ISPRS Journal of Photogrammetry and Remote Sensing 129 (2017) 86–102

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Robust point cloud classification based on multi-level semantic relationships for urban scenes

Qing Zhu [a,b], Yuan Li [c], Han Hu [a,c,*], Bo Wu [c]

[a] Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Gaoxin West District, Chengdu, China
[b] Collaborative Innovation Center for Geospatial Technology, 129 Luoyu Road, Wuhan, China
[c] Department of Land Surveying and Geo-Informatics, The Polytechnic University of Hong Kong, Hum Hom, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 23 November 2016
Received in revised form 27 April 2017
Accepted 27 April 2017

Keywords:
Point clouds
3D semantics
Classification
Markov random field

ABSTRACT

The semantic classification of point clouds is a fundamental part of three-dimensional urban reconstruction. For datasets with high spatial resolution but significantly more noises, a general trend is to exploit more contexture information to surmount the decrease of discrimination of features for classification. However, previous works on adoption of contexture information are either too restrictive or only in a small region and in this paper, we propose a point cloud classification method based on multi-level semantic relationships, including point–homogeneity, supervoxel–adjacency and class–knowledge constraints, which is more versatile and incrementally propagate the classification cues from individual points to the object level and formulate them as a graphical model. The point–homogeneity constraint clusters points with similar geometric and radiometric properties into regular-shaped supervoxels that correspond to the vertices in the graphical model. The supervoxel–adjacency constraint contributes to the pairwise interactions by providing explicit adjacent relationships between supervoxels. The class–knowledge constraint operates at the object level based on semantic rules, guaranteeing the classification correctness of supervoxel clusters at that level. International Society of Photogrammetry and Remote Sensing (ISPRS) benchmark tests have shown that the proposed method achieves state-of-the-art performance with an average per-area completeness and correctness of 93.88% and 95.78%, respectively. The evaluation of classification of photogrammetric point clouds and DSM generated from aerial imagery confirms the method's reliability in several challenging urban scenes.

1. Introduction

Automatic three-dimensional (3D) city modeling has generated significant attention in the urban planning, analysis and design community. Despite the procedural approach (Dang et al., 2015; Esri, 2016; Vanegas et al., 2010), which uses predefined rules/-grammar and two-dimensional (2D) footprints to generate detailed 3D models, considerable efforts have been devoted to automatic reconstruction from point clouds (Lafarge and Mallet, 2012; Poullis, 2013; Xiong et al., 2015; Zhou and Neumann, 2010). Airborne laser scanning (ALS) is an important source of massive point clouds. Another important source is dense image matching (DIM), especially DIM using oblique images through multi-view stereo (MVS) pipelines (Furukawa and Ponce, 2010; McClune et al., 2016; Vu et al., 2012), which is quite popular in the field of photogrammetry currently. However, except for the generation of textured triangulated meshes, the automatic generation of 3D polygonal models remains an open problem that is being actively researched (Musialski et al., 2013). Recent advances in automatic urban reconstruction have revealed that enriching the raw point clouds or meshes with semantic segments and then reconstructing each segment, is an effective and scalable paradigm for large-scale reconstruction (Lafarge and Mallet, 2012; Poullis, 2013; Verdie et al., 2015).

However, the semantic segmentation or classification of point clouds, the focus of this paper, is considered non-trivial work in complex urban scenes (Bláha et al., 2016). The two cornerstones of classification are discriminative features and proper classifiers, both of which are generally obtained locally (e.g., a point and its neighborhood) (Chehata et al., 2009; Hackel et al., 2016), and only pairwise interactions (Niemeyer et al., 2014; Weinmann et al., 2015) are considered. However, due to the obvious defects of point

* Corresponding author at: Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Gaoxin West District, Chengdu, China.
E-mail address: huhan8807@gmail.com (H. Hu).

http://dx.doi.org/10.1016/j.isprsjprs.2017.04.022
0924-2716/

cloud data (e.g., noise, loss of sharp features and outliers), such methods are not resilient and thus require extensive drudgery in the form of manual quality control, especially for photogrammetric point clouds (Hu et al., 2016; Nex and Gerke, 2014). Compared to Light Detection and Ranging (LiDAR) point clouds, the photogrammetric are generally more noise-laden, which will dramatically decrease the discriminations of features derived from small local regions and consequently lead to the failure of classification. To achieve robust classification of point clouds, larger context range must be incorporated into the workflow to surmount the noise. In fact, the exploitation of larger context is a trend of scene classification. For example, by plane segmentation (Xu et al., 2014; Zhang and Lin, 2012) or second or higher order Markov random field (MRF) or conditional random fields (CRF) (Niemeyer et al., 2014, 2016; Lafarge and Mallet, 2012; Sengupta and Sturgess, 2015).

However, the adoption of contexture information is either too restrictive that requires perfect segmentation of planes (Zhang and Lin, 2012) or only involves a small local region through point-level interactions (Niemeyer et al., 2014; Lafarge and Mallet, 2012). Therefore, we propose a point cloud classification method that propagates the classification cues from a single point to the object level using flexible multi-level semantic relationships based on an intermediate representation of point clouds — the "supervoxel". The "supervoxel" in this paper is an extension of "superpixel" (Ren and Malik, 2003; Achanta et al., 2012) from 2D to 3D, but unlike the in the enumerative space of a 2D image, the "supervoxel" refers to a fixed-size cluster of unorganized points generated through space partitioning, and the points in each cluster maintain the original geometries individually but together constitute a regular shape. The proposed method involves three constraints derived from different entity levels. (1) the point–homogeneity constraint represents the semantic relationships between points and clusters homogenous points into over-segmented supervoxels designed to not cross object boundaries. (2) the supervoxel–adjacency constraint encodes pairwise interactions between supervoxels. (3) the class–knowledge constraint represents the global relationships of supervoxels at the object scale. These three relationships can be modeled using the unary, pairwise and high-order cliques (Li, 2009) in a MRF, and a two-step inference strategy is adopted to solve the labels of each supervoxel.

The remainder of this paper is organized as follows. Section 2 provides a brief literature review of the existing point classification methods. In Section 3, the classification method using multi-level semantic relationships is demonstrated in detail. The performance of the proposed methods is then evaluated and analyzed in Section 4, using both the ISPRS benchmark dataset (Rottensteiner et al., 2012) and photogrammetric point clouds derived from a penta-view multiple camera system (Petrie, 2009). The concluding remarks and future works are presented in Section 5.

2. Related works

According to the type of entity used for classification, existing methods can be categorized as point- or object/segment-based (Gerke and Xiao, 2014; Zhou et al., 2012). Below, we briefly review previous methods and demonstrate the rationale for the proposed method.

2.1. Point-based methods

Point-based methods generally extract point-wise features locally from the neighborhood defined by a sphere or cylinder, and then supervised or unsupervised classifiers are used. Therefore, such methods usually focus on the selection of discriminative features and effective classifiers. For instance, (Lodha et al., 2007) merged airborne LiDAR with images to extract more discriminative features, including geometric features from LiDAR and radiometric features from images. Then, based on these features, the points were divided into four classes with AdaBoost (Chehata et al., 2009). In addition to the geometric features from points and radiometric features from images, more sophisticated features are also used. For example, the full waveform LiDAR provides useful information for feature extractions (Mallet et al., 2008), spectral information within the feature selection framework shows promising results (Guo et al., 2011; Mallet et al., 2011) and hierarchical features exhibit superior performance in large-scale urban environments (Hackel et al., 2016). With regard to classifiers, despite the boosting method mentioned above, other popular methods such as Random Forests (RFs) (Breiman, 2001; Gislason et al., 2006) and support vector machines (SVMs) (Mountrakis et al., 2011) are also used for point cloud classification.

In the abovementioned methods, the points are labeled individually in the feature space without considering relationships, which often leads to discontinuities in the classification results. To avoid this, other point-based methods take advantage of contextual information. This type of semantic relationship at the point level leads to the use of graphical models, such as MRF or CRF (Kumar and Hebert, 2006). For instance, (Lafarge and Mallet, 2012) proposed an unsupervised method with an MRF framework, where the Potts model (Li, 2009) is introduced to define the pairwise interactions between neighboring points, with discriminative features from each point used to compute a potential classification result. Then, a graph cut-based algorithm (Boykov et al., 2001) was used to quickly reach an approximate solution close to the global optimum of energy. (Niemeyer et al., 2014) integrated an RF classifier into a CRF framework, where the unary and pairwise potentials of CRF were based on probabilities computed by RF.

Although these methods benefit from using contextual information, their effects have been very limited because they only consider the coherences between points within a small neighborhood. This limitation renders point-based methods less resilient to issues of data quality—such as noises and density anisotropy—such that these methods can generally only be applied to accurate point clouds (e.g., LiDAR), and require other ancillary datasets to extract discriminative features. As such, a great deal of manual parameter tuning and interactive post-processing work is required to refine the results. Furthermore, in cases of large-scale urban scenes, a typical site contains millions or more points, such that even a state-of-the-art inference method is challenged by graphical models with only pairwise interactions.

2.2. Object-based methods

Object-based methods choose a point cluster in which points share homogeneous properties as the entity to be classified. The features are generally extracted from the points first, and then split into different clusters, from which more discriminative features are extracted. The clusters are then classified based on these object-based features using a proper classifier, such as a probability distribution function (PDF) (Poullis, 2013), SVM (Zhang and Lin, 2012), RM or MRF (Gerke and Xiao, 2014). When compared with the point-based method, the major difference lies in the step for generating the clusters.

One strategy for generating clusters is to make each cluster contain as many homogeneous points (points that have similar colors, normals, curvatures, etc.) as possible through a segmentation process, so that each segment corresponds to a certain component of the objects, such as a façade or a roof. Based on geometric features, (Xu et al., 2014) segmented the point cloud into planar and irregular segments using surface growing (an instantiation of region

growing) and mean-shift. Then, based on the features extracted from each surface, the segments were divided into four classes. (Gerke and Xiao, 2014) considered spectral features during region growing by fusing LiDAR with images, so that objects that were very close to each other could be separated. (Poullis, 2013) clustered the point cloud based on a hierarchical analysis comprising two stages: segmenting the point cloud into non-overlapping patches, and merging the patches into surfaces according to their PDFs. A similar hierarchical method was used in (Zhang et al., 2016), except that the features were extracted with latent Dirichlet allocation and sparse coding.

Another strategy is to partition the point cloud into regular-shaped clusters known as voxels (Aijazi et al., 2013) or supervoxels (Lim and Suter, 2009; Papon et al., 2013). Early works on voxelization were based on the distances between points. For instance, (Aijazi et al., 2013) defined a supervoxel's size using radius research. Zhou et al. (2012) and Babahajiani et al. (2015) clustered points within a given radius into one supervoxel and limited the number of points to avoid bestriding object boundaries. In addition, (Lim and Suter, 2009) used normalized RGB values as color constraints during voxelization. The supervoxels were then merged into larger components labeled as different classes based on their features (Aijazi et al., 2013; Babahajiani et al., 2015; Zhou et al., 2012), or directly used for classification in a CRF framework (Kim et al., 2013; Lim and Suter, 2009; Wolf et al., 2015).

A recent trend in the classification of point clouds is to exploit contextual information. For example, Xu et al. (2014) proposed a multiple-entity based classification method, which segmented the LiDAR point cloud into semantically connected parts and labeled each part independently. However, this imposed restrictions to the segmentation, which is hard for photogrammetric point cloud in the first place. Contexture information on the point level is also adopted, including pairwise interaction through CRF (Niemeyer et al., 2014), relatively longer interactions through higher order Potts model (Niemeyer et al., 2016) or from multiscale representation of point clouds (Hackel et al., 2016). However, these methods extracted features at point level, which only shown application for LiDAR point cloud with relatively better quality. Concurrent with our work, Rouhani et al. (2017) proposed a semantic segmentation methods for triangulated meshes from MVS pipeline, which involved a similar primitive termed as superfacets, the sizes of the superfacets were not uniform, which may induce some difficulties on feature extraction and small superfacets are significantly biased and prone to misclassification. To better exploit contexture information for noisy point cloud, in this paper, we propose a point classification methods based on multi-level semantic relationships, which uses uniform sized supervoxels and the interactions between primitives are modeled through the MRF.

3. Point cloud classification based on multi-level semantic relationships

3.1. Overview of the approach

The framework starts with an adaptive surface filter (ASF) algorithm proposed by (Hu et al., 2014) to separate the original point cloud into ground and non-ground points. This paper mainly focuses on the classification of non-ground point cloud using the multi-level semantic constraints, which are integrated into a higher order MRF framework, as follows,

$$\min_{\mathbf{y}} E(\mathbf{y}) = \sum_{i\in\mathbf{v}} E(y_i) + \sum_{(i,j)\in\mathbf{e}} E(y_i, y_j) + \sum_{i\in\mathbf{c}} E(y_i) \quad (1)$$

where $\mathbf{y}$ is a labeling configuration and for each $y \in \mathbf{y}$ the value space is Ω = *{roof, façade, vegetation, clutter}* and $\mathbf{v}$, $\mathbf{e}$ and $\mathbf{c}$ represent the set of vertices, edges and connected components in a graphical model, respectively. The first term on the right side is the unary energy, which is measured by the features extracted from supervoxels and constrained by the point–homogeneity. The second term is the pairwise energy, which represents the supervoxel–adjacency constraint. The third term is the high-order cliques of supervoxel clusters, in which the class–knowledge constraint is enforced. The purpose is to find an optimal labeling configuration $\mathbf{y}$ for each supervoxel, which minimizes the global energy $E(\mathbf{y})$. However, as it is difficult to minimize the energy of general purpose higher order MRF, to solve the energy minimization problem with higher order cliques, we resort to a two-step strategy that minimizes pairwise energy through graph cut and post-processing refinement which reduces the energy in higher order cliques.

Fig. 1 illustrates the framework of the proposed method. The column on the left shows the multi-level entities in the point cloud. The middle column shows the corresponding entities in the graphical model, and between them are the links between these two types of entities. The right column represents the multi-level semantic constraints with the arrows pointing to the entities that they enforce. The red solid and red dotted lines represent the constraints that take major and minor effects, respectively.

3.2. Graphical model generation of the Markov random field

Rather than a single point, we use the supervoxel as the fundamental entity to build the MRF model for three reasons. First, the supervoxel provides more discriminative features than a single point due to the larger meaningful contexts involved in feature extraction. Second, for discrete point representation, the adjacency relationships are vague and more densely connected; however, the adjacency relationships between supervoxels are much clearer, and this merit is vital for the construction of a graphical model and energy optimization. Third, using supervoxels as the basic entity significantly reduces computational complexity, which is critical for large-scale urban scenes. Based on the supervoxels, a supervoxel–adjacency graph is created, where each vertex corresponds to a supervoxel, each edge corresponds to the pairwise connection between two adjacent supervoxels and each connected component corresponds to a set of supervoxels that represent one complete object. Fig. 2 illustrates our graphical model, and the elements of the three different order-level cliques for Eq. (1).

3.2.1. Supervoxel generation for first-order cliques

Based on the assumption that points with high homogeneity are more likely to correspond to the same object, the non-ground points are firstly clustered into supervoxels with the point–homogeneity constraint. The point homogeneity is measured in three spaces—coordinate space, color space and normal space—and formulated as,

$$h = h_{Coord} + h_{Color} + h_{Normal} \quad (2)$$

where h_{Coord} is the Euclidean distance calculated by spatial coordinates x, y and z; h_{Color} is also a Euclidean distance calculated by the color components R, G and B; and h_{Normal} is the angle between two normal vectors. The normal vectors are estimated using the method proposed by Boulch and Marlet (2012), with the neighborhood defined by K (K is set 10–15 in this paper) nearest points. To balance the effects of these three parts, they are normalized to *0–1* as in Eq. (3), where r is the size of the supervoxel (as illustrated in Fig. 3(a), and in this paper r is set to 3–5 m depending on the density of the point cloud and the size of the smallest object in the scene), m is the maximum color gradient and $\theta(n_1 \bullet n_2)$ represents the angle between two normal vectors n_1 and n_2. The three factors h_{Coord}, h_{Color} and h_{Normal} together determine the homogeneity h

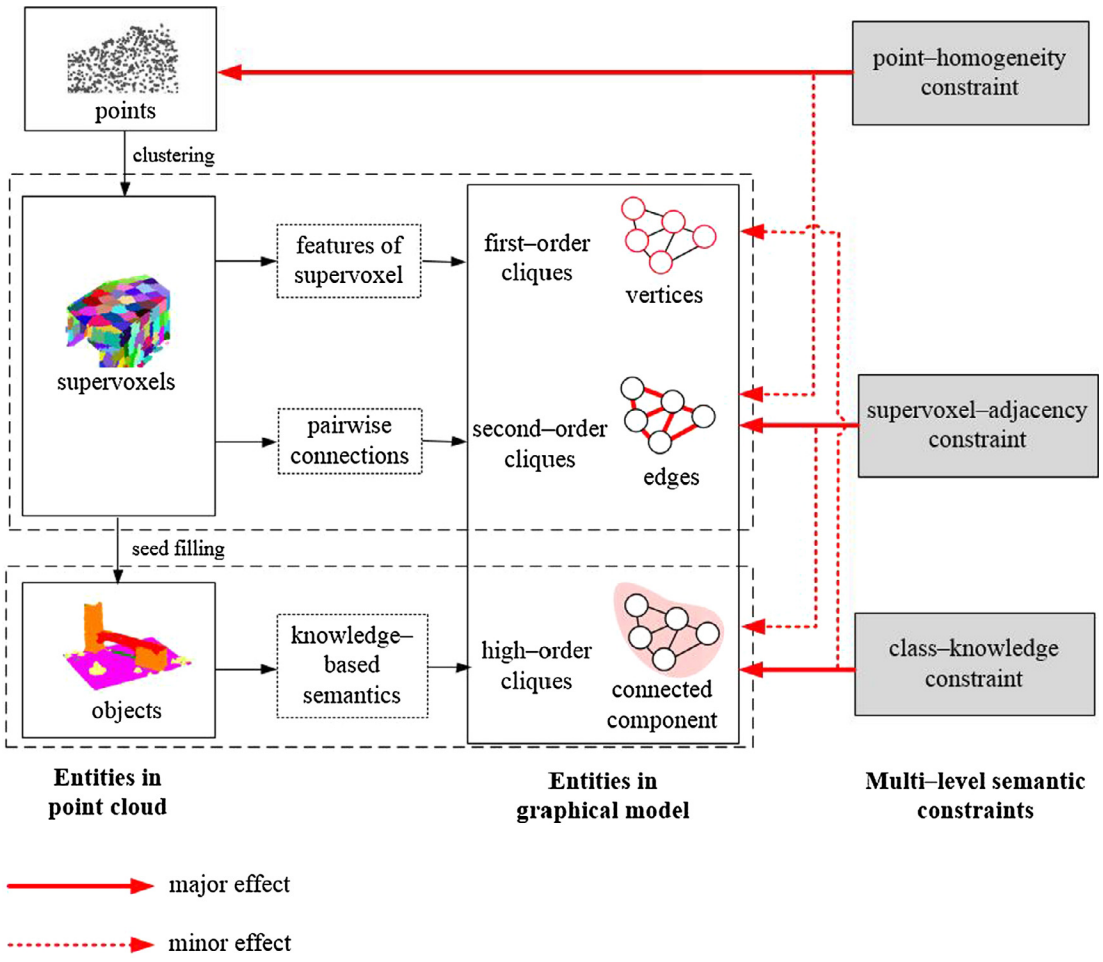


Fig. 1. MRF framework of non-ground point cloud classification.

between points to avoid striding boundaries, and the less h between two points is, the more homogeneous they are.

$$\begin{cases} h_{Coord} = \frac{\sqrt{\Delta x^2+\Delta y^2+\Delta z^2}}{r} \\ h_{Color} = \frac{1}{m}\sqrt{\frac{\Delta R^2+\Delta G^2+\Delta B^2}{3}} \\ h_{Normal} = \frac{\theta(n_1 \bullet n_2)}{\pi} \end{cases} \quad (3)$$

The non-ground points are clustered into supervoxels by an octree-based region growing strategy (Papon et al., 2013; Vo et al., 2015), during which the point–homogeneity constraint judges whether a new point will be contained or not in a supervoxel. The supervoxels on different objects exhibit different characteristics (as shown in Fig. 3(b)), which can help to extract more discriminative features and compute the first-order cliques, as described in Section 3.3.1.

3.2.2. Pairwise connection of second-order cliques

In most of the previous studies that used graphical models for classification, such as (Niemeyer et al., 2014) and (Zhou et al., 2012), the edges were determined based on the distance between two vertices. However, this type of adjacency cannot describe the relationships accurately, especially when point density varies. In this paper, as the supervoxels are generated based on octree, explicit pairwise connections can be easily obtained from the octree structure. We define 26 adjacent leaf-nodes in the octree (see Fig. 3(a)) that are maintained between the supervoxels grown from corresponding valid leaf-nodes, so that the pairwise connection in the graph is explicit according to the adjacency relationships between supervoxels. The pairwise connection is a key element of second-order cliques in Eq. (1), which makes the labeling results consistent.

3.2.3. Connected components for high-order cliques

A single supervoxel is only a small part of an object, and cannot represent the relationships between different classes. In reality, there are vital semantic relationships between objects that belong to different classes, such as roofs must be higher than facades, and facades must be connected by the roof and the ground. To benefit from such semantic knowledge, we define a subgraph without any extra connected vertex as a connected component, and the vertices in a connected component sharing the same label correspond to an independent object, as shown in Fig. 2(d), where different independent objects are marked with different background colors. According to the pairwise connections described above, the connected components—which can be easily obtained using a two-pass or seed-filling algorithm—are used to compute the high-order cliques. A connected component can contain more than one independent objects, and the independent objects contained in one connected component together determine the energy of the corresponding higher order clique, based on their semantic features and relationships.

3.3. MRF energies

3.3.1. Supervoxel-based features for unary terms

Before introducing the method to compute the unary terms in MRF, we first demonstrate the features extracted from supervoxels.

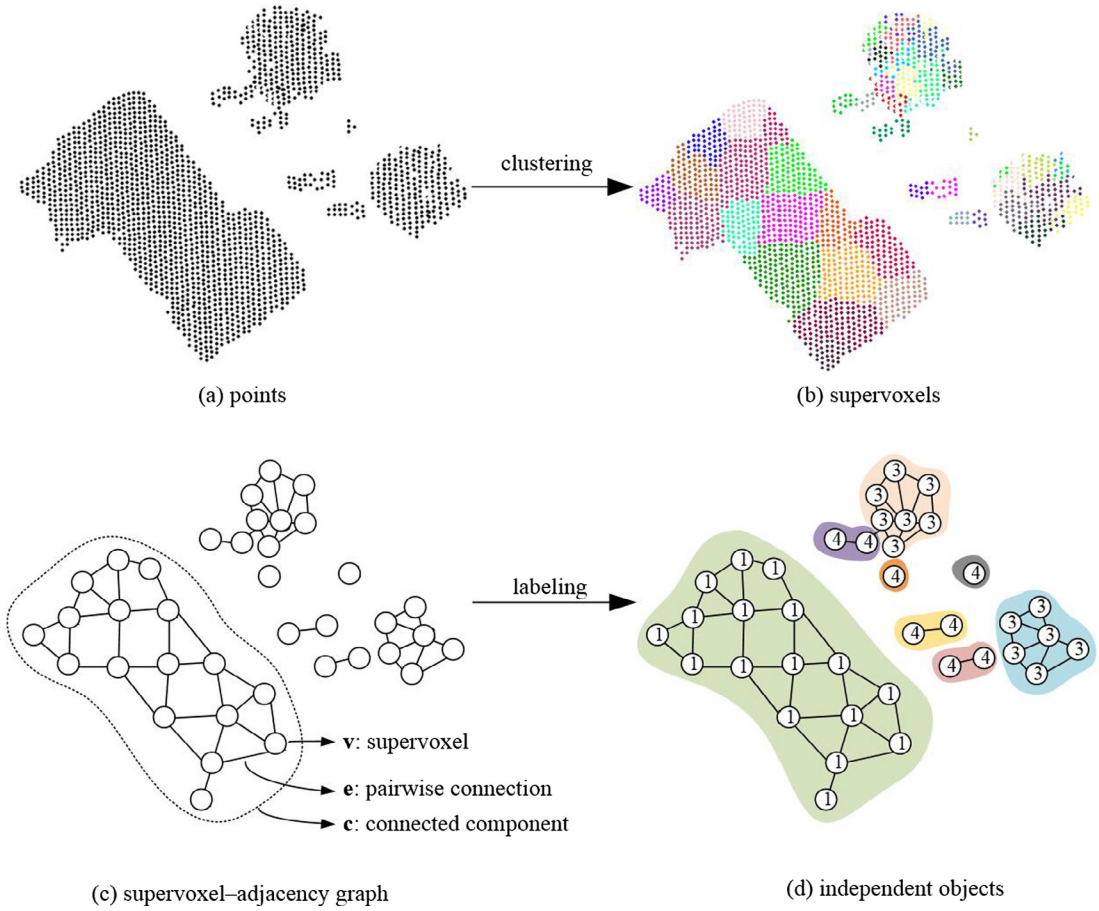


Fig. 2. Graphical model for MRF: (a) non-ground point set, (b) color-coded supervoxels (c) supervoxel–adjacency graph and the three level elements in the graph. (d) independent objects marked with different background colors. The numbers 1–4 simulate a labeling configuration for the supervoxels. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

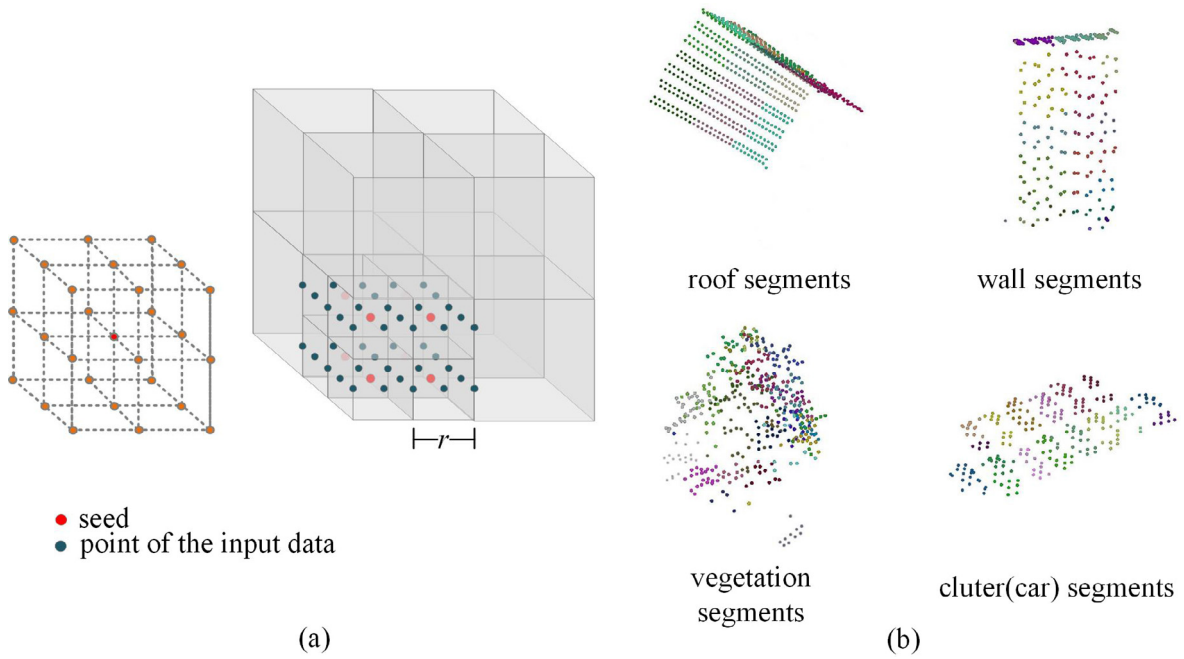


Fig. 3. Supervoxel generation: (a) schematic diagram of the octree structure and (b) clustering results of four different kinds of object.

As Fig. 3(b) shows, supervoxels corresponding to different objects are quite distinctive, as the roof and façade segments are rectangular in shape and well-distributed, but they differ in direction because the clutter segments are rarely neighbored and the vegetation segments are extremely irregular in both shape and distribution. In this paper, six features are extracted from each supervoxel:

(1) F_e: elevation above ground measured by the height difference between the barycenter of the supervoxel and ground elevation interpolated from the filtered ground points (Hu et al., 2014). This feature is used to distinguish the roofs and clutters, as even the lowest roofs are usually higher than the clutters.

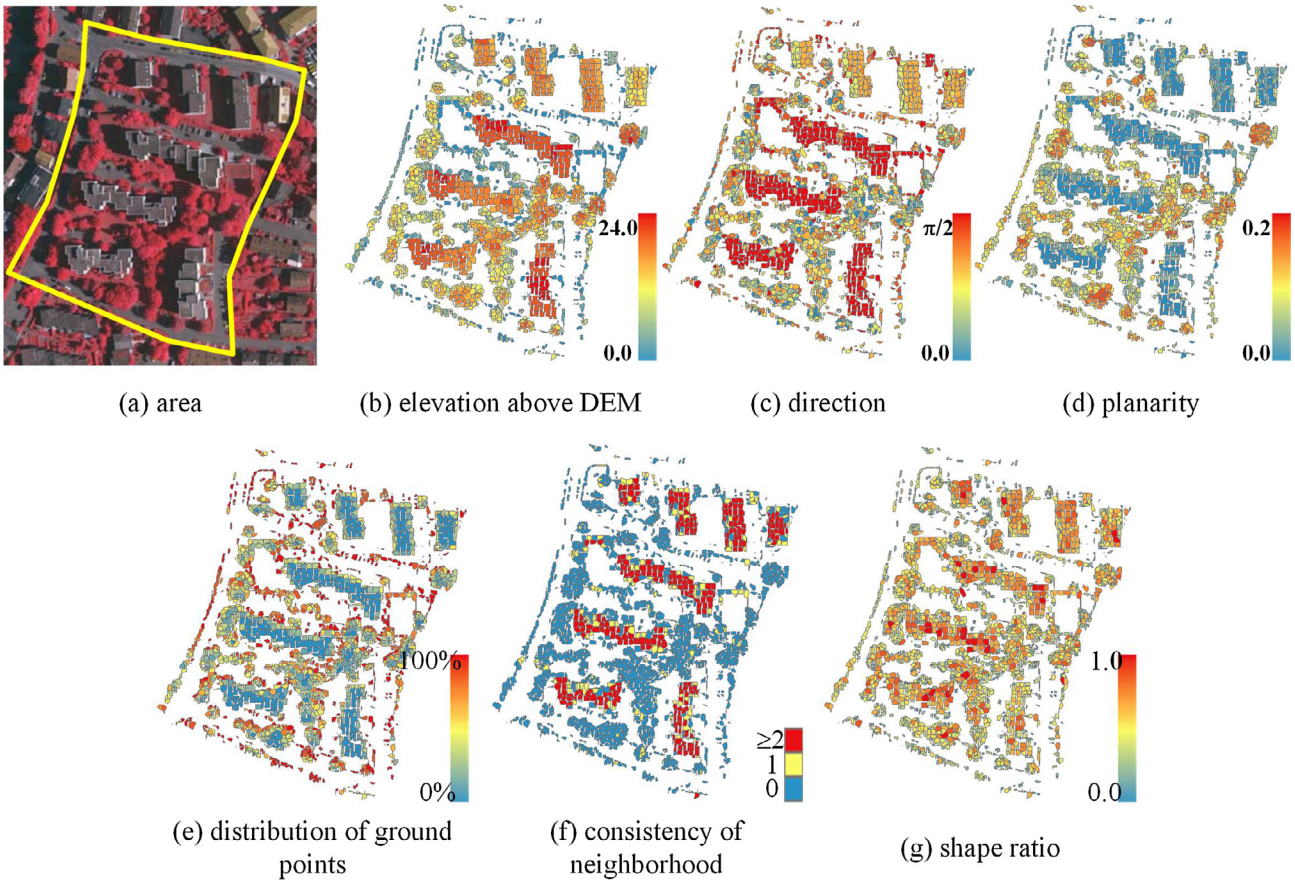


Fig. 4. Each feature's capacity to distinguish: (a) the area shown on ortho-image, (b)–(g) the value distribution of each feature in the area with color-coded feature values, where blue represents a low and red a high value. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

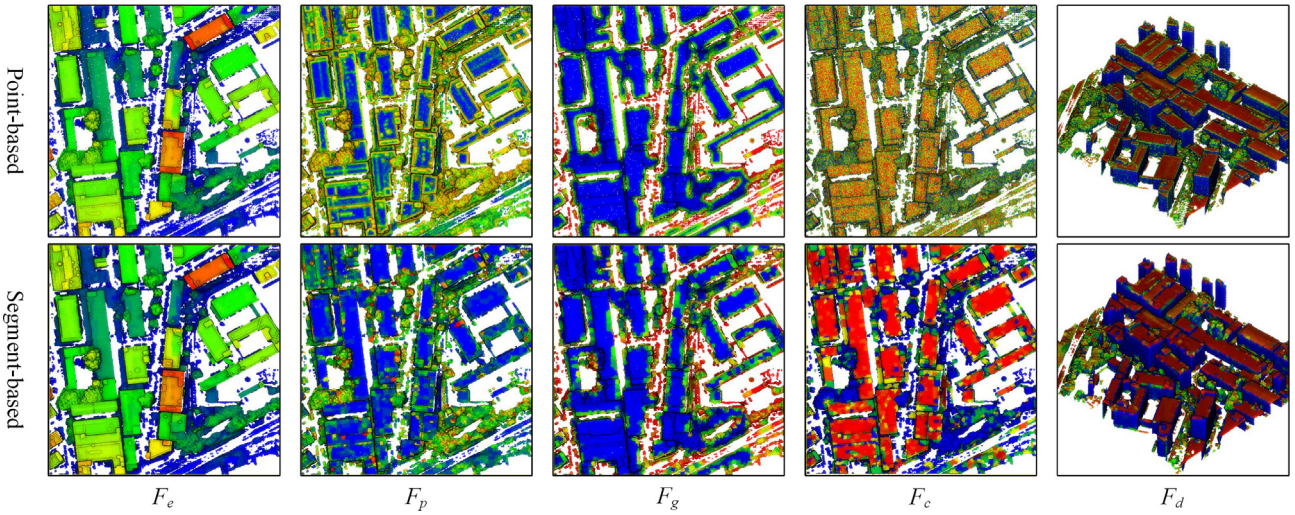


Fig. 5. The differences between features extracted from points and those extracted from segments.

Table 1
Discrimination of the features for each class. The symbols "+", "−" and "/" indicate the feature value tends to be large, small, and arbitrary in the corresponding class, respectively.

	F_e	F_d	F_p	F_g	F_c	F_s
Roofs	+	+	−	−	+	+
Facades	/	−	−	/	+	−
Vegetation	/	/	+	/	−	/
Clutters	−	/	−	+	−	−

(2) F_d: direction of the supervoxel measured by the angles between the normal vectors of the supervoxel and the horizontal plane. This feature takes a small value for supervoxels corresponding to facades.
(3) F_p: planarity measured by the variance in distances between the points and the plane by least squares fitting. This feature can distinguish flat surfaces, such as roofs and facades, from other rough objects, such as vegetation and clutters.

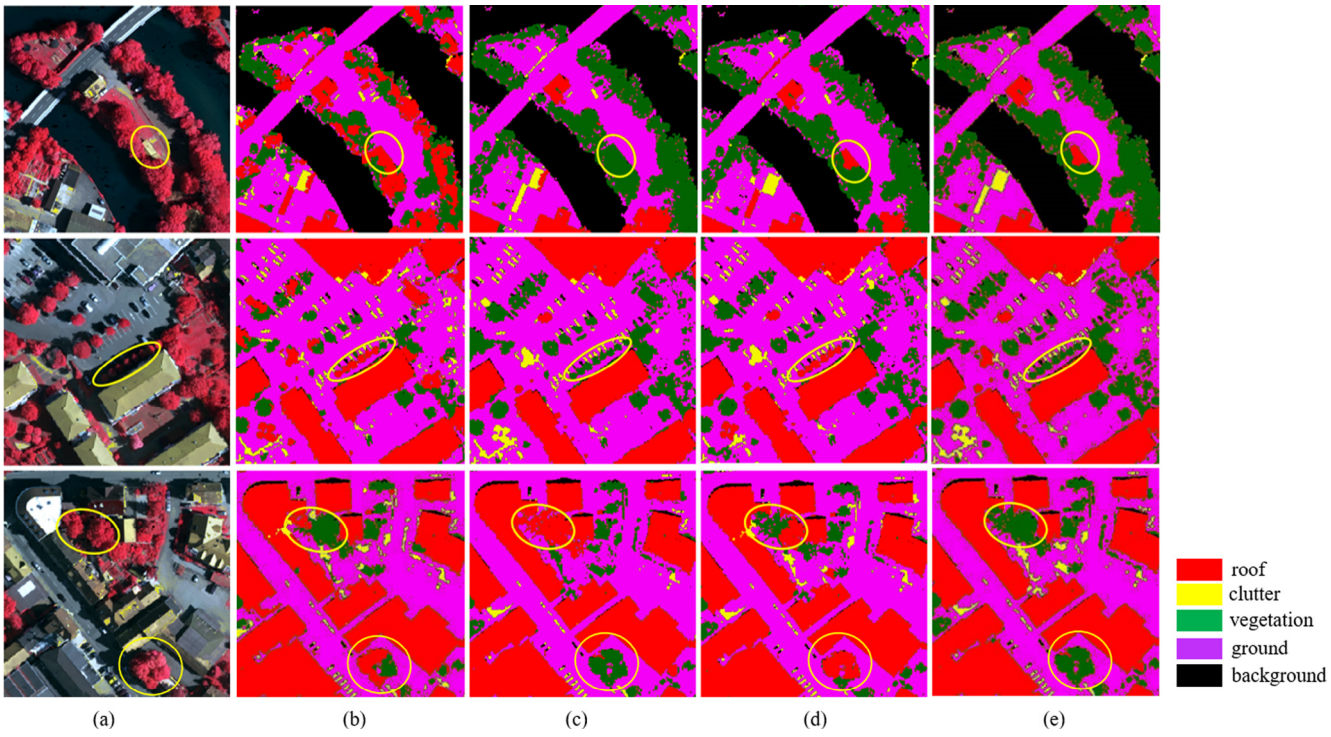


Fig. 6. Comparison of different definitions of κ. From left to right: (a) image content, (b) initial labeling result, (c) $\kappa = 1$, (d) $\kappa = normalize\ (distance)$, (e) $\kappa = h(i, j)$.

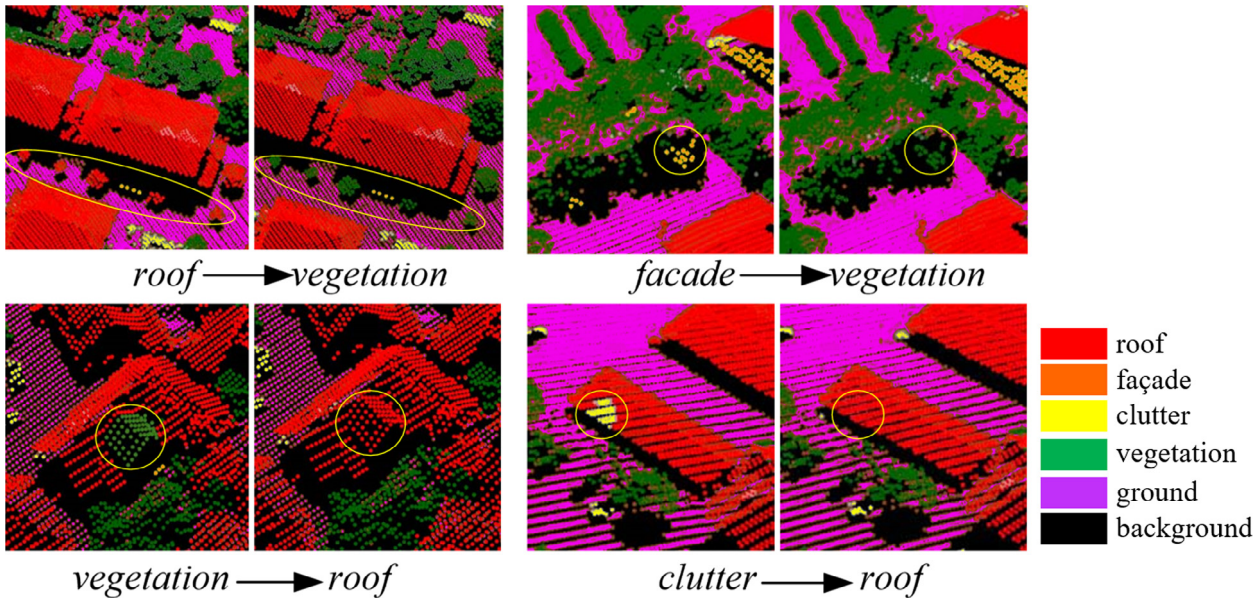


Fig. 7. Four types of misclassification corrections through class–knowledge-based semantic rules. Color code: red represents *roof*, yellow represents *façade*, green represents *vegetation*, gray represents *clutter* and brown represents *ground*. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(4) F_g: distribution of ground points around the supervoxel (Sánchez-Lopera and Lerma, 2014) measured by a circular region with radius r surrounding the supervoxel, which is divided into 16 angular bins and represented by the percentage of the bins containing ground points. This feature can distinguish between small (clutters, vegetation) and large (buildings) objects.

(5) F_c: consistency between neighborhoods measured by the number of adjacent supervoxels, which share similar normal vectors with the current supervoxel. If the angle is less than threshold τ, they are regarded to be consistent, and $\tau = 5°$ is generally adopted in this paper. The supervoxels corresponding to roofs and facades usually have at least one consistent adjacent supervoxel, whereas those corresponding to vegetation and clutters barely have any consistent neighbors.

(6) F_s: compactness of shape, measured by the area divided by the square of the perimeter of the convex hull, on the 2D points projected orthogonally onto the horizontal plane. If the supervoxel corresponds to erect objects such as facades and fences, its projected 2D polygon will approximate to a linear shape with a small compactness.

As described above, F_e and F_s are only suitable for supervoxel representation. They have proven effective in distinguishing one or more classes from the others. Fig. 4 shows each feature's capacity to distinguish, with color-coded feature values. Fig. 5 illustrates the differences between the features $F_e \sim F_c$ extracted from points, and those extracted from supervoxels. It can be noted that supervoxel-based features are much more stable than point-based features at the boundaries of objects, which can significantly improve the initial labeling results.

Before computing the unary energy of MRF defined in Eq. (1), the features, except for F_c, are normalized to [0,1] by truncated normalization to balance their contributions as

$$x = \begin{cases} 0, & \text{if } F < F_{\min} \\ 1, & \text{if } F > F_{\max} \\ (F - F_{\min})/(F_{\max} - F_{\min}), & \text{otherwise} \end{cases} \quad (4)$$

where x is the normalized value, F_{min} and F_{max} are truncating thresholds which are determined based on the distribution of the original feature value F.

The feature F_c, unlike other features, is discrete, as in Fig. 4(f), and F_c takes 0 for vegetation and clutters, 1 for the borders of roofs and facades and a value larger than 1 for the inside portions of roofs and facades. Thus, F_c is scaled as

$$x_c = 1 - \exp(-\gamma \cdot F_c) \quad (5)$$

where $\gamma = 0.7$ is a constant to adjust the normalization. Furthermore, because some features are only discriminated and contributed for specified classes, we summarize the discrimination of all of the features with regard to the label space in Table 1, in which "+" means that the feature tends to take a large value in this class, "−" means a small value and "/" means that the feature is not discriminative for this class. To simplify the description, we define the following operator $\oplus$ with regard to a normalized feature x and each label $y \in$ as,

$$x \oplus y = \begin{cases} 1 - x, & \text{if "+"} \\ x, & \text{if "−"} \\ 0, & \text{if "/"} \end{cases} \quad (6)$$

Furthermore, based on common knowledge, several semantic rules are also applied in energy computation, including: (1) F_e for

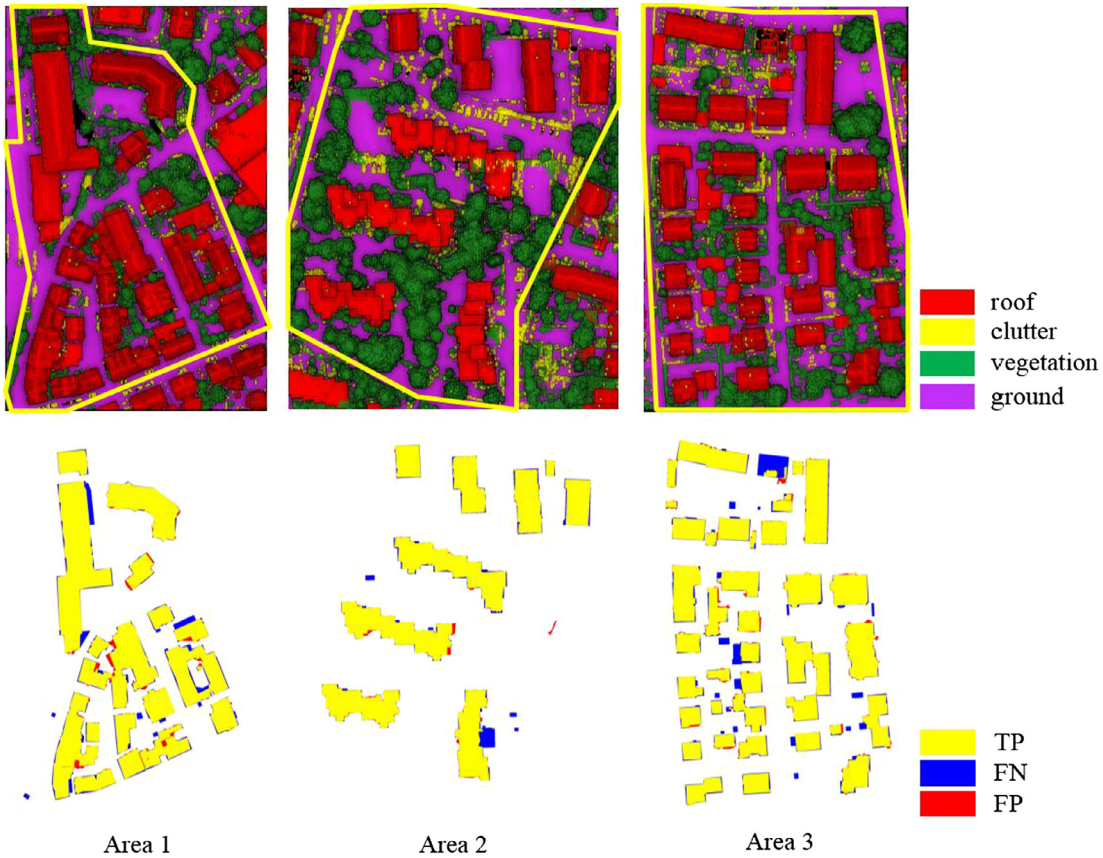


Fig. 8. Classification results and visualized evaluation on per-pixel level of three Vaihingen test sites with ALS point clouds.

roofs must be larger than a threshold τ_e and 3.0 m is used; (2) F_e for clutters must be smaller than τ_e; (3) F_d for roofs must be smaller than a threshold τ_d and 45° is used; and (4) F_c for roofs and facades must be larger than 1. The values of τ, τ_e and τ_d are semantically defined based on common knowledge, and generally are suitable in all scenes. The unary energy is defined as,

$$D(\mathbf{x},y)=\begin{cases}\infty, & \text{if against rules}\\ \frac{1}{n}\sum_{x\in\mathbf{x}}x\oplus y, & \text{otherwise}\end{cases}$$
$$\sum_{i\in\mathbf{v}}E(y_i)=\sum_{i\in\mathbf{v}}D(\mathbf{x}_i,y_i) \tag{7}$$

where $\mathbf{x}=[x_e, x_d, x_r, x_g, x_c, x_s]^T$ is the feature vector for a supervoxel.

3.3.2. Weighted Potts model

The penalties of pairwise connections make the labeling results piecewise smooth by favoring the continuities in labels $\mathbf{y}$. The Potts model has been shown to work well for this purpose in many previous studies (Gerke and Xiao, 2014; Lafarge and Mallet, 2012; Poullis, 2013), and is given as

$$E(y_i,y_j)=w_{ij}\mathbf{1}[y_i,y_j] \tag{8}$$

where $(i,j)\in\mathbf{e}$ is a pair of connected vertices in the graph, w_{ij} is the weights for the edge and $\mathbf{1}[\bullet]$ is the binary function (which is 1 if $y_i=y_j$ and 0 otherwise). There are several strategies for determining weights, including constant values (Gerke and Xiao, 2014), by distance (Zhou et al., 2012) and by distance along the angle between normal vectors (Verdie et al., 2015). Because the point homogeneity, as shown in Eqs. (2) and (3), has shown its capability in clustering supervoxels and measuring the probability that two adjacent segments belong to the same class, we also extend the homogeneity measurement for weight determination. For a vertex i, its coordinate, color and normal can be regarded as the mean value of the points in the corresponding supervoxel, and for a pair of connected vertices $(i,j)\in\mathbf{e}$, the homogeneity weight is given as

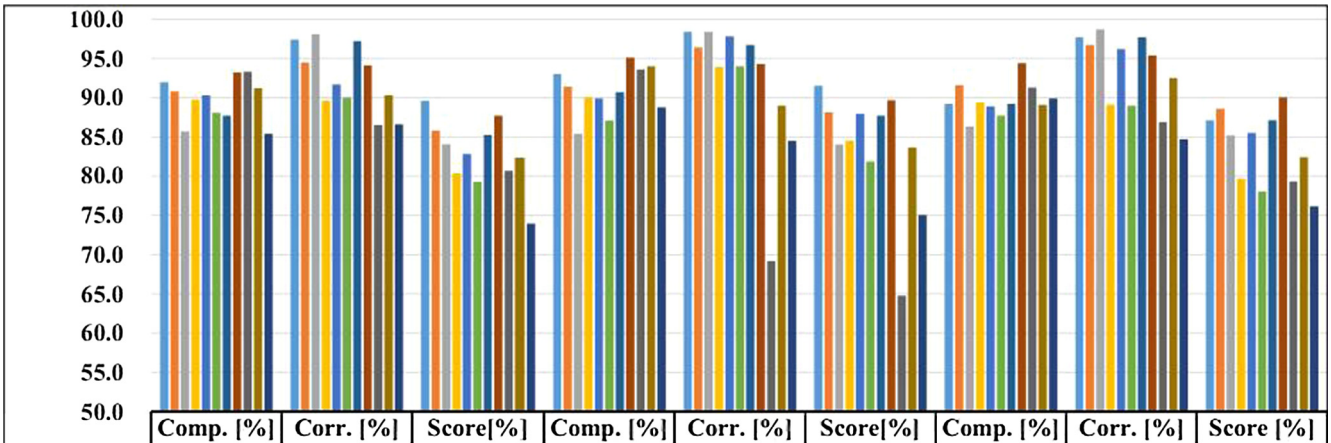
$$w_{ij}=h(i,j)=h_{Coord}(i,j)+h_{Color}(i,j)+h_{Normal}(i,j) \tag{9}$$

where h_{Coord}, h_{Color} and h_{Normal} has the same meaning and are computed as in Eqs. (2) and (3).

Fig. 6 compares the effects of different weight determination strategies on the classification results, and it can be noted that the point homogeneity measurements proposed in this paper perform best in making the results consistent and not crossing the object boundaries.

3.3.3. Knowledge-based semantics

Because MRF with general purpose, high-order cliques are difficult to solve, knowledge-based semantics are used to find the misclassified connected components and refine the classification as a post-processing step. Specifically, the unary and pairwise energies are minimized using Graph Cut (Boykov et al., 2001), and then the labels are refined according to rules enforced in the high-order cliques detected by connected components. Because the ground and non-ground are separated before classification, connected supervoxels are more likely to belong to the same object and thus share the same labels. The connected components in the supervoxel–adjacency graph are extracted, as shown in Fig. 2. Additional knowledge-based rules are enforced for single components, which have shown promising capability in improving the overall classification accuracy (Xu et al., 2014). The rules used in this paper include:



	Area1			Area2			Area3		
	Comp. [%]	Corr. [%]	Score[%]	Comp. [%]	Corr. [%]	Score[%]	Comp. [%]	Corr. [%]	Score [%]
Ours	92.0	97.4	**89.6**	93.0	98.4	**91.5**	89.2	97.7	87.1
HANC3	90.8	94.5	85.8	91.4	96.4	88.1	91.6	96.7	88.6
VSK	85.7	98.1	84.1	85.4	98.4	84.0	86.3	98.7	85.2
WHUY2	89.7	89.6	80.4	90.0	93.9	84.5	89.4	89.1	79.7
MAR_2	90.3	91.7	82.8	89.9	97.8	87.9	88.9	96.2	85.5
MON2	88.1	90.0	79.3	87.1	94.0	81.9	87.7	89.0	78.1
CAL2	87.7	97.2	85.2	90.7	96.7	87.7	89.2	97.7	87.1
LJU1	93.2	94.1	87.7	95.1	94.3	89.7	94.4	95.4	**90.1**
ITCM	93.3	86.5	80.7	93.6	69.2	64.8	91.3	86.9	79.3
ITCR	91.2	90.3	82.4	94.0	89.0	83.7	89.1	92.5	82.4
FED_2	85.4	86.6	74.0	88.8	84.5	75.0	89.9	84.7	76.1

Fig. 9. Comparison of our method and other point cloud classification methods. HANC3 to MON2 denote methods using only point clouds, and CAL2 to FED_2 denote methods using both point clouds and images.

(1) in a connected component, the total 2D area of connected supervoxels labeled as *roof* must be larger than threshold τ_A (e.g., 20m^2). Here, the value of τ_A is conservatively defined, to ensure this semantic rule is true in most cases;

(2) in a connected component, the total 2D area of connected supervoxels labeled as *clutter* must be smaller than threshold τ_A;

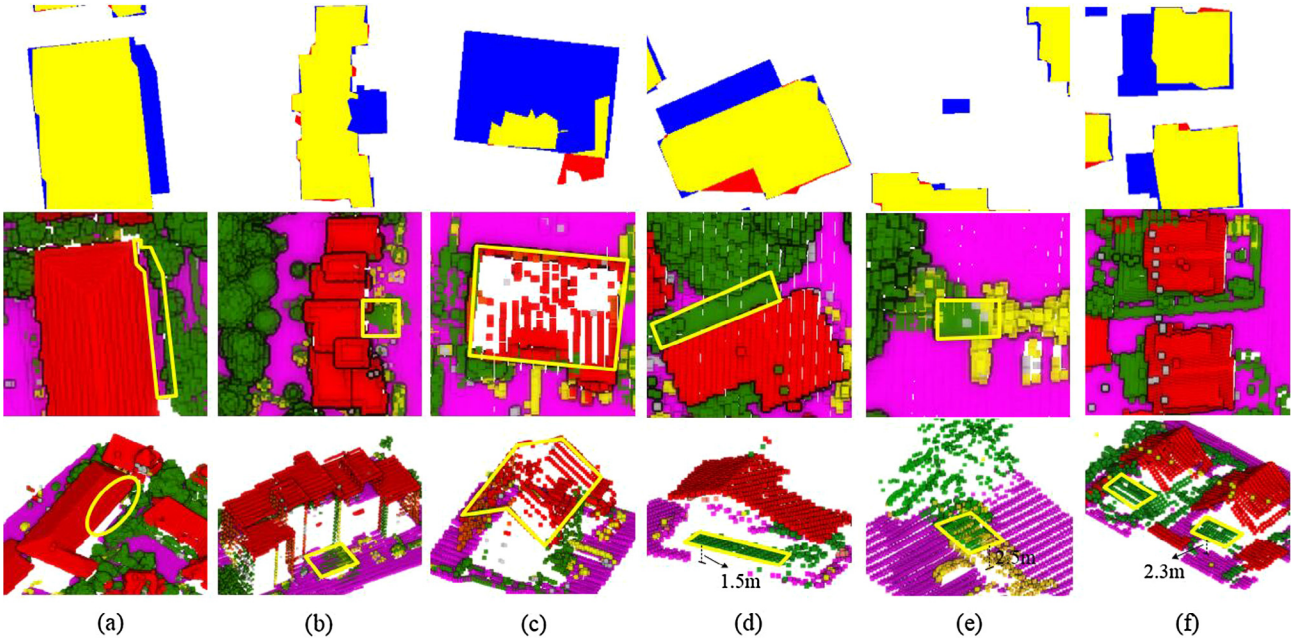


Fig. 10. Analysis of misclassified regions in Area1-3. The first row is the misclassified regions shown in the benchmark result, the second row is the corresponding regions in the classified point clouds from the top view, and the last row is the corresponding regions in the classified point clouds from the perspective view. (a)–(c) are large (>50 m^2) false-negative detections in Area1-3, and (d)–(f) are small (<50 m^2) false-negative detections with low height differences in Area1-3.

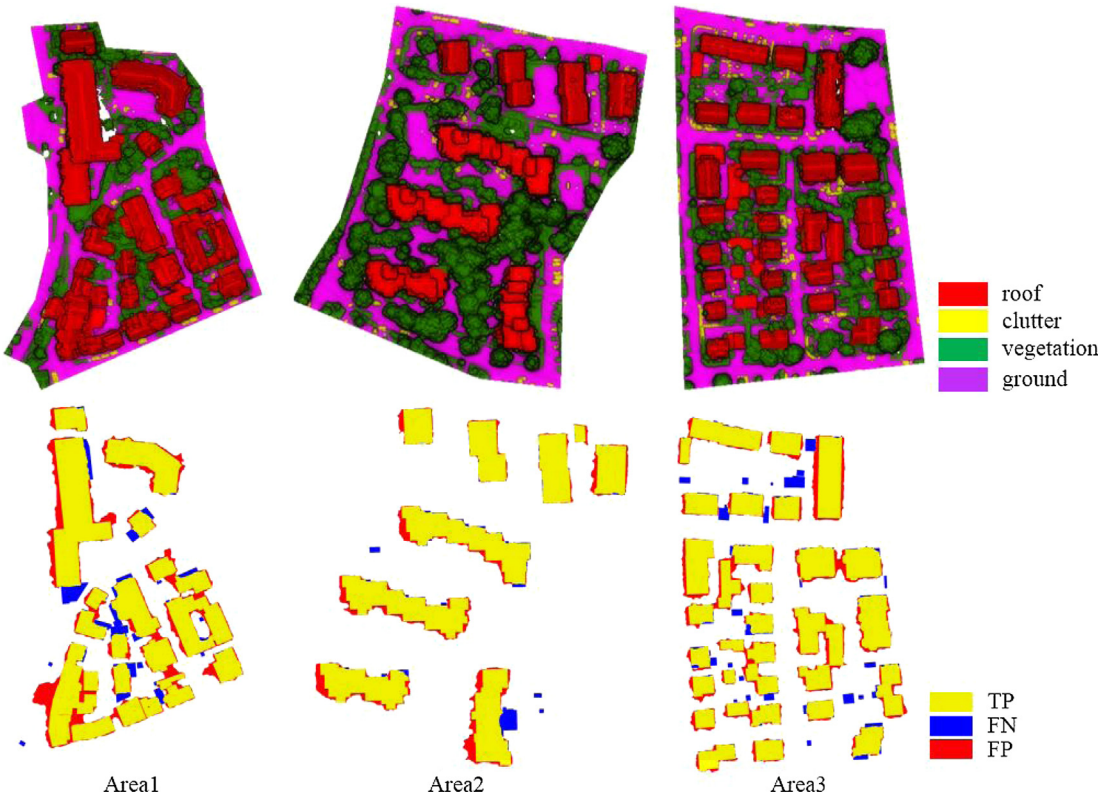


Fig. 11. Classification results and visualized evaluation on per-pixel level of the Vaihingen DSM data.

(3) if a connected component contains supervoxels labeled as *façade*, it must also contain supervoxels labeled as *roof*, and these two types supervoxels are directly connected;
(4) if a connected component contains vegetation, the total number of connected supervoxels labeled as *vegetation* must be larger than the number of adjacent supervoxels labeled as *façade* or *roof*; and
(5) if a connected component contains clutter, the total number of connected supervoxels labeled as *clutter* must be larger than the number of adjacent supervoxels labeled as *façade* or *roof*.

If a connected component does not meet the abovementioned semantic rules, the corresponding supervoxels in it will be relabeled as other classes according to their unary energy, excluding the labels rejected by the rules. According to Verdie et al. (2015), this rule-based refinement process can effectively correct local misclassification caused by roof superstructures (e.g., chimneys, dormers) and similarities between the vertical components of vegetation and facades. The result of refinement is shown in Fig. 7. After refinement, the labeled non-ground points are fused with ground points to generate the final classified point cloud.

Table 2
Completeness and correctness of the classification results on per-pixel level of the Vaihingen DSM data.

Area1		Area2		Area3	
Comp. [%]	Corr. [%]	Comp. [%]	Corr. [%]	Comp. [%]	Corr. [%]
91.8	80.3	95.8	85.9	92.2	82.6

4. Experimental analysis

To verify the efficiency of the proposed method, two tests with different datasets are reported in this paper. The first test uses the ISPRS benchmark data from Vaihingen, Germany and Toronto, Canada, as provided by the ISPRS benchmark (Rottensteiner et al., 2012), and the corresponding benchmark results are carried out by the organizer. This dataset contains three test sites in Vaihingen (Area1 through Area3) and two test sites in Toronto (Area4 and Area5). The second dataset is the dense-matched point clouds produced by the MVS pipeline (Vu et al., 2012) using the penta-view oblique images over Shenzhen, China. The details of evaluation protocols and datasets are described below.

4.1. Evaluation using the ISPRS benchmark data

4.1.1. Vaihingen sites

The Vaihingen sites (Cramer, 2010) have relatively lower buildings, which are generally lower than 20 m with classical architecture. However, the scenes from all three sites are complex, as the objects are close to, or even overlap each other in both the horizontal and vertical directions. Even so, our method performs well in classifying the point clouds of these sites. Fig. 8 shows the classification results from ALS point clouds and corresponding benchmark evaluations. Fig. 9 shows the quantitative statistics with respect to the classified areas for the first three areas, because rather than an object-level evaluation, an area evaluation is directly related to the classification accuracy. Both correctness and completeness are evaluated for each area, and the score is determined by the product of the two factors. For the Vaihingen sites, we compare our method with ten others, the first five of which use point cloud data exclu-

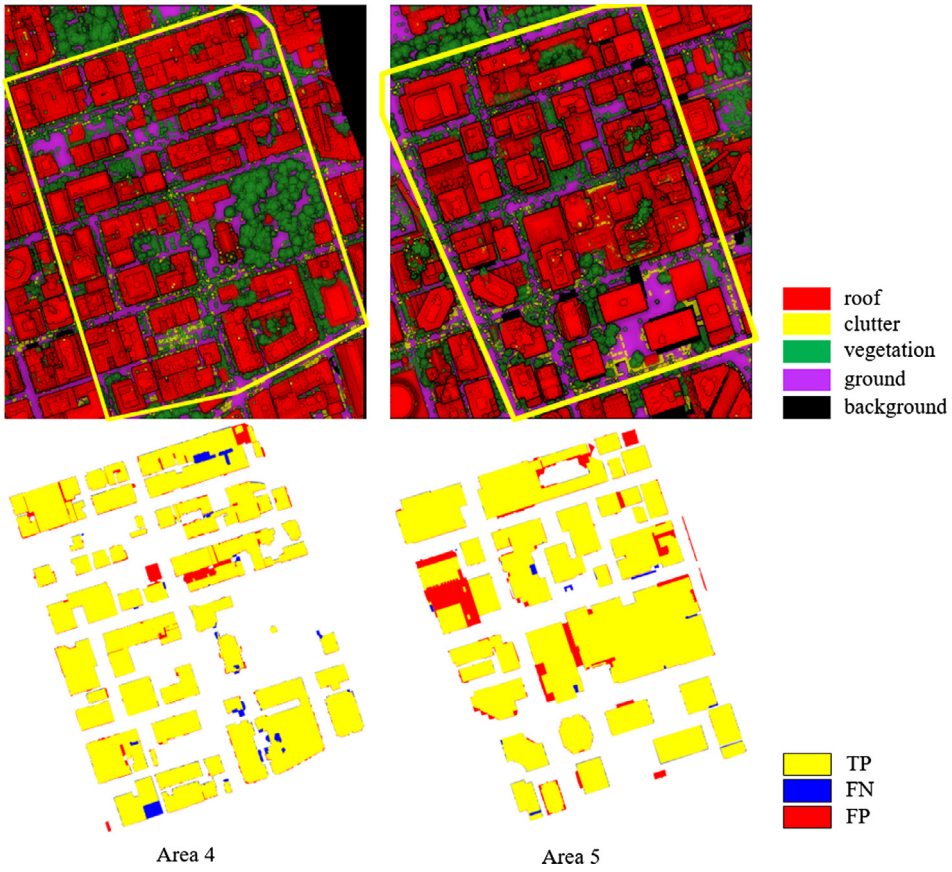
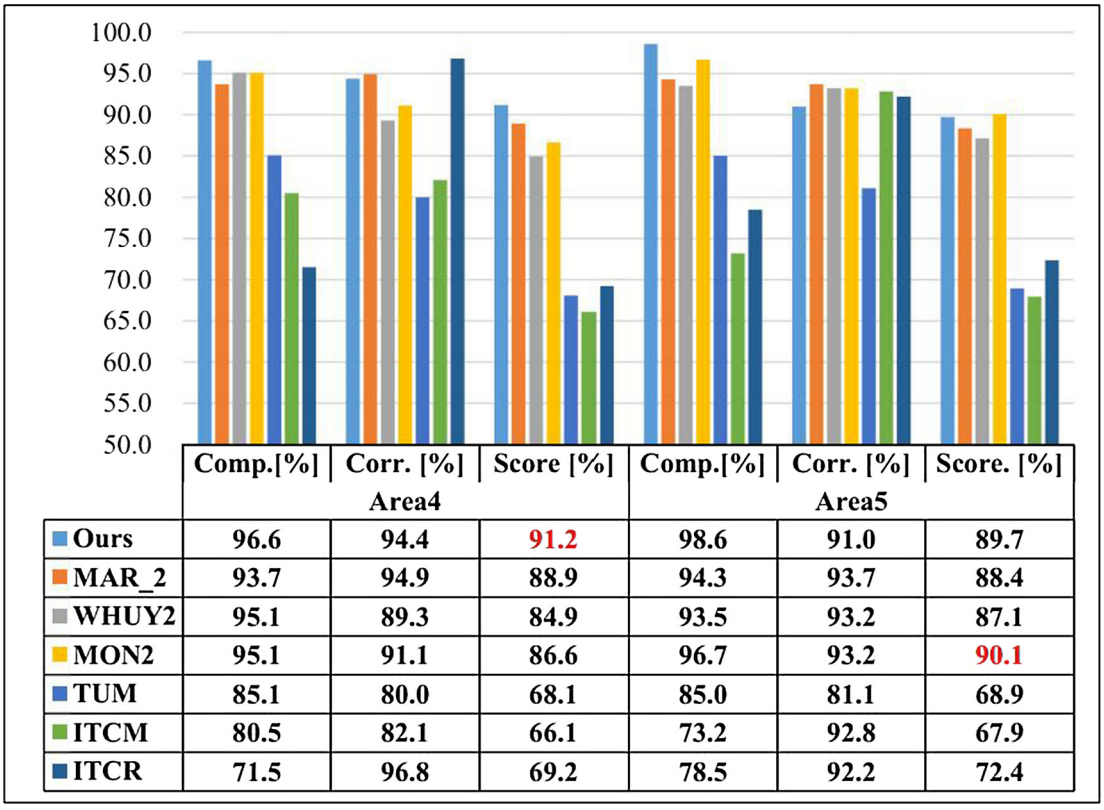


Fig. 12. Classification results of two Toronto test sites and visualized evaluation on per-pixel level.



	Area4			Area5		
	Comp.[%]	Corr. [%]	Score [%]	Comp.[%]	Corr. [%]	Score. [%]
Ours	96.6	94.4	**91.2**	98.6	91.0	89.7
MAR_2	93.7	94.9	88.9	94.3	93.7	88.4
WHUY2	95.1	89.3	84.9	93.5	93.2	87.1
MON2	95.1	91.1	86.6	96.7	93.2	**90.1**
TUM	85.1	80.0	68.1	85.0	81.1	68.9
ITCM	80.5	82.1	66.1	73.2	92.8	67.9
ITCR	71.5	96.8	69.2	78.5	92.2	72.4

Fig. 13. Comparison of our method and other point cloud classification methods. MAR_2 to MON2 denote methods using only point clouds, and TUM ~ ITCR denote methods using both point clouds and images.

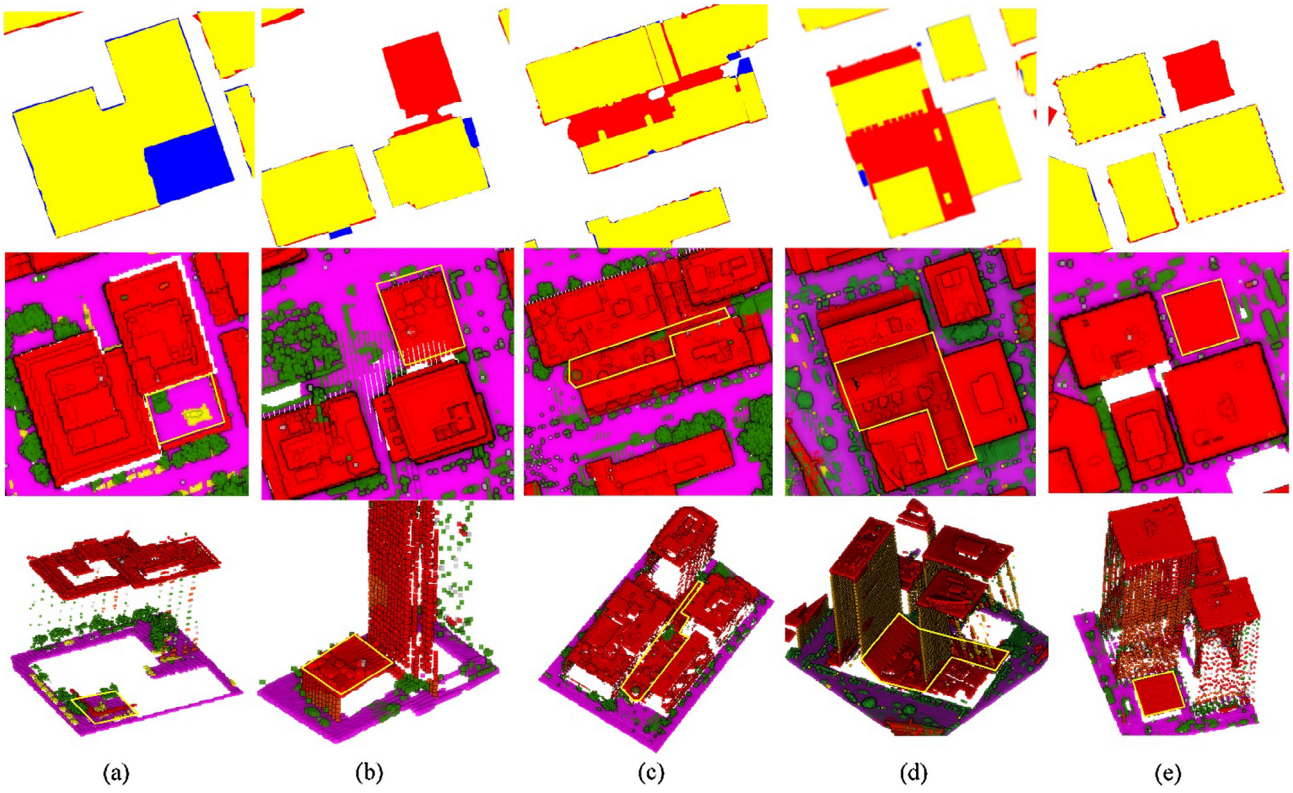


Fig. 14. Analysis of misclassified region. The first row is the misclassified regions shown in the benchmark result, the second row is the corresponding regions in the classified point clouds from the top view, and the last row is the corresponding regions in the classified point clouds from the perspective view. (a) is the main false-negative detection in area4, and (b)–(e) are the false-positive detections in area4 and area5.

sively, the same as the proposed method, and the latter five of which use both point clouds and the corresponding orthophotos with four spectral bands.

The scores in Fig. 9 suggest that the proposed method performs best in two out of three areas, and is also one of the best in area3. The correctness which is at least 97% also indicates the efficiency of our method in these three areas. With respect to the completeness factor it can be noted that in these three areas buildings larger than 50 m^2 are detected correctly, except those shown in Fig. 10(a)–(c). The interested region in Fig. 10(a) is correctly classified, but somehow the missing out building structure is not represented in the point cloud. In Fig. 10(b) the false-negative part is a platform with height lower than 1.0 m and is firstly mis-labeled as ground in the filter stage. And the building in the interested region in Fig. 10(c) is missing because the point cloud is sparse and does not integrate into a polygon for benchmark evaluation. In addition, objects smaller than 50m^2 are more likely to be omitted, especially when their heights are close to or lower than the height threshold τ_e as shown in Fig. 10(d)–(f). This may result from two reasons. First, there are inadequate points located at these small objects and therefore discriminative features cannot be derived from the corresponding supervoxles. Second, the low elevation differences above ground do not conform to the semantic rules set for buildings, and consequently the points are labeled as other classes.

The digital surface model (DSM) with 9 cm resolution of Vaihingen is also used to verify the efficiency of our method, and both the qualitative and quantitatative evaluations are shown in Fig. 11 and Table 2, respectively. Compared with the ALS classification results, while the completeness values almost maintain the same, the correctness values are obviously lower due to the increasing false-positive detections around building boundaries. Considering the smooth change of gradient at the object edge and the defects of the DSM, the DSM classification results are also acceptable.

4.1.2. Toronto sites

The two Toronto sites feature high-rise buildings with podium gardens on some building roofs, which increase the difficulties of the scene classification. The classification and testing results for the benchmark are shown in Fig. 12. Further comparisons with other methods are summarized in Fig. 13. Because there are fewer submissions on the Toronto dataset, we only compare our method with six others, the first three of which only use point clouds and

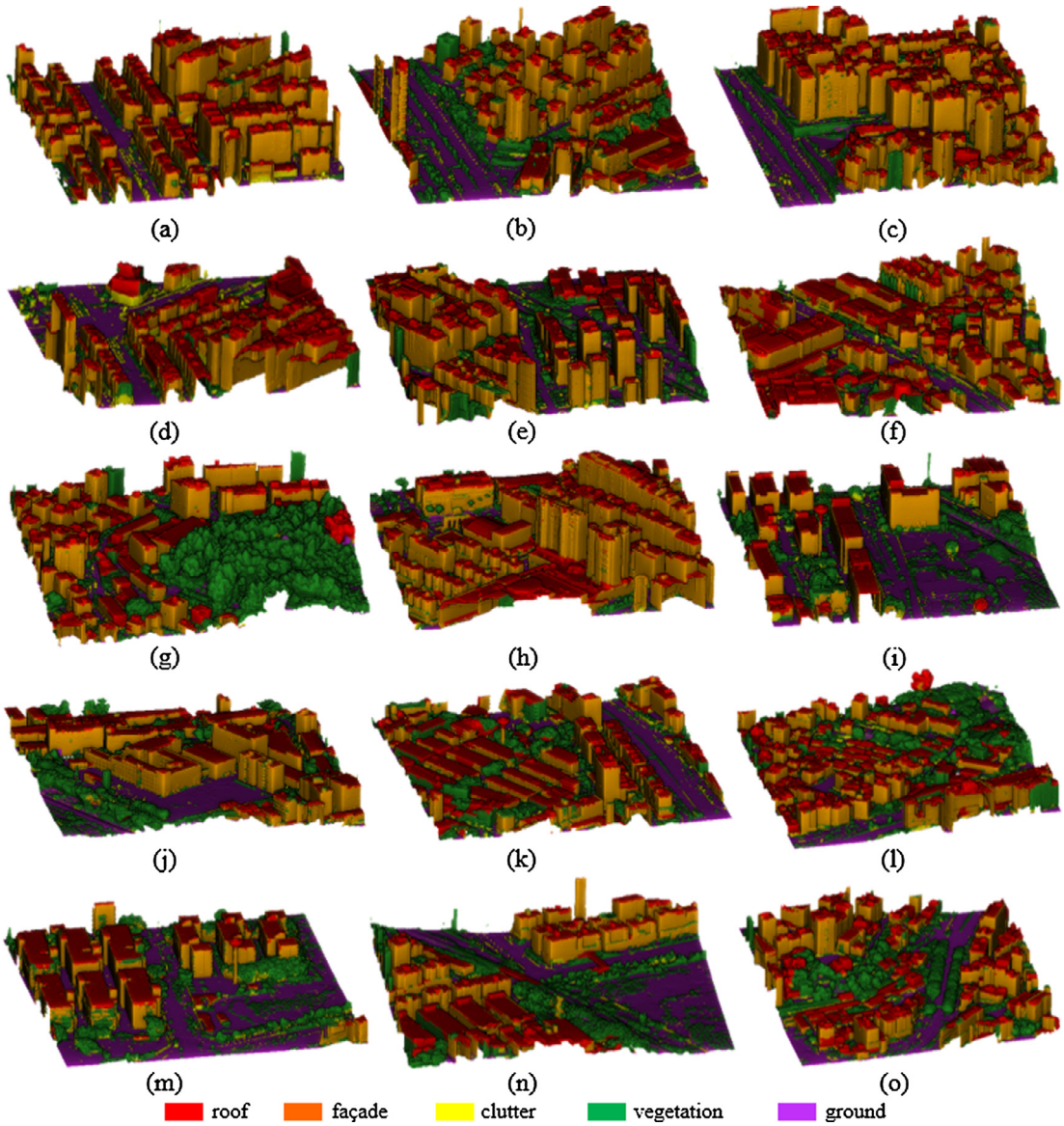


Fig. 15. Classification results of 15 dense-matched point cloud tiles.

the other three of which use both orthophotos and point clouds. The orthophotos in the Toronto dataset do not provide the near infrared band, so only the pseudo-NDVI is available. This explains the performance degradation of methods integrating point clouds and images in the Toronto dataset. The proposed method is also among the best for this dataset.

The visualized evaluation indicates that the main false-negative detection of Toronto sites locates at the bottom of area4, details of which are shown in Fig. 14(a). From Fig. 14(a) it can be seen that the main missing part corresponds to a courtyard having the same height with ground, and without texture information this part is most likely to be considered as ground. There are more false-positive detections in area4 and area5, which lead to lower area-correctness in Fig. 13 compared with the foregoing three areas. Further analyses of the specified areas are shown in Fig. 14(b)–(e). It can be noted that the ground is misclassified as building roofs. However, these parts in fact are extruded ground attached to the buildings, and 3*m* higher than the ground in the neighborhood. Thus, this is a misclassification of the ground filter (Hu et al., 2014), rather than non-ground point classification. The only reasonable label for this area is building roofs.

4.2. Evaluation using photogrammetric data

Because our major interest is processing photogrammetric point clouds from aerial oblique images, massive experiments on the tiled point clouds generated by the existing MVS pipeline are evaluated in this paper. The oblique images have a ground sample distance (GSD) of about 8 *cm* for the nadir views and 6–16 cm for the four oblique views. The point clouds are generated in a tiled way, with each tile covering an area of about 250 m × 250 m. The merit of the photogrammetric point clouds is that facades are better observed and recovered. However, the point clouds are defects-laden and sharp features are not well preserved, compared to the ALS point clouds. We select 15 tiles of typical scenarios and the results are shown in Fig. 15, according to which, most of the surface objects are classified correctly, including the high towers in (i), (k) and (m), and the bridges in (n), which we generally group into the "building" category. To evaluate the results quantitatively, four tiles ((a), (h), (j) and (m) in Fig. 15) are selected and benchmarked based on our self-labeled ground-truth reference as shown in Fig. 16. The first two tiles ((a) and (h)) feature dense buildings and little vegetation, and the last two tiles ((j) and (m)) have more vegetation close to buildings, while the buildings are much further from each other. The completeness and correctness evaluations are shown in Table 3.

Compared with the ALS results of Toronto sites, which feature similar construction style with Shenzhen, the average completeness is relatively lower for the following reasons. First, although the photogrammetric point clouds have quite high density compared to ALS, and small objects are also observed (e.g., parapets and staircases), the low accuracies and noises lead to high roughness of building points, as shown in Fig. 17(a). Second, objects at the edge of each tile are more likely to be classified incorrectly because they are usually cut into parts and the incomplete parts may not conform to our semantic rules (e.g., a small part of a building may not conform to the rule of area constraint). With respect to the correctness, the reasons lead to false-positive detections are as the follows. First, rather than the intuitive sharp corner between roofs and facades/vegetation, in practice, the transition from facades/vegetation to roofs in photogrammetric point cloud is generally smoothed like shown in Fig. 17(a), and thus causes problems in the connectivity graph for MRF. Second, unlike ALS, which can penetrate vegetation and recover points beneath it, the existing MVS pipeline does not recover points under vegetation, which produces a smoothed 2.5D depth surface in the camera view, as shown in Fig. 17(b), making vegetation less distinctive from building points. Third, although oblique images can obtain the details of objects from different angles, occlusions are inevitable. Therefore, some unexpected points may be generated during dense-matching. These incorrectly dense-matched points change objects into structures that do not conform to the semantics, reducing the accuracy, as denoted in Fig. 17(c)–(e).

To investigate the impact of oblique views on the classification results, DSMs of 0.1 m resolution are generated from nadir views for the four selected tiles and classified using the proposed method. The visualized and quantitative evaluations of the DSM results are also shown in Fig. 16 and Table 3 respectively. An interesting finding is that classification results of DSM generally have lower completeness but higher correctness. The reason for lower completeness is because of omit of facades in DSM. However, the higher correctness is counter-intuitive. After detailed analyses,

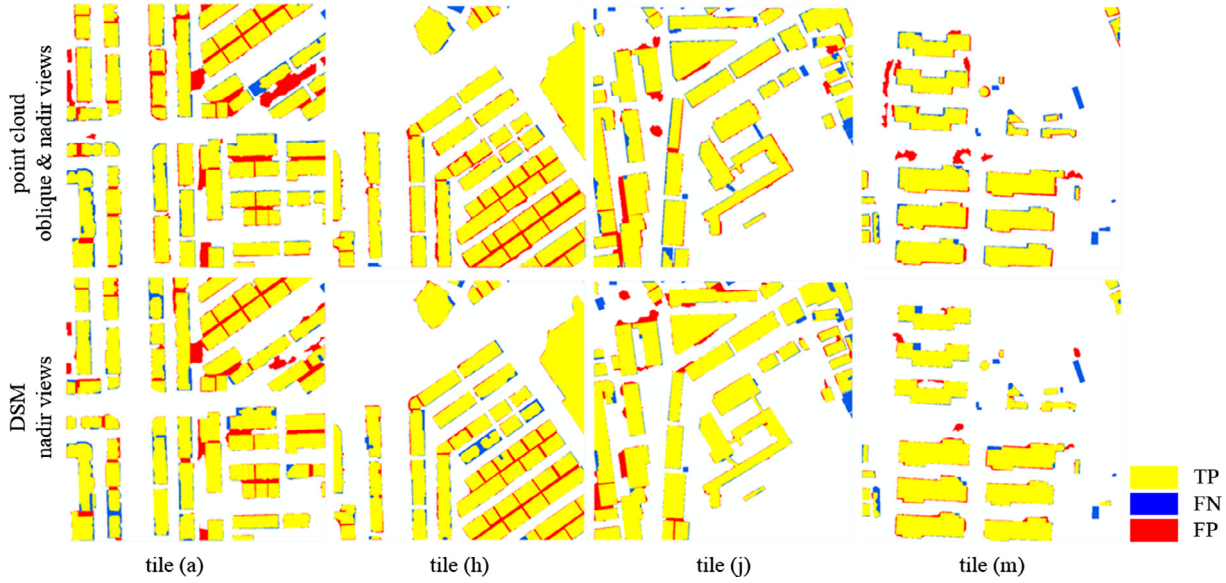


Fig. 16. Visualized evaluation on per-pixel level. The first row shows the results of the dense-matched point clouds from oblique views as well as nadir views, and the second row shows the results of the 0.1 m DSM generated from the nadir views.

we found that this is because that in the occluded regions, especially those under the trees as shown in Fig. 18, the point clouds generated from existing MVS (Vu et al., 2012) pipeline are interpolated from the triangulated meshes and thus is almost planar and recognized as buildings. In fact, in the MVS pipeline (Vu et al., 2012) the initial triangulated meshes are first created from a sparse (or semi-dense as denoted in the paper) point clouds and refined by photometric consistency. The point clouds are just sampled

Table 3
Completeness and correctness of the classification results on per-pixel level of the dense-matched point cloud and DSM data.

	Tile (a)		Tile (h)		Tile (j)		Tile (m)	
	Comp. [%]	Corr. [%]	Comp. [%]	Corr. [%]	Comp. [%]	Corr. [%]	Comp. [%]	Corr [%]
Point clouds	95.6	84.4	97.7	89.8	96.1	90.6	90.9	87.3
DSM	94.3	87.8	95.9	92.2	94.7	92.7	94.3	94.2

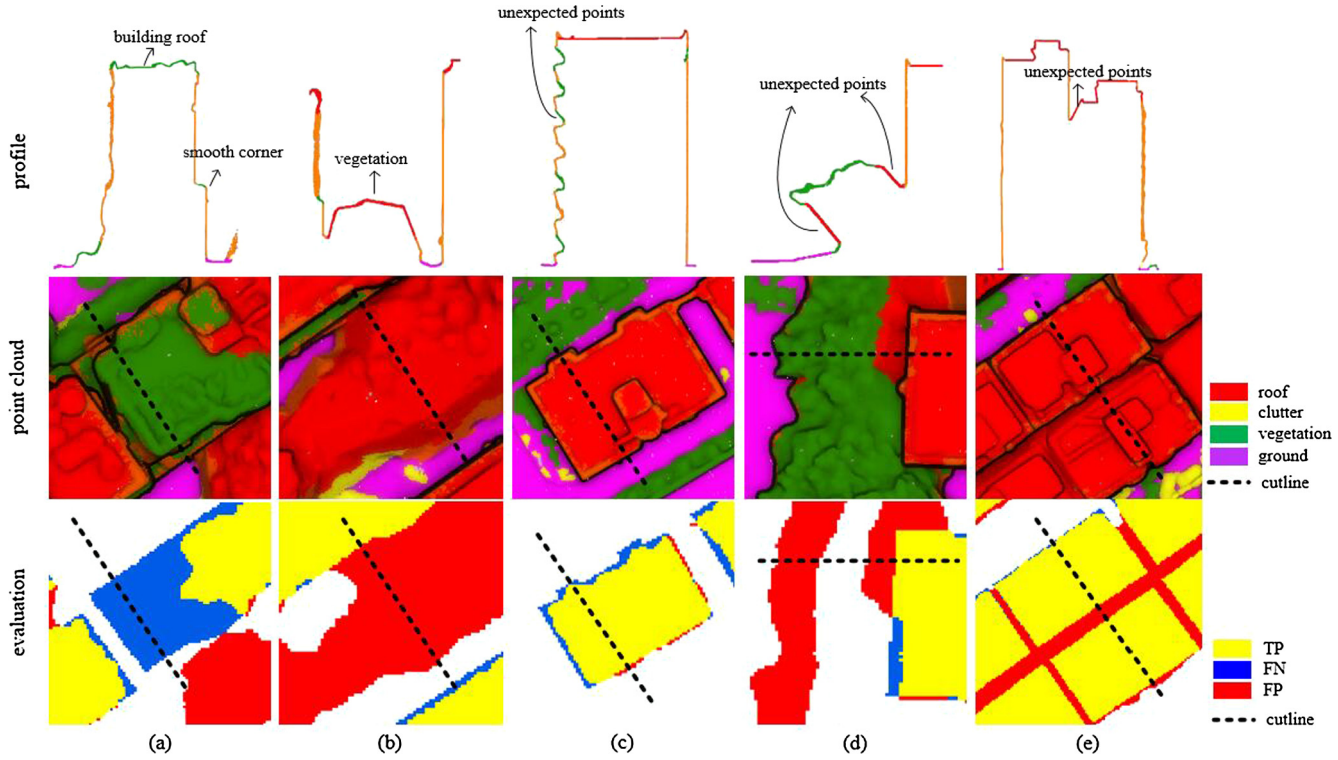


Fig. 17. Profile analysis of the classified dense-matched point clouds. (a) denotes the false-negative detection caused by the rough expression of building roof, (b) denotes the false-positive detection caused by the smooth expression of vegetation surface, and (c)–(e) show the unexpected points caused by occlusions.

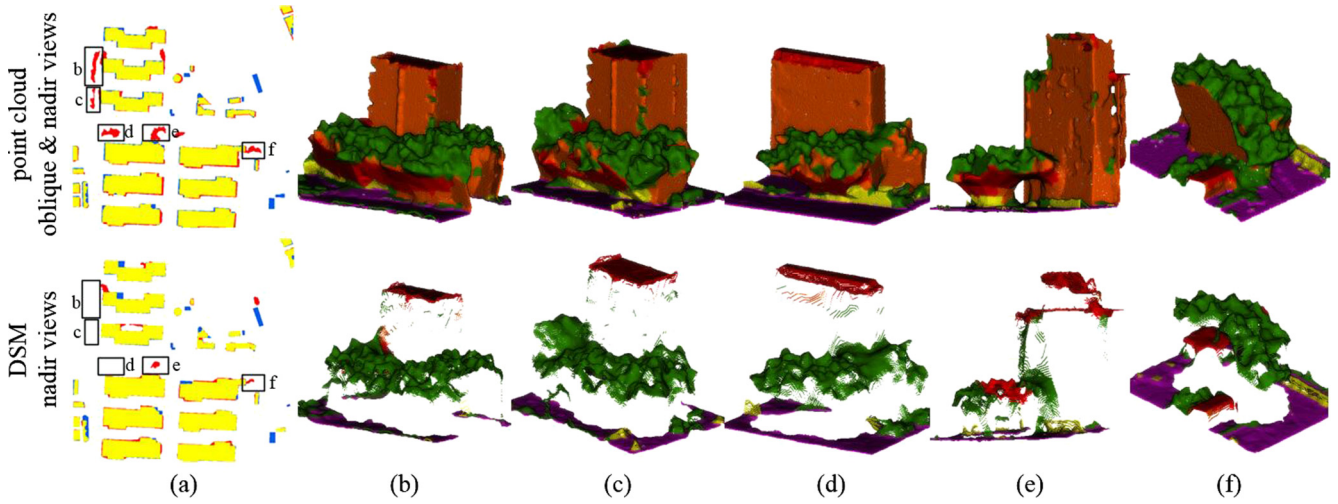


Fig. 18. Impact of occlusions on the classification results. (a) is the visualized evaluations of dense-matched point cloud result and DSM result, and (b)–(f) shows the details of the classified point cloud/DSM in the rectangle areas in (a).

from the mesh and enriched by color information by re-projection to the image.

In general, considering the above factors, the overall performance of our method is quite satisfactory because the supervoxel-based method reduces the influence of noise and exploits more contextual information. Although due to the differences between the scenes and data qualities, the thresholds used for truncated normalization are set differently for each dataset, nonetheless they are set exactly the same for each area or tile in the same dataset. In fact, our method provides an interactive way for adjusting the thresholds between different datasets by visualizing the discriminations of features, which makes our method transferable in practical applications. Furthermore, in practice, manual quality control is inevitable and the supervoxel-based approach is more friendly for interactive editing. Compared to classical point cloud software, in which profile analysis and inspection are the standard approach for point cloud editing in 2D views, the proposed method can represent a supervoxel with a polygon patch that can be directly selected and annotated in 3D space. This provides a unified and seamless interface for single-click inspection and editing.

5. Conclusions

In this paper, we propose a novel supervoxel-based method for the automatic classification of point clouds data in urban scenes. The main contribution of our work is the combination and comprehensive uses of multi-level semantic relationships, including point–homogeneity, supervoxel–adjacency and class–knowledge constraints. These multi-level constraints are modeled in a high-order graphical model, and take effects interactively in a MRF framework to generate 3D labeled point cloud automatically.

Two experiments are performed with ALS data provided by ISPRS and photogrammetric point clouds obtained from aerial images, respectively. The quantized evaluation of the ALS dataset is compared with other homeochronous methods, and the high rank reveals the effectiveness of the proposed method. For the second experiment, although the low accuracy, excessive noise, and additional clusters of the photogrammetric dataset will significantly increase the difficulty of classification, the proposed method is indicated to be robust and transferable by the high correctness and completeness in the evaluations.

In this paper there are still some limitations. Due to the difficulty of minimizing high-order energy, we resort to a two-step strategy that minimizes pairwise energy and post-processing refinement. Future work on classification may be focused on exploiting recent discrete optimization methods based on mixed integer programming (Boukouvala et al., 2016) to solve high-order energy problems. Furthermore, based on the semantically enriched point clouds, the polygonal reconstruction of buildings (Arikan et al., 2013) will also be investigated.

Acknowledgments

This study was supported by grants from the National Natural Science Foundation of China (NO. 441501421, 41631174, 61602392), the Foundation of Key Laboratory for Geo-Environmental Monitoring of Coastal Zone of the National Administration of Surveying, Mapping and Geoinformation, Open Research Fund of State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (NO. 15I01) and a grant from the Research Grants Council of Hong Kong (Project No. PolyU 5330/12E). The Vaihingen dataset was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) [Cramer, 2010]: http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html.

References

Achanta, R. et al., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34 (11), 2274–2282.
Aijazi, A.K., Checchin, P., Trassoudaine, L., 2013. Segmentation-based classification of 3D urban point clouds: a super-voxel based approach with evaluation. Remote Sens. 5 (4), 1624–1650.
Arikan, M., Schwärzler, M., Flöry, S., Wimmer, M., Maierhofer, S., 2013. O-snap: optimization-based snapping for modeling architecture. ACM Trans. Graph. (TOG) 32 (1), 6.
Babahajiani, P., Fan, L., Kamarainen, J., Gabbouj, M., 2015. Automated super-voxel-based features classification of urban environments by integrating 3D point cloud and image content. In: Proc. IEEE International Conference on Signal and Image Processing Applications (ICSIPA), IEEE, Kuala Lumpur, Malaysia.
Bláha, M., Vogel, C., Richard, A., Wegner, J.D., Pock, T., Schindler, K., 2016. Large-scale semantic 3D reconstruction: An adaptive multi-resolution model for multi-class volumetric labeling. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA.
Boulch, A., Marlet, R., 2012. Fast and robust normal estimation for point clouds with sharp features. In: Proc. Computer Graphics Forum. Wiley Online Library, pp. 1765–1774.
Boukouvala, F., Misener, R., Floudas, C.A., 2016. Global optimization advances in mixed-integer nonlinear programming, MINLP, and constrained derivative-free optimization, CDFO. Eur. J. Oper. Res. 252 (3), 701–727.
Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. 23 (11), 1222–1239.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Chehata, N., Guo, L., Mallet, C., 2009. Airborne LiDAR feature selection for urban classification using random forests. Int. Arch. Photogramm., Remote Sens. Spatial Inform. Sci. XXXVIII (Part 3), 207–212.
Cramer, M., 2010. The DGPF-test on digital airborne camera evaluation overview and test design. Photogramm. Fernerkundung Geoinform. 7 (2), 73–82.
Dang, M., Lienhard, S., Ceylan, D., Neubert, B., Wonka, P., Pauly, M., 2015. Interactive design of probability density functions for shape grammars. ACM Trans. Graph. (TOG) 34 (6), 206.
Esri, 2016. City Engine <http://www.esri.com/software/cityengine> (accessed 14th September, 2016).
Furukawa, Y., Ponce, J., 2010. Accurate, dense, and robust multi-view stereopsis. IEEE Trans. Pattern Anal. Mach. Intell. 32 (8), 1362–1376.
Gerke, M., Xiao, J., 2014. Fusion of airborne laser scanning point clouds and images for supervised and unsupervised scene classification. ISPRS J. Photogramm. Remote Sens. 87, 78–92.
Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random forests for land cover classification. Pattern Recogn. Lett. 27 (4), 294–300.
Guo, L., Chehata, N., Mallet, C., Boukir, S., 2011. Relevance of airborne LiDAR and multispectral image data for urban scene classification using random forests. ISPRS J. Photogramm. Remote Sens. 66 (1), 56–66.
Hackel, T., Wegner, J.D., Schindler, K., 2016. Fast semantic segmentation of 3D point clouds with strongly varying density. ISPRS Ann. Photogramm., Remote Sens. Spatial Inform. Sci. III 3, 177–184.
Hu, H., Chen, C., Wu, B., Yang, X., Zhu, Q., Ding, Y., 2016. Texture-aware dense image matching using ternary census transform. ISPRS Ann. Photogramm., Remote Sens. Spatial Inform. Sci. III 3, 59–66.
Hu, H., Ding, Y., Zhu, Q., Wu, B., Lin, H., Du, Z., Zhang, Y., Zhang, Y., 2014. An adaptive surface filter for airborne laser scanning point clouds by means of regularization and bending energy. ISPRS J. Photogramm. Remote Sens. 92, 98–111.
Kim, B., Kohli, P., Savarese, S., 2013. 3D scene understanding by Voxel-CRF. In: Proc. IEEE International Conference on Computer Vision (ICCV), Sydney, Australia.
Kumar, S., Hebert, M., 2006. Discriminative random fields. Int. J. Comput. Vision 68 (2), 179–201.
Lafarge, F., Mallet, C., 2012. Creating large-scale city models from 3D-point clouds: A robust approach with hybrid representation. Int. J. Comput. Vision 99 (1), 69–85.
Li, S.Z., 2009. Markov Random Field Modeling in Image Analysis. Springer Science & Business Media, pp. 29–39.
Lim, E.H., Suter, D., 2009. 3D terrestrial LiDAR classifications with super-voxels and multi-scale Conditional Random Fields. Comput. Aided Des. 41 (10), 701–710.
Lodha, S.K., Fitzpatrick, D.M., Helmbold, D.P., 2007. Aerial LiDAR data classification using AdaBoost. In: Proc. 6th International Conference on 3-D Digital Imaging and Modeling (3DIM'07), IEEE, Montreal, Canada.
Mallet, C., Bretar, F., Roux, M., Soergel, U., Heipke, C., 2011. Relevance assessment of full-waveform LiDAR data for urban area classification. ISPRS J. Photogramm. Remote Sens. 66 (6), 71–84.
Mallet, C., Bretar, F., Soergel, U., 2008. Analysis of full-waveform LiDAR data for classification of urban areas. Photogramm. Fernerkundung Geoinform. 5, 337–349.
McClune, A.P., Mills, J.P., Miller, P.E., Holland, D.A., 2016. Automatic 3D building reconstruction from a dense image matching dataset. Int. Arch. Photogramm., Remote Sens. Spatial Inform. Sci. XLI B3, 641–648.
Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: a review. ISPRS J. Photogramm. Remote Sens. 66 (3), 247–259.

Musialski, P., Wonka, P., Aliaga, D.G., Wimmer, M., Gool, L., Purgathofer, W., 2013. A survey of urban reconstruction. Comput. Graph. Forum 32 (6), 146–177.
Nex, F., Gerke, M., 2014. Photogrammetric DSM denoising. Int. Arch. Photogramm., Remote Sens. Spatial Inform. Sci. XL 3, 231–238.
Niemeyer, J., Rottensteiner, F., Soergel, U., 2014. Contextual classification of LiDAR data and building object detection in urban areas. ISPRS J. Photogramm. Remote Sens. 87, 152–165.
Niemeyer, J., Rottensteiner, F., Soergel, U., Heipke, C., 2016. Hierarchical higher order CRF for the classification of airborne LiDAR point clouds in urban areas. ISPRS Int. Arch. Photogramm., Remote Sens. Spatial Inform. Sci. XLI B3, 655–662.
Papon, J., Abramov, A., Schoeler, M., Worgotter, F., 2013. Voxel cloud connectivity segmentation-supervoxels for point clouds. In: Proc. the IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA.
Petrie, G., 2009. Systematic oblique aerial photography using multiple digital cameras. Photogramm. Eng. Remote Sens. 75 (2), 102–107.
Poullis, C., 2013. A framework for automatic modeling from point cloud data. IEEE Trans. Pattern Anal. Mach. Intell. 35 (11), 2563–2575.
Ren, X., Malik, J., 2003. Learning a Classification Model for Segmentation. In: Proc. Computer Vision, 2013. Proceedings. Ninth IEEE International Conference on, pp. 10–17.
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. ISPRS Ann. Photogramm., Remote Sens. Spatial Inform. Sci. I 3, 293–298.
Rouhani, M., Lafarge, F., Alliez, P., 2017. Semantic segmentation of 3D textured meshes for urban scene analysis. ISPRS J. Photogramm. Remote Sens. 123, 124–139.
Sánchez-Lopera, J., Lerma, J.L., 2014. Classification of LiDAR bare-earth points, buildings, vegetation, and small objects based on region growing and angular classifier. Int. J. Remote Sens. 35 (19), 6955–6972.
Sengupta, S., Sturgess, P., 2015. Semantic octree: Unifying recognition, reconstruction and representation via an octree constrained higher order MRF. In: Proc. 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, USA, pp. 1874–1879.
Vanegas, C.A., Aliaga, D.G., Wonka, P., Müller, P., Waddell, P., Watson, B., 2010. Modelling the appearance and behaviour of urban spaces. Comput. Graph. Forum 29 (1), 25–42.
Verdie, Y., Lafarge, F., Alliez, P., 2015. LOD Generation for urban scenes. ACM Trans. Graph. (TOG) 34 (3), 15.
Vo, A., Truong-Hong, L., Laefer, D.F., Bertolotto, M., 2015. Octree-based region growing for point cloud segmentation. ISPRS J. Photogramm. Remote Sens. 104, 88–100.
Vu, H., Labatut, P., Pons, J., Keriven, R., 2012. High accuracy and visibility-consistent dense multi-view stereo. IEEE Trans. Pattern Anal. Mach. Intell. 34 (5), 889–901.
Weinmann, M., Jutzi, B., Hinz, S., Mallet, C., 2015. Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers. ISPRS J. Photogramm. Remote Sens. 105, 286–304.
Wolf, D., Prankl, J., Vincze, M., 2015. Fast semantic segmentation of 3D point clouds using a dense CRF with learned parameters. In: Proc. IEEE International Conference on Robotics and Automation (ICRA), IEEE, Washington, USA.
Xiong, B., Jancosek, M., Elberink, S.O., Vosselman, G., 2015. Flexible building primitives for 3D building modeling. ISPRS J. Photogramm. Remote Sens. 101, 275–290.
Xu, S., Vosselman, G., Elberink, S.O., 2014. Multiple-entity based classification of airborne laser scanning data in urban areas. ISPRS J. Photogramm. Remote Sens. 88, 1–15.
Zhang, J.X., Lin, X.G., 2012. Object-based classification of urban airborne LiDAR point clouds with multiple echoes using SVM. ISPRS Ann. Photogramm., Remote Sens. Spatial Inform. Sci. I 3, 135–140.
Zhang, Z., Zhang, L., Tong, X., Mathiopoulos, P.T., Guo, B., Huang, X., Wang, Z., Wang, Y., 2016. A multilevel point-cluster-based discriminative feature for ALS point cloud classification. IEEE Trans. Geosci. Remote Sens. 54 (6), 3309–3321.
Zhou, Q., Neumann, U., 2010. 2.5D dual contouring: A robust approach to creating building models from aerial LiDAR point clouds. In: Proc. European Conference on Computer Vision, Springer, Crete, Greece.
Zhou, Y., Yu, Y., Lu, G., Du, S., 2012. Super-segments based classification of 3D urban street scenes. Int. J. Adv. Rob. Syst. 9, 1–8.