J., Fisher, J., 2015. Semantically-aware aerial reconstruction from multi-modal data, in: IEEE International Conference on Computer Vision (ICCV).

Chauve, A.L., Pons, J.P., Audibert, J.Y., Keriven, R., 2009. Transductive segmentation of textured meshes, in: Asian Conference on Computer Vision (ACCV).

Chen, X., Golovinskiy, A., Funkhouser, T., 2009. A Benchmark for 3D Mesh Segmentation. ACM Transactions on Graphics Proc. of SIGGRAPH.

Cohen-Steiner, D., Alliez, P., Desbrun, M., 2004. Variational shape approximation. ACM Trans. on Graphics 23, 905–914.

Galleguillos, C., Rabinovich, A., Belongie, S.J., 2008. Object categorization using co-occurrence, location and appearance, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Haene, C., Zach, C., Cohen, A., Angst, R., Pollefeys, M., 2013. Joint 3D scene reconstruction and class segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

He, H., Upcroft, B., 2013. Nonparametric semantic segmentation for 3d street scenes, in: International Conference on Intelligent Robots and Systems.

Kähler, O., Reid, I.D., 2013. Efficient 3d scene labeling using fields of trees, in: IEEE International Conference on Computer Vision (ICCV).

Kalogerakis, E., Hertzmann, A., Singh, K., 2010. Learning 3d mesh segmentation and labeling. ACM Trans. on Graphics 29, 102:1–102:12.

Krähenbühl, P., Koltun, V., 2011. Efficient inference in fully connected crfs with gaussian edge potentials, in: Advances in Neural Information Processing Systems (NIPS).

Kundu, A., Li, Y., Dellaert, F., Li, F., Rehg, J.M., 2014. Joint semantic segmentation and 3d reconstruction from monocular video, in: European Conference on Computer Vision (ECCV).

Ladicky, L., Russell, C., Kohli, P., Torr, P.H.S., 2013. Inference methods for crfs with co-occurrence statistics. International Journal of Computer Vision (IJCV) 103, 213–225.

Lafarge, F., Keriven, R., Bredif, M., 2010. Insertion of 3D-Primitives in Mesh-Based Representations: Towards Compact Models Preserving the Details. IEEE Trans. on Image Processing 19, 1683–1694.

Lafarge, F., Keriven, R., Bredif, M., Vu, H.H., 2013. A hybrid multi-view stereo algorithm for modeling urban scenes. IEEE Trans. on Pattern Analysis and Machine Intelligence 35.

Lai, K., Bo, L., Fox, D., 2014. Unsupervised feature learning for 3D scene labeling, in: IEEE International Conference on Robotics and Automation (ICRA).

Liu, K., Zhang, J., Yang, P., Huang, K., 2015. GRSA: generalized range swap algorithm for the efficient optimization of mrfs, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Lv, J., Chen, X., Huang, J., Bao, H., 2012. Semi-supervised mesh segmentation and labeling. Computer Graphics Forum 31, 2241–2248.

Martinovic, A., Knopp, J., Riemenschneider, H., Van Gool, L., 2015. 3d all the way: Semantic segmentation of urban scenes from start to end in 3d, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

"Montoya-Zegarra", J., Wegner, J.D., Ladicky, L., Schindler, K., 2015. Semantic segmentation of aerial images in urban areas with class-specific higher-order cliques, in: Photogrammetric Image Analysis (PIA).

Mottaghi, R., Chen, X., Liu, X., Cho, N., Lee, S., Fidler, S., Urtasun, R., Yuille, A.L., 2014. The role of context for object detection and semantic segmentation in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Musialski, P., Wonka, P., Aliaga, D., Wimmer, M., Van Gool, L., Purgathofer, W., 2013. A survey of urban reconstruction. Computer Graphics Forum 32.

Myeong, H., Chang, J.Y., Lee, K.M., 2012. Learning object relationships via graph-based context model, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Niemeyer, J., Rottensteiner, F., Soergel, U., 2014. Contextual classification of lidar data and building object detection in urban areas. ISPRS Journal of Photogrammetry and Remote Sensing 87, 152–165.

Page, D.L., Koschan, A.F., Abidi, M.A., 2003. Perception-based 3d triangle mesh segmentation using fast marching watersheds, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., Breitkopf, U., 2012. The ISPRS benchmark on urban object classification and 3d building reconstruction, in: Proc. of the ISPRS congress.

Savinov, N., Ladicky, L., Hane, C., Pollefeys, M., 2015. Discrete optimization of ray potentials for semantic 3d reconstruction, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Schindler, K., 2012. An overview and comparison of smooth labeling methods for land-cover classification. IEEE Trans. on Geoscience and Remote Sensing 50, 4534–4545.

Seitz, S., Curless, B., Diebel, J., Scharstein, D., Szeliski, R., 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Sengupta, S., Greveson, E., Shahrokni, A., Torr, P.H.S., 2013. Urban 3d semantic modelling using stereo vision, in: IEEE International Conference on Robotics and Automation (ICRA), pp. 580–585.

Shamir, A., 2008. A survey on mesh segmentation techniques. Computer Graphics Forum 27.

Shlafman, S., Tal, A., Katz, S., 2002. Metamorphosis of polyhedral surfaces using decomposition. Computer Graphics Forum 21, 219–228.

Shotton, J., Johnson, M., Cipolla, R., 2008. Semantic texton forests for image categorization and segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Theologou, P., Pratikakis, I., Theoharis, T., 2015. A comprehensive overview of methodologies and performance evaluation frameworks in 3d mesh segmentation. Computer Vision and Image Understanding (CVIU) 135, 49–82.

Valentin, J.P.C., Sengupta, S., Warrell, J., Shahrokni, A., Torr, P.H.S., 2013. Mesh based semantic modelling for indoor and outdoor scenes, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Van Kaick, O., Tagliasacchi, A., Sidi, O., Zhang, H., Cohen-Or, D., Wolf, L., Hamarneh, G., 2011. Prior knowledge for part correspondence. Computer Graphics Forum 30, 553–562.

Verdie, Y., Lafarge, F., Alliez, P., 2015. LOD generation for urban scenes. ACM Trans. on Graphics 34, 30.

Vineet, V., Miksik, O., Lidegaard, M., Niesner, M., Golodetz, S., Prisacariu, V.A., Kahler, O., Murray, D.W., Izadi, S., Perez, P., Torr, P.H.S., 2015. Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction, in: IEEE International Conference on Robotics and Automation (ICRA).

Volpi, M., Ferrari, V., 2015. Semantic segmentation of urban scenes by learning local class interactions, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Vu, H., Keriven, R., Labatut, P., Pons, J., 2009. Towards high-resolution large-scale multiview, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Xiao, J., Quan, L., 2009. Multiple view semantic segmentation for street view images, in: IEEE International Conference on Computer Vision (ICCV).

Yao, J., Fidler, S., Urtasun, R., 2012. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Zhang, C., Wang, L., Yang, R., 2010a. Semantic segmentation of urban scenes using dense depth maps, in: European Conference on Computer Vision (ECCV), pp. 708–721.

Zhang, H., van Kaick, O., Dyer, R., 2010b. Spectral mesh processing. Computer Graphics Forum 29, 1865–1894.