Higher-order CRF Structural Segmentation of 3D Reconstructed Surfaces

Jingbo Liu Jinglu Wang Tian Fang Chiew-Lan Tai Long Quan
The Hong Kong University of Science and Technology
{jingbo, jwangae, tianft, taicl, quan}@ust.hk

Abstract

In this paper, we propose a structural segmentation algorithm to partition multi-view stereo reconstructed surfaces of large-scale urban environments into structural segments. Each segment corresponds to a structural component describable by a surface primitive of up to the second order. This segmentation is for use in subsequent urban object modeling, vectorization, and recognition.

To overcome the high geometrical and topological noise levels in the 3D reconstructed urban surfaces, we formulate the structural segmentation as a higher-order Conditional Random Field (CRF) labeling problem. It not only incorporates classical lower-order 2D and 3D local cues, but also encodes contextual geometric regularities to disambiguate the noisy local cues. A general higher-order CRF is difficult to solve. We develop a bottom-up progressive approach through a patch-based surface representation, which iteratively evolves from the initial mesh triangles to the final segmentation. Each iteration alternates between performing a prior discovery step, which finds the contextual regularities of the patch-based representation, and an inference step that leverages the regularities as higher-order priors to construct a more stable and regular segmentation.

The efficiency and robustness of the proposed method is extensively demonstrated on real reconstruction models, yielding significantly better performance than classical mesh segmentation methods.

1. Introduction

Modern multi-view stereo (MVS) algorithms are capable of reconstructing a large-scale 3D urban surface with unprecedented scalability and accuracy [1, 6, 7, 16, 27]. Segmenting the reconstructed surfaces is fundamental to higher level operations like object modeling and understanding a scene. Different from general clean meshes, the reconstructed meshes contain high geometrical and topological noises due to the imprecision of the reconstruction process. Besides, the reconstructed surface is under-sampled due to the insufficient image resolution or occlusion. Therefore, the segmentation problem is challenging despite the existence of a vast body of mesh segmentation literature. In this paper, we investigate the problem of a precise structural partition of a reconstructed urban surface. Each segment is intended to possess homogeneous textures and can be described by a second order surface patch. Often, such a segment corresponds to a structural component, such as a façade, roof, dome or column, as shown in Figure 1.

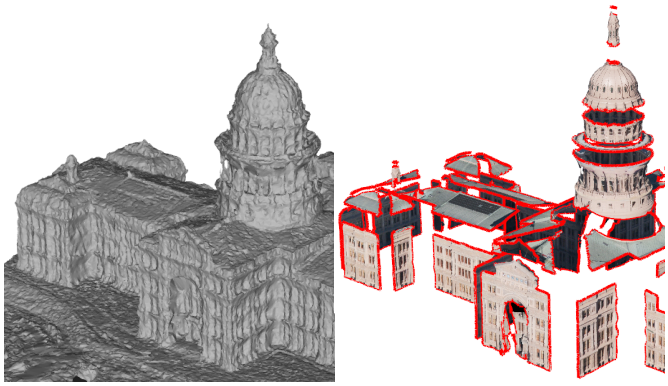


Figure 1: An example of the structural segmentation decomposing the reconstructed surface (left) into components, such as roofs, façades, domes, spires, and chimneys and so on (right).

Related work. A significant number of works have been devoted to partitioning urban scenes into structural segments. The input can be 2D images [25], 3D points from LiDAR [20], or 2D and 3D jointly [9, 16, 19]. The point clouds used in these approaches are unstructured thus these approaches are less aware of the topology. The approaches in [9, 16] are closest to ours since we all focus on the MVS reconstructed surface.

While mesh segmentation has rich literature, existing methods handle relatively clean input meshes. General purpose methods such as [10, 21] are able to produce semantically meaningful segmentation, however do not guarantee man-made characteristics like piecewise-qudraic, symmetry and the like. Our approach is highly related to the

primitive-fitting based mesh segmentation methods, which can be roughly divided into two approaches: the fast greedy approaches [2, 18], and the accurate but time-consuming variational approaches [3, 28]. In Section 5, our evaluation shows that the performance of these approaches declines with the existence of reconstruction noises.

CRFs, which encode powerful probabilistic formulations, are ubiquitously applied to the segmentation problem. Pairwise CRFs which only model local interactions between variables can produce good segmentation results, but may fail at boundaries, especially in the presence of noise, incompleteness and ambiguities. Therefore the use of higher-order potentials is motivated in a general manner, to encode priors like label consistency [11, 12], co-occurrence statistics [14, 15] and general pattern-based potential [13].

Different from the motivation of the above approaches, we target a structural segmentation of reconstructed meshes in urban environments, which possesses strong urban priors, such as piecewise-planarity [8], the Manhattan-world assumption [5, 26], inter-element relations [16, 31], and co-occurrence patterns [30]. Inspired by the pattern-based potentials of CRFs [13], we encode the urban regularity priors in a higher-order potential in the form of the P^n Potts model, which is later demonstrated to significantly improve the segmentation accuracy.

Our contributions are as follows:

- A new higher-order CRF formulation for the 2D-3D joint segmentation.
- An implicit patch-based surface representation (Figure 2 bottom) that reduces the complexity, which makes the higher-order CRF solvable in an efficient manner. It embeds 2D and 3D information, and dynamically evolves from an input triangle mesh to structural components as the algorithm proceeds.
- An efficient algorithm involving intermediate graph-matching to impose higher-order shape priors, that tolerates defects and produces results of high regularity.

2. Higher-order CRF Formulation

We start from a triangular mesh $\mathcal{S}$, reconstructed from multi-view images $\mathcal{I}$, using large-scale MVS methodologies (*e.g.*, [7, 17, 27]). Our target is to decompose the surface into structural components like façades, roofs, domes, and the like, as shown in Figure 1, 6. Every component possesses homogeneous textures, and fits a simple geometric primitive (a quadratic surface in our method). We call such decomposition a **structural segmentation** as it reveals the architectural structure.

The surface $\mathcal{S}$ is represented by a set of disjoint patches called surface units $\{x_0, x_1, ..., x_n\}$. The structural segmentation is formulated as a labeling problem: surface units that belong to the same structural component are assigned with the same label. Unfortunately, the existence of noises severely degrades the local properties of the surface units, making them unable to recover the correct label. The neighboring region, as well as the global context, needs to be taken into consideration. This motivates us to adopt the CRF framework, a structured probabilistic model, to predict the most probable label.

Data: mesh surface $\mathcal{S}$, images $\mathcal{I}$, threshold ϵ.
Result: structural segmentation represented by a graph G (Each node $x \in G$ represents a disjoint surface patch, and y is the label of x that represents a quadric.).
Initialize $G^{(0)}$ as the dual graph of $\mathcal{S}$;
repeat
  Ensure local photo-geometry consistency using lower order potentials ϕ_i and ψ_{ij} (Eqn 2, 3);
  Find regular patterns $\{R\}$ in $G^{(t)}$ (Eqn 6);
  foreach R **do**
    Identify every subgraph $G_c \subseteq G^{(t)}$ that closely matches R;
    Conform G_c to R by imposing higher order potential ψ_R (Eqn 7);
  end
  Infer the labeling $\mathbf{y}^{(t)}$ by minimizing $E^{(t)}$ (Eqn 1);
  Reduce $G^{(t)}$ to $G^{(t+1)}$ by merging nodes with the same label, and volume discrepancy is less than ϵ (as shown in Figure 2);
  Update the label set $\mathcal{L}_y$ by fitting new nodes $\{x^{(t+1)}\}$ with quadrics;
until $G^{(t+1)} = G^{(t)}$;

Algorithm 1: Overview of the iterative algorithm

The goal of the CRF labeling is to assign every surface unit x_i with a labeling y_i, which indicates the most probable fitted quadric. We use a tuple (p, η) to describe a quadric, where p is a parameterized *quadratic primitive* (the shape attribute), and η the pose parameter (position and orientation) of the quadratic primitive. The label set is composed of a set of such quadrics $\mathcal{L}_y = \{(p_1, \eta_1), (p_2, \eta_2), ..., (p_M, \eta_N)\}$. These quadratic primitives p_i are constructed by quantifying the geometrically best fitted quadrics of all surface units, using the DBSCAN [4] method.

Our approach progressively aggregates small surface units into a larger one when they belong to the same structural component. Initially, each surface unit represents a mesh triangle; at every iteration, neighboring surface units that fit identical quadrics are merged; eventually, each surface unit is described by a unique quadric.

At t-th iteration, the CRF is defined on a patch-based surface representation $G^{(t)} = (\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \mathbf{a}^{(t)}, \mathbf{b}^{(t)}, \mathcal{C}^{(t)})$ (*e.g.*,

Figure 2 bottom), where $\mathbf{x}^{(t)} = \{x_i^{(t)}\}$ is the set of nodes representing the surface units, $\mathbf{y}^{(t)} = \{y_i^{(t)}\}$ is the labeling of $\mathbf{x}^{(t)}$, $\mathbf{a}^{(t)} \subseteq \mathbf{x}^{(t)} \times \mathbf{x}^{(t)}$ is the set of edges representing the 3D demarcation line intersected by neighboring surface units, $\mathbf{b}^{(t)}$ is the labeling of $\mathbf{a}^{(t)}$ ($\mathbf{b}^{(t)}$ is an auxiliary variable to find regularities, detailed later in Section 4), and $\mathcal{C}^{(t)}$ is the set of *cliques*. A *clique* c, which may also be viewed as a hyper-edge of $G^{(t)}$, contains a set of surface units that are conditionally dependent on each other.

The CRF energy function takes the following form (the superscript $^{(t)}$ is omitted hereafter when we discuss each individual iteration):

$$E = \sum_{x_i \in \mathcal{S}} \phi_i(y_i) + \sum_{x_j \in \mathcal{N}(x_i)} \psi_{ij}(y_i, y_j) + \sum_{c \in \mathcal{C}} \psi_R(\mathbf{y}_c) \quad (1)$$

where the unary potential ϕ_i measures the photometric coherence of x_i to the input multi-view images $\mathcal{I}$, the pairwise potential ψ_{ij} encodes the 2D-3D joint domain discontinuity constraints at the intersection of x_i and x_j, and the higher-order potential ψ_R imposes regularity priors to enforce a regular and clean partition.The potentials are further elaborated on in the following sections. Parameters involved in the potentials are learned by cross validation. As the penalty function is semi-metric, $\alpha\beta$-swap is used for inference.

We assign the optimal labels to $\mathbf{y}$ by minimizing the energy function. When neighboring units are assigned to the same label (effectively, the same quadric), they are evaluated for merging by the volume discrepancy measurement $\int_{x_i \cup x_j} ||\hat{\mathbf{n}}||_2 \, \mathrm{d}S$, $||\hat{\mathbf{n}}||_2$ denotes the length of the difference vector $\hat{\mathbf{n}}$ to the fitted quadric along the normal direction. Taking plane, cylinder and sphere, the three most common primitives in a man-made scene as examples: to fit a plane, $\hat{\mathbf{n}} = \mathbf{n} \cdot (\overline{S} - \overline{x})$; for cylinders, $\hat{\mathbf{n}} = \|\overline{S} - \overline{x}\|_2 - r$; for spheres, $\hat{\mathbf{n}} = \|(\overline{S} - \overline{x}) \times \mathbf{n}\|_2 - r$.

After merging, a new set of surface units $\mathbf{x}^{(t+1)}$ is consolidated for the next iteration, and $\mathbf{y}^{(t)}$ is used as the initial labeling of $\mathbf{x}^{(t+1)}$, as illustrated in Algorithm 1. The patch-based surface representation is simplified after each iteration, as shown in Figure 2. Our segmentation approach is equivalent to a progressive graph reduction. As the graph reduces, the labeling problem complexity decreases as well.

3. Lower-order Potentials

The first two lower-order potentials of the energy function, which incorporate local photometric and geometric properties, are defined in this section.

Unary Potential. Structural components are mostly simple by design, for construction purposes. Therefore a structural component can be well described by a quadratic surface. The unary potential measures how likely a surface unit x_i can be approximated by a quadric (p_m, η_n) (associated with the label y_i) in terms of photo-consistency energy [16]. This term computes the aggregated dissimilarity of the projections in all pairs of the input images with respect to the approximated quadrics.

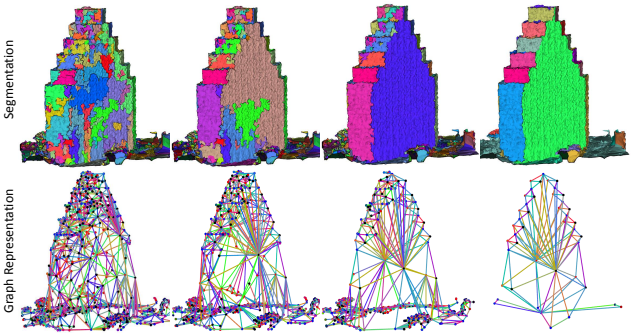


Figure 2: The progressive segmentation and graph reduction. Every surface partition (top) corresponds to a patch-based graphical representation (bottom). Since the segmentation process is constrained by regularities, the reduced graph becomes visually regular.

$$\phi_i(y_i) = \sum_{j,k} \int_{\Omega_{jk}^{q_i}} f(I_j, I_k)(s) \mathrm{d}s \quad (2)$$

where $f(I_j, I_k)(s)$ is implemented using the sum of normalized cross correlations (NCCs) for multiple sizes of the neighborhood of s [16], $\Omega_{jk}^{S_i}$ the intersected domain of projections in images I_j and I_k induced by the approximated surface q_i.

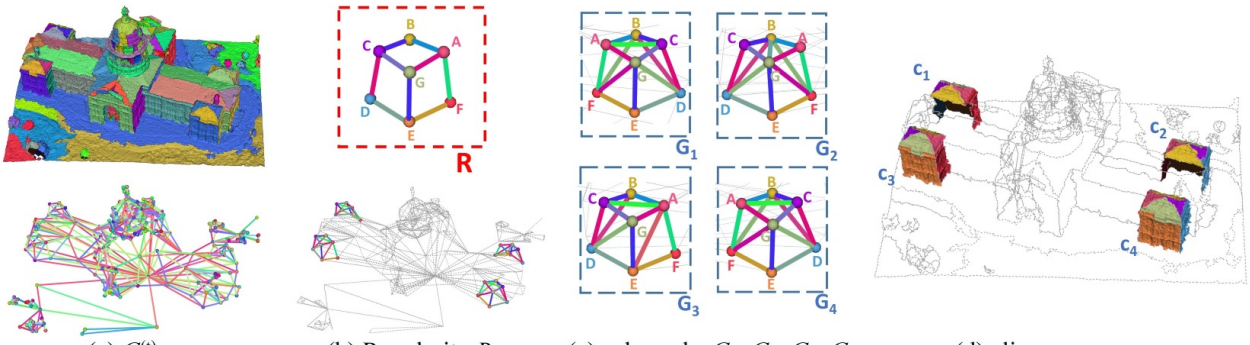
Pairwise Potential. Strong correspondences can be observed between photometric and geometric edges. Due to illumination change and silhouette, the boundary of surface units can be projected onto gradient domain edges in images. Each pair of adjacent surface unit x_i and x_j share a common boundary a_{ij}. The pairwise potential imposes the joint-domain smoothness within a surface unit x_i, while allowing strong bends at adjacency a_{ij} , and is given by:

$$\psi_{ij}(y_i, y_j) = \begin{cases} \theta_e(1 - e^{-f_e(a_{ij})}) & y_i = y_j \\ \theta_{e'} e^{-f_e(a_{ij})} & y_i \neq y_j \end{cases} \quad (3)$$

where $f_e(a_{ij})$ measures the edge coherence of both 2D-3D feature edges w.r.t. the adjacency boundary a_{ij}, θ_e and $\theta_{e'}$ the balancing parameters.

$$f_e(a) = \int_a (\sum_{I \in \mathcal{I}} f_{2d}(a(t), I) + \beta f_{3d}(a(t))) \, \mathrm{d}t \quad (4)$$

The 2D edge measurement f_{2d} is defined similarly to the directional edge filter proposed in [5]: $f_{2d}(\vec{a}, I) = \Omega(\vec{a}) \, \nabla_{\mathbf{d}} I(s) \, \mathrm{d}s \, / \int_{\Omega(\vec{a})} \nabla_{\mathbf{d}^\perp} I(s) \, \mathrm{d}s$, where $\Omega(\vec{a}) = \Pi_m \circ \vec{a}$



(a) $G^{(t)}$ (b) Regularity R (c) subgraphs G_1, G_2, G_3, G_4 (d) cliques c_1, c_2, c_3, c_4

Figure 3: (a) Top: a surface partition at t-th iteration. Bottom: the corresponding patch-based surface representation $G^{(t)}$. (b) R is a *regularity* that occurs four times in $G^{(t)}$. Nodes and edges are color coded by the labeling. (c) The the labeling of the four subgraphs which contain R is *regular*. (d) The four cliques are encouraged to preserve their labeling. Translational, rotational (*e.g.*, c_2 and c_3) and reflective (*e.g.*, c_3 and c_4) symmetry are detected.

is the 2D projection of an infinitesimal 3D edge $\vec{a}$ onto image I, $\mathbf{d}$ the normalized direction vector of the corresponding 2D edge, and $\mathbf{d}^{\perp}$ the direction vector perpendicular to $\mathbf{d}$. The 3D edge term f_{3d} measures the ridges and valleys value [22] in the direction of a.

4. Higher-order Regularity Potential

Reconstructed meshes inherently contain geometrical and topological noises (*e.g.*, distortions, holes, and genus are ubiquitous in Figure 1, 6), which affect the reliability of local geometric properties. An urban assumption, that man-made objects of urban scenes are usually regular by design, is utilized to disambiguate those noisy local properties. Our key observation here is that the repetition of structural patterns is a distinctive urban regularity. For instance, the parallelism is a repetition of surface orientation or edge direction; symmetry is a repetition of geometric shape; and architecture style is a repetition of the inter-relationship among a combination of structural elements. Close to the idea in [31], through favoring repetition, and suppressing the minor details and noises, we are able to obtain a more regular surface partition. The labeling of a clique (a set of surface units that are conditionally dependent on each other) is regarded *regular* when it conforms to a frequent labeling pattern, and the pattern is called a *regularity*.

We first discuss the finding of the *regularities* using subgraph isomorphism, and then we propose a higher-order potential to encode the *regularities* constraints as priors, which eventually improves the segmentation quality. Our proposed regularity-constrained potential encourages every clique to take a sufficiently frequent labeling pattern.

$$\psi_R(\mathbf{y}_c) = \begin{cases} 0 & \text{if } y_c \text{ is } \textit{regular} \\ \psi_c(\mathbf{y}_c) & \text{otherwise} \end{cases}$$

Contextual Regularities. Different from a restrictive approach of using parallelism and orthogonality [5, 16, 26], or further heuristically enumerating more pairwise regularities [31], we define *regularity* intrinsically from the input. For example, Figure 3 and 4 show *regularities* discovered from a palace and a high-rise building. A *regularity* is a sufficiently frequent graphical pattern R =($\mathbf{x}_R$,$\mathbf{y}_R$,$\mathbf{a}_R$,$\mathbf{b}_R$) in the patch-based surface representation. This definition characterizes the inter surface units relations arising from human design and construction. Thanks to the intrinsic nature of the *regularity*, our approach is adaptive to a large variety of input.

The recurrence of a graphical pattern is defined by a graph theory concept, *subgraph isomorphism*, which is a structure-preserving bijection between two labeled graphs. Both nodes' labeling $\mathbf{y}$ and edges' labeling $\mathbf{b}$ contribute to the isomorphism. Similar to the node label set $\mathcal{L}_y$, the edge label set is comprised of tuples $\mathcal{L}_b = \{(q, \theta)\}$, where $\{q\}$ is constructed by quantifying $\mathbf{a}$ as parametric conics, $\{\theta\}$ by quantifying the dihedral angles between adjacent surface units' orientations, *i.e.*, $\theta_{ij} = \arccos(n_i \cdot n_j)$. The edge labeling $\mathbf{b}$ implicitly encodes pairwise geometric relations between surface units. Even better, $\mathbf{a}$ and $\mathbf{b}$ are translational, rotational and reflective invariant. Thus, partial symmetry of the surface $\mathcal{S}$ can be detected through a tailored isomorphism [1] (*e.g.*, Figure 3(d)).

The concept of *regularity* can be formally defined with an exact subgraph isomorphism. Let G_c denote the vertex-

[1] Given two graphs g, g', the tailored isomorphism "$\simeq$" is a bijective function $f: \mathbf{x} \mapsto \mathbf{x}'$ satisfies,

$$\forall i, \quad l_b(x_i).p = l_b(f(x_i)).p, \quad \textit{and} \tag{5a}$$

$$\forall a_{ij} \in \mathbf{a}, \quad (f(x_i), f(x_j)) \in \mathbf{a}' \quad \textit{and} \quad l_b(a_{ij}) = l_b((f(x_i), f(x_j))). \tag{5b}$$

where $l_y: x \mapsto \mathcal{L}_y$ and $l_a: a \mapsto \mathcal{L}_b$ are the label functions, $l_b(x).p$ is the quadratic primitive.

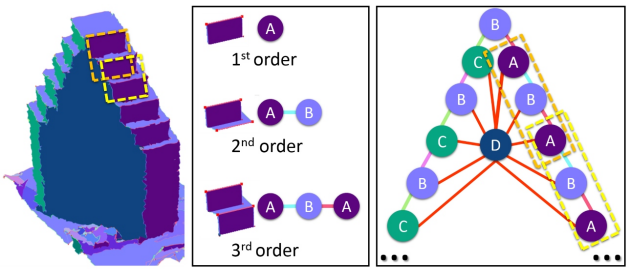


Figure 4: Left: a surface partition. Middle: examples of 1st, 2nd, 3rd order *regularities*. Right: an illustrative graph of the patch-based surface representation.

induced subgraph formed by the clique c (*e.g.*, in Figure 3, G_i and c_i), and the set of all vertex-induced subgraphs $\mathcal{G}$=$\{G_c\}$. Given a minimum repetition frequency threshold, $minSup$, R satisfies,

$$\varsigma(R, G_i) = \begin{cases} 1 & \exists g \simeq R,\ VIS(g) = G_i \\ 0 & \text{otherwise} \end{cases} \tag{6a}$$

$$\sigma(R, G) = \sum_{G_i \in \mathcal{G}} \varsigma(R, G_i) \geq minSup \tag{6b}$$

$\sigma(R, G)$ denotes the occurrence frequency of R in G, $\simeq$ the tailored isomorphism, and $VIS(g)$ the vertex-induced subgraph induced from g.

The *regularities* are found by using an exact subgraph isomorphism based method, *gSpan* [29]. Every subgraph $G_i \in \mathcal{G}$ is assigned a canonical label (identical canonical labels lead to isomorphism), which is a minimum depth-first search (DFS) code in this case. Through constructing a DFS code tree based on the label set $\mathcal{L}_y$ and $\mathcal{L}_b$, all k-length frequent minimum DFS codes, or equivalently, *regularities* consisting of k nodes, can be found.

Regularity-constrained potential. The higher-order potential encodes the *regularities* we just discovered. These *regularities* are far more expressive priors than local geometric properties, on which conventional mesh segmentation approaches rely. Meanwhile, higher-order potentials allow multiple interactions among surface units to be captured. The inter-relations between the surface units can disambiguate the unreliable local cues, and lead to a far more accurate partition.

A *regularity* containing $|\mathbf{c}|$ nodes is a *regularity of order* $|c|$, illustrated in Figure 4. The subgraph G_c, which is vertex-induced from the clique c, is matched against all *regularities of order* $|c|$, to find a mapping to a *regular* labeling $\mathcal{R}\colon (\mathbf{x}_c) \mapsto \mathcal{L}_y^{|c|}$.

The straight forwards case is that G_c contains a *regularity* R, meaning the clique's initial labeling is identical to $\mathbf{y}_R$. $\mathbf{y}_c$ is already *regular* and is encouraged to be preserved, *i.e.*, $\mathcal{R}(\mathbf{x}_c) = \mathbf{y}_c$. Otherwise, we use a volume discrepancy measurement to define an inexact match [23] between G_c and the closest *regularity* R. In that case, $\mathcal{R}$ maps every G_c's node to a label so that the labeling conforms to R, $\mathcal{R}(\mathbf{x}_c) = \mathbf{y}_R$.

The regularity potential takes the form of the P^n Potts model [11]:

$$\psi_c(\mathbf{y}_c) = \begin{cases} 0 & \text{if } \mathbf{y}_c = \mathcal{R}(\mathbf{x}_c) \\ \theta_p^h |c|^{\theta_\alpha} & \text{otherwise} \end{cases} \tag{7}$$

where $|c| = \sum_{i \in c} (1_{y_i \neq \mathcal{R}(x_i)})$ measures the number of mismatches between y_c and the labeling of the closest *regularity*, θ_p^h and θ_α are the parameters of the P^n Potts model.

Regularities essentially act as exemplars for frequently occurring structures at the clique level. Once we have recovered the *regularities*, we can use them to perform inexact matching to find other imperfect instances, based on the fact that: a few damaged instances of one *regularity* might become difficult to recognize due to defects; meanwhile, the other instances are still precise therefore can be used to discover that *regularity*. Therefore, we first compute *regularities* through exact matching, which can be effectively solved by exact graph isomorphism. Then we use these *regularities* to inexactly match them against all subgraphs in $\mathcal{G}$. These regularity priors enables our method to cope with the noises. By utilizing this redundancy nature of the urban scene, our approach tolerates data imperfections. It also outperforms previous methods when dealing with ambiguities, which are inevitable in stereo reconstructed meshes.

5. Implementation and Experiments

We implemented the proposed algorithm in C++ and run it on a PC equipped with 3.10GHz Intel Dual Core i5 CPU with 16 GB RAM.

Complexity Reduction. Enumerating all subgraphs of G to find the regularities triggers an exponential explosion, which becomes the algorithm's bottleneck. However, most such subgraphs are weakly connected and present no regularities. We only consider strongly connected disk-like subgraphs, like a sliding window on graph G. A breath-first search starting from $x_i \in G$ is performed. The first $|c|$ visited nodes forms a disk-like tree centered at x_i. The nodes form a clique c, and a vertex-induced subgraph G_c. The enumeration complexity is reduced to $\mathcal{O}(|c||\mathbf{x}^{(t)}|)$, meanwhile, nearly all regular patterns are preserved.

The complexity of each iteration is proportional to the size of G and the order of *regularities*. To balance the complexity between G and R, we search up to t-th order regularities at the t-th iteration for efficiency. At the early iterations, G has a huge size and is comprised of over-segmented surface units which are unlikely to represent complex regularities. Thus, we only search the local (lower-order) regularities. Later as G is significantly reduced, the segmen-

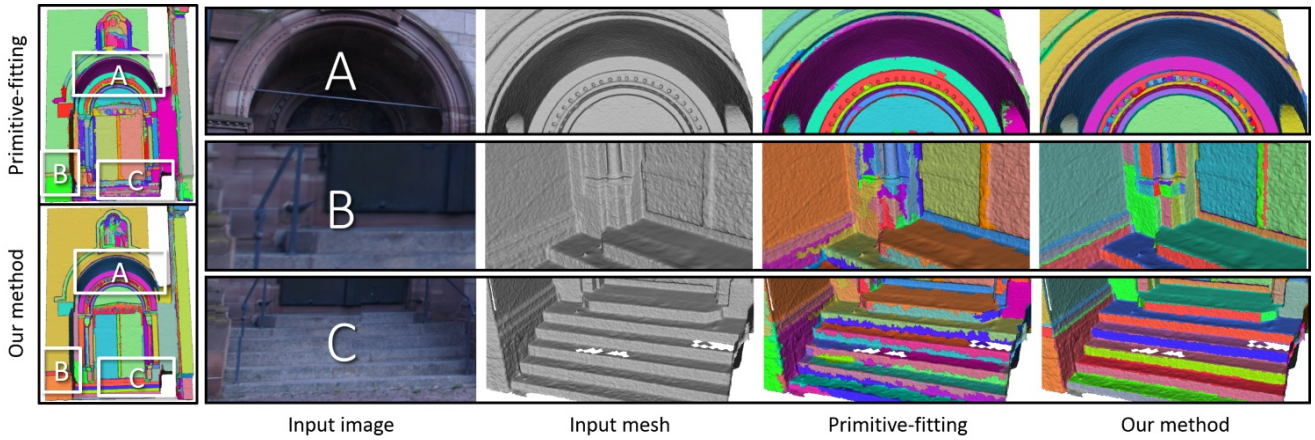


Figure 5: Comparison with [2] on a public LiDAR dataset "*Herz-Jesu-P8*" [24] with close-ups shown in region **A**, **B**, and **C**. Due to noises and ambiguities, primitive-based methods disturbs around the boundaries. In contrast, our method generates a segmentation with clean and crisp border. Specifically, in **B**, our method successfully recovers tiny structures.

Table 1: Statistics of the experiments. #I: the number of images used for reconstruction; Area: the covered area of the reconstructed meshes in km^2; #TRI: the triangle number of the reconstructed meshes. #SEG: the number of final segments; #R: the number of discovered regularities at the final iteration; ours: the time spent for segmentation using method (in minutes); vsa: the time spent using [28]; prim: using [2]; crf: using a naive pairwise crf.

data	#I	Area	#TRI	#SEG	#R	ours	vsa	prim	crf
Capitol	68	0.03	203K	247	26	5	25	2	3
Hez	12	-	248K	320	8	3	53	2	2
Stadium	100	0.08	40K	286	12	4	31	1	3
Dualwing	32	0.01	50K	76	62	2	6	1	1
CityA	2,177	2.8	2,437K	7,408	247	85	-	21	78
CityB	2,092	8.2	3,669K	11,092	163	136	-	42	119

tation becomes abstractive, and we explore more sophisticated regularities for shape priors. This bottom-up approach shares same insight with the Apriori algorithm: any subgraph of a high order regularity must be a regularity as well. Exploring only low order regularity in the early iterations will not cause us to lose any high order ones in later iterations. However, it saves us tremendous computation time. As described in Section 4, the *regularities* will be retained and aggregated through the iterations. This property enables our approach to find regular features at multiple scales.

Datasets. We performed both empirical and quantitative evaluation on six datasets: a public dataset "*Herz-Jesu-P8*" [24], and challenging real world reconstruction datasets include three buildings "*Capitol*", "*Stadium*", "*Dualwing*" and two city-scale "*CityA*" and "*CityB*". The real world reconstruction contains high geometric and topological noises. As shown in Figure 6 "*CityB*" has a lot of topological holes which are usually very difficult to handle in general mesh segmentation methods. The same pipeline is used for all the real world datasets to get the rough meshes. We use [17] to estimate the poses of the input images, [7] to reconstruct the surfaces, and [27] to refine the reconstructed surfaces. Statistics of the datasets are listed in Table 1. The volume discrepancy threshold ϵ is set to 0.001 for all the datasets.

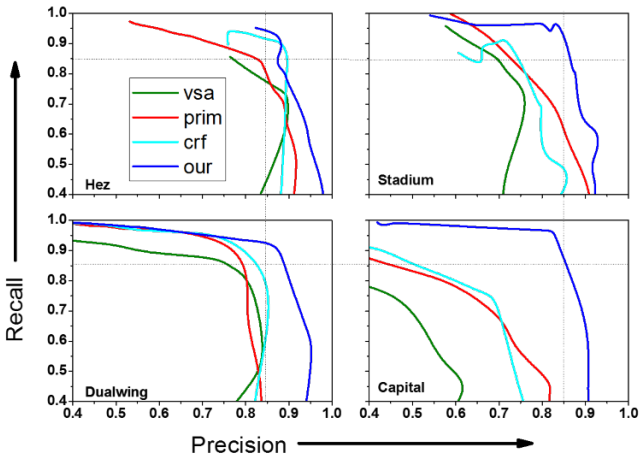


Figure 7: *Precision-Recall* evaluation of four methods: variational shape approximation (*vsa*), primitive-fitting based segmentation (*prim*), a naive joint segmentation using pairwise CRF (*crf*), and our method (*ours*). Lines further to the upper right represent better matches to the ground truth.

Performance comparison. We compare our results against two representative quadric-fitting based segmentation methods: a greedy approach (*prim*) [2], and a variational approach (*vsa*) [28]. Furthermore, we perform a con-

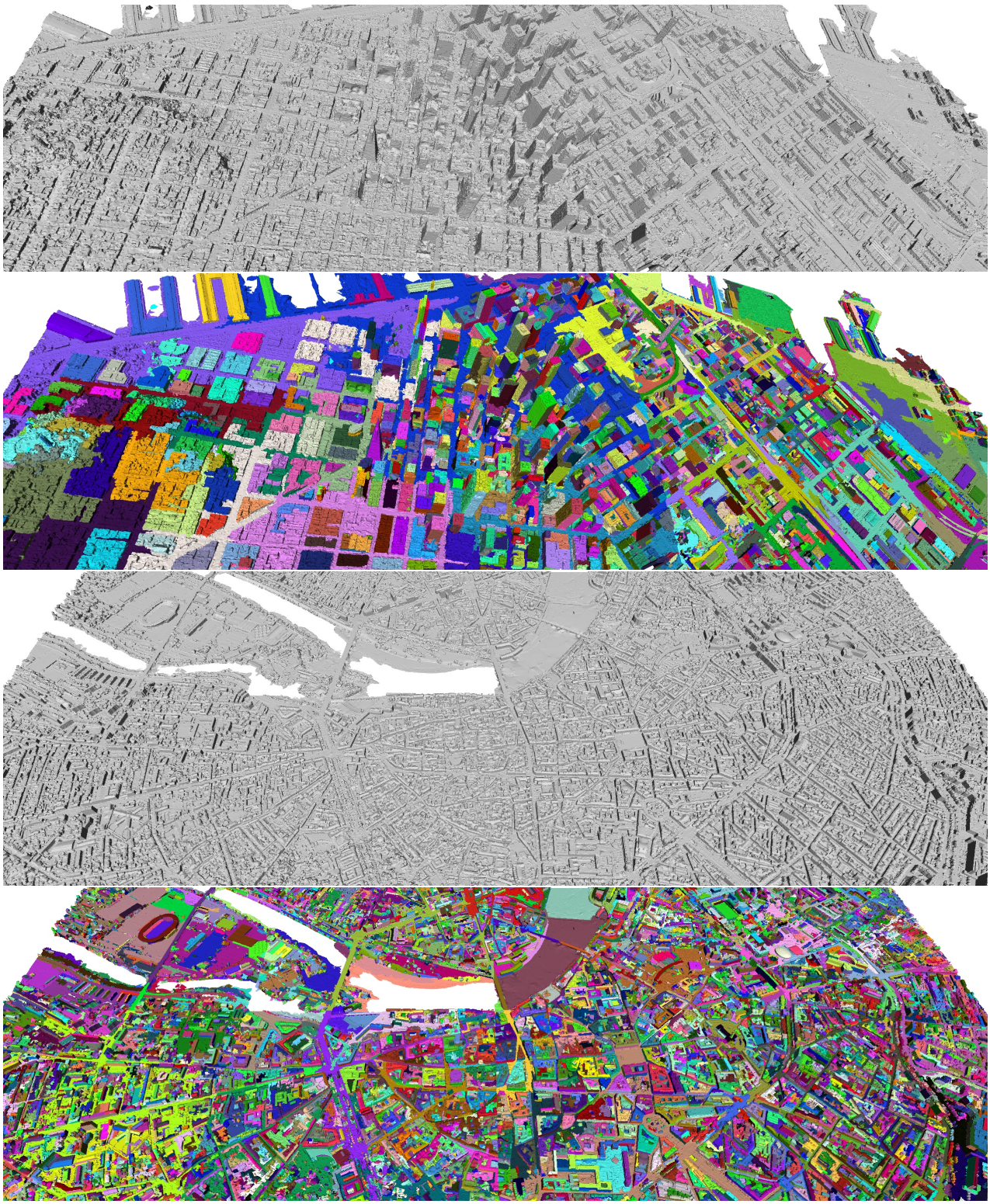


Figure 6: The reconstructed surface and segmentation of *CityA* downtown (top), *CityB* downtown (bottom). Our approach can be easily scaled up to handle city-scale input. (Please refer to the color print for better visualization.)

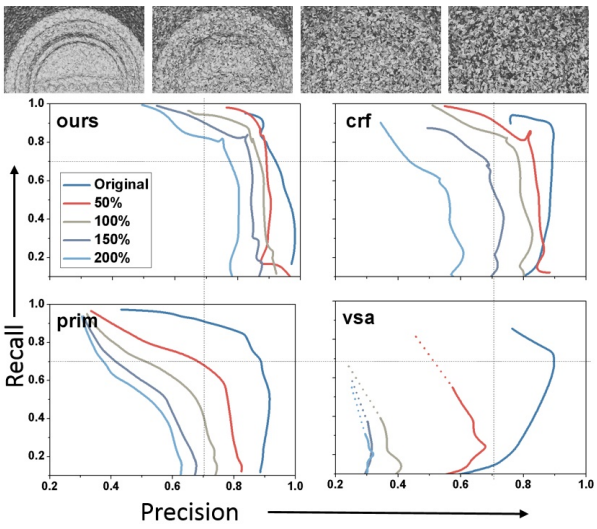


Figure 8: *Robustness to noises* evaluation of the methods on the original "Herz-Jesu" data, and four artificial noise levels (50/100/150/200% of the average edge length). A close view of the meshes is shown on the top. The *precision-recall* curves of our method are less influenced as the noise level increases.

trolled evaluation to see how the higher-order component affects our method. That is, we evaluate its performance if the higher-order potential is removed (*crf*). To the best of our knowledge, there is no publicly available benchmark for structural segmentation of urban scenes. We manually label four datasets as the ground truth. In order to quantitatively compare the results of the algorithms against the ground truth, we use the information retrieval statistics of *precision* and *recall*. Precision is defined as the fraction of segment boundaries in the algorithm's result that are near any segment boundaries of the ground truth. Recall is defined as the fraction of segment boundaries in the ground truth that are near any boundary of the algorithm's result. We define "near" by choosing a geodesic distance threshold (twice the average mesh edge is used in the evaluation).

Figure 7 shows the results of the precision-recall evaluation. The variational approach is very sensitive to noises [3, 28]. The *vsa* approach underperforms when compared to the other three in all four test sets. With an appropriate number of segments (middle of the precision-recall curve), *crf* outperforms *prim* thanks to the joint formulation. However, when the surface is extremely over/under-segmented (two ends of the precision-recall curve), the *prim* approach performs better. This may be due to the image gradients being computed at a fixed scale, in a range around the appropriate number of segments.

To evaluate the robustness of the four approaches towards noise, we apply the random vertex displacement noise of four levels (50/100/150/200% of the average edge length) to the "Herz-Jesu-P8" data, and measure four approaches' performances (Figure 8). The accuracy of all methods drops as the noise increases, while our methods is the least susceptible. It maintains a precision-recall accuracy of above 70%-70% at the 200% noise level. The *vsa* approach (bottom-right) is computationally expensive. Every *vsa* sample consumes more than six hours when the number of segments go beyond 800. Therefore we use the dashed lines to indicate the trend after 800 segments.

An example of an empirical evaluation is shown in Figure 5. The primitive-fitting based 3D segmentation method can recover several dominant structural primitives on *Herz-Jesu-P8*. Due to the lack of regularity awareness to enforce a clear segment boundary, the boundaries are usually perturbed. On the contrary, our method can successfully recover a more precise boundary. (For more results of empirical evaluation, please refer to the supplementary materials.)

Apart from the robustness in handling noise, our approach can easily be applied to a large-scale reconstruction surface, as shown in Figure 6. On the contrary, *vsa* has high time complexity, and *prim* has high space complexity. Thus it is difficult for them to adapt to large-scale data.

In summary, our approach has an obvious advantage in all test sets, showing the effectiveness of the urban regularity prior. The advantage of considering contextual regularities prevails in tackling ambiguities and defects.

6. Conclusion

We have proposed an approach to jointly segment reconstructed urban scenes from multi-view stereo. Structural segmentation not only takes the textural and structural information into account, but also unveils and investigates the relationship between structural segments to further regularize the results. One of our key innovations is that the contextual information of the urban scene structure is encoded in the higher-order potential of the CRF. Solving such an optimization gives us superior results over the state-of-the-art related segmentation approaches. Such topology-aware structural segmentation provides a powerful representation for manipulating and editing the unstructured reconstructed meshes. We believe our topology-constrained structural segmentation can be applied to improve and innovate a variety of research and applications.

Acknowledgment

This work was supported by RGC-GRF 16208614, 16209514, 619611, and ITC-PSKL12EG02. We appreciate the helpful comments from the anonymous reviewers. We would further like to thank F. Lafarge for his help in conducting the evaluation, and Shauna Dalton for the proofreading and editing.

References

[1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski. Building Rome in a day. *Communications of the ACM*, 54(10):105, Oct. 2011. 1

[2] M. Attene, B. Falcidieno, and M. Spagnuolo. Hierarchical mesh segmentation based on fitting primitives. *The Visual Computer*, 22(3):181–193, Feb. 2006. 2, 6

[3] D. Cohen-Steiner, P. Alliez, and M. Desbrun. Variational shape approximation. *ACM Transactions on Graphics*, 23(3):905, Aug. 2004. 2, 8

[4] M. Ester, H.-p. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *KDD*, 96:226–231, 1996. 2

[5] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski. Manhattan-world stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1422–1429. IEEE, June 2009. 2, 3, 4

[6] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Towards Internet-scale multi-view stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1434–1441. IEEE, June 2010. 1

[7] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–76, Aug. 2010. 1, 2, 6

[8] D. Gallup, J.-m. Frahm, and M. Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1418–1425. IEEE, June 2010. 2

[9] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3D Scene Reconstruction and Class Segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 97–104, June 2013. 1

[10] E. Kalogerakis, A. Hertzmann, and K. Singh. Learning 3D mesh segmentation and labeling. *ACM Transactions on Graphics*, 29(4):1, July 2010. 1

[11] P. Kohli, M. P. Kumar, and P. H. S. Torr. P3 & beyond: Solving energies with higher order cliques. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 2, 5

[12] P. Kohli, L. Ladický, and P. H. S. Torr. Robust Higher Order Potentials for Enforcing Label Consistency. *International Journal of Computer Vision*, 82(3):302–324, Jan. 2009. 2

[13] N. Komodakis and N. Paragios. Beyond pairwise energies: Efficient optimization for higher-order mrfs. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2985–2992, 2009. 2

[14] H. S. Koppula, A. Anand, T. Joachims, and A. Saxena. Semantic labeling of 3d point clouds for indoor scenes. In *Advances in Neural Information Processing Systems*, pages 244–252, 2011. 2

[15] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Graph cut based inference with co-occurrence statistics. In *European Conference on Computer Vision*, pages 239–253. Springer, 2010. 2

[16] F. Lafarge, R. Keriven, M. Brédif, and H.-h. Vu. A hybrid multiview stereo algorithm for modeling urban scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):5–17, Jan. 2013. 1, 2, 3, 4

[17] M. Lhuillier and L. Quan. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):418–33, Mar. 2005. 2, 6

[18] Y. Li, X. Wu, Y. Chrysathou, A. Sharf, D. Cohen-Or, and N. J. Mitra. GlobFit. *ACM Transactions on Graphics*, 30(4):1, 2011. 2

[19] Y. Li, Q. Zheng, A. Sharf, D. Cohen-Or, B. Chen, and N. J. Mitra. 2D-3D fusion for layer decomposition of urban facades. *International Conference on Computer Vision*, 1:882–889, Nov. 2011. 1

[20] H. Lin, J. Gao, Y. Zhou, G. Lu, M. Ye, C. Zhang, L. Liu, and R. Yang. Semantic decomposition and reconstruction of residential scenes from LiDAR data. *ACM Transactions on Graphics*, 32(4):1, July 2013. 1

[21] J. Lv, X. Chen, J. Huang, and H. Bao. Semi-supervised Mesh Segmentation and Labeling. *Computer Graphics Forum*, 31(7):2241–2248, Sept. 2012. 1

[22] Y. Ohtake, A. Belyaev, and H.-P. Seidel. Ridge-valley lines on meshes via implicit surface fitting. *ACM Transactions on Graphics*, 23(3):609, Aug. 2004. 4

[23] K. Riesen and H. Bunke. *Managing and Mining Graph Data*, volume 40 of *Advances in Database Systems*. Springer US, Boston, MA, 2010. 5

[24] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, June 2008. 6

[25] Z. Tu, X. Chen, A. L. Yuille, and S. C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of Computer Vision*, 63(2):113–140, 2005. 1

[26] C. A. Vanegas, D. G. Aliaga, and B. Benes. Building reconstruction using manhattan-world grammars. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 358–365. IEEE, June 2010. 2, 4

[27] H.-H. Vu, P. Labatut, J.-P. Pons, and R. Keriven. High accuracy and visibility-consistent dense multiview stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):889–901, May 2012. 1, 2, 6

[28] D.-M. Yan, W. Wang, Y. Liu, and Z. Yang. Variational mesh segmentation via quadric surface fitting. *Computer-Aided Design*, 44(11):1072–1082, Nov. 2012. 2, 6, 8

[29] X. Yan and J. Han. gSpan: graph-based substructure pattern mining. In *IEEE International Conference on Data Mining*, volume 1, pages 721–724. IEEE Comput. Soc, 2002. 5

[30] J. Yuan, M. Yang, and Y. Wu. Mining discriminative co-occurrence patterns for visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2777–2784, 2011. 2

[31] Q.-Y. Zhou and U. Neumann. 2.5D building modeling by discovering global regularities. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–333. IEEE, June 2012. 2, 4