



**HAL**  
open science

# Large-scale Point Cloud Semantic Segmentation with Superpoint Graphs

Loic Landrieu, Simonovsky Martin

► **To cite this version:**

Loic Landrieu, Simonovsky Martin. Large-scale Point Cloud Semantic Segmentation with Superpoint Graphs. 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Jun 2018, Salt Lake City, United States. hal-01801186

**HAL Id: hal-01801186**

**<https://hal.science/hal-01801186>**

Submitted on 28 May 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Large-scale Point Cloud Semantic Segmentation with Superpoint Graphs

Loic Landrieu<sup>1\*</sup>, Martin Simonovsky<sup>2\*</sup>

<sup>1</sup> Université Paris-Est, LASTIG MATIS IGN, ENSG

<sup>2</sup> Université Paris-Est, Ecole des Ponts ParisTech

loic.landrieu@ign.fr, martin.simonovsky@enpc.fr

## Abstract

*We propose a novel deep learning-based framework to tackle the challenge of semantic segmentation of large-scale point clouds of millions of points. We argue that the organization of 3D point clouds can be efficiently captured by a structure called superpoint graph (SPG), derived from a partition of the scanned scene into geometrically homogeneous elements. SPGs offer a compact yet rich representation of contextual relationships between object parts, which is then exploited by a graph convolutional network. Our framework sets a new state of the art for segmenting outdoor LiDAR scans (+11.9 and +8.8 mIoU points for both Semantic3D test sets), as well as indoor scans (+12.4 mIoU points for the S3DIS dataset).*

## 1. Introduction

Semantic segmentation of large 3D point clouds presents numerous challenges, the most obvious one being the scale of the data. Another hurdle is the lack of clear structure akin to the regular grid arrangement in images. These obstacles have likely prevented Convolutional Neural Networks (CNNs) from achieving on irregular data the impressive performances attained for speech processing or images.

Previous attempts at using deep learning for large 3D data were trying to replicate successful CNN architectures used for image segmentation. For example, SnapNet [5] converts a 3D point cloud into a set of virtual 2D RGBD snapshots, the semantic segmentation of which can then be projected on the original data. SegCloud [46] uses 3D convolutions on a regular voxel grid. However, we argue that such methods do not capture the inherent structure of 3D point clouds, which results in limited discrimination performance. Indeed, converting point clouds to 2D format comes with loss of information and requires to perform surface reconstruction, a problem arguably as hard as semantic segmentation. Volumetric representation of point clouds is

inefficient and tends to discard small details.

Deep learning architectures specifically designed for 3D point clouds [39, 45, 42, 40, 10] display good results, but are limited by the size of inputs they can handle at once.

We propose a representation of large 3D point clouds as a collection of interconnected simple shapes coined superpoints, in spirit similar to superpixel methods for image segmentation [1]. As illustrated in Figure 1, this structure can be captured by an attributed directed graph called the superpoint graph (SPG). Its nodes represent simple shapes while edges describe their adjacency relationship characterized by rich edge features.

The SPG representation has several compelling advantages. First, instead of classifying individual points or voxels, it considers entire object parts as whole, which are easier to identify. Second, it is able to describe in detail the relationship between adjacent objects, which is crucial for contextual classification: cars are generally above roads, ceilings are surrounded by walls, etc. Third, the size of the SPG is defined by the number of simple structures in a scene rather than the total number of points, which is typically several order of magnitude smaller. This allows us to model long-range interaction which would be intractable otherwise without strong assumptions on the nature of the pairwise connections. Our contributions are as follows:

- We introduce superpoint graphs, a novel point cloud representation with rich edge features encoding the contextual relationship between object parts in 3D point clouds.
- Based on this representation, we are able to apply deep learning on large-scale point clouds without major sacrifice in fine details. Our architecture consists of PointNets [39] for superpoint embedding and graph convolutions for contextual segmentation. For the latter, we introduce a novel, more efficient version of Edge-Conditioned Convolutions [45] as well as a new form of input gating in Gated Recurrent Units [8].
- We set a new state of the art on two publicly available datasets: Semantic3D [14] and S3DIS [3]. In particu-

\* Both authors contributed equally to this work.

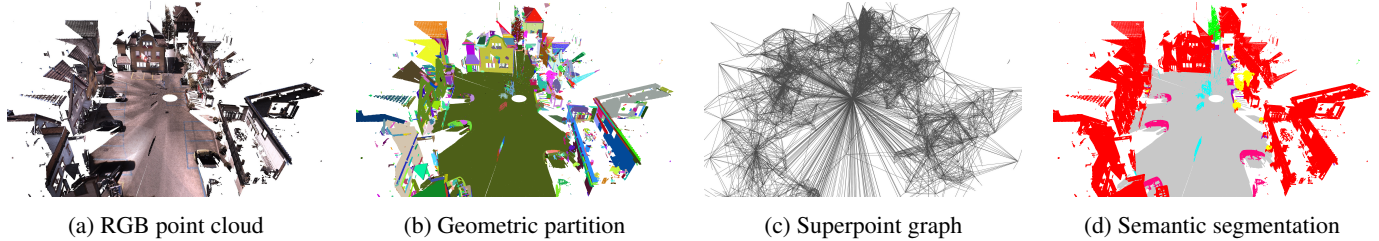


Figure 1: Visualization of individual steps in our pipeline. An input point cloud (a) is partitioned into geometrically simple shapes, called superpoints (b). Based on this preprocessing, a superpoints graph (SPG) is constructed by linking nearby superpoints by superedges with rich attributes (c). Finally, superpoints are transformed into compact embeddings, processed with graph convolutions to make use of contextual information, and classified into semantic labels.

lar, we improve mean per-class intersection over union (mIoU) by 11.9 points for the Semantic3D reduced test set, by 8.8 points for the Semantic3D full test set, and by up to 12.4 points for the S3DIS dataset.

## 2. Related Work

The classic approach to large-scale point cloud segmentation is to classify each point or voxel independently using handcrafted features derived from their local neighborhood [48]. The solution is then spatially regularized using graphical models [37, 24, 34, 44, 21, 2, 38, 35, 49] or structured optimization [27]. Clustering as preprocessing [16, 13] or postprocessing [47] have been used by several frameworks to improve the accuracy of the classification.

**Deep Learning on Point Clouds.** Several different approaches going beyond naive volumetric processing of point clouds have been proposed recently, notably set-based [39, 40], tree-based [42, 23], and graph-based [45]. However, very few methods with deep learning components have been demonstrated to be able to segment large-scale point clouds. PointNet [39] can segment large clouds with a sliding window approach, therefore constraining contextual information within a small area only. Engelmann *et al.* [10] improves on this by increasing the context scope with multi-scale windows or by considering directly neighboring window positions on a voxel grid. SEGCloud [46] handles large clouds by voxelizing followed by interpolation back to the original resolution and post-processing with a conditional random field (CRF). None of these approaches is able to consider fine details and long-range contextual information simultaneously. In contrast, our pipeline partitions point clouds in an adaptive way according to their geometric complexity and allows deep learning architecture to use both fine detail and interactions over long distance.

**Graph Convolutions.** A key step of our approach is using graph convolutions to spread contextual information. Formulations that are able to deal with graphs of variable sizes can be seen as a form of message passing over graph

edges [12]. Of particular interest are models supporting continuous edge attributes [45, 36], which we use to represent interactions. In image segmentation, convolutions on graphs built over superpixels have been used for post-processing: Liang *et al.* [32, 31] traverses such graphs in a sequential node order based on unary confidences to improve the final labels. We update graph nodes in parallel and exploit edge attributes for informative context modeling. Xu *et al.* [50] convolves information over graphs of object detections to infer their contextual relationships. Our work infers relationships implicitly to improve segmentation results. Qi *et al.* [41] also relies on graph convolutions on 3D point clouds. However, we process large point clouds instead of small RGBD images with nodes embedded in 3D instead of 2D in a novel, rich-attributed graph. Finally, we note that graph convolutions also bear functional similarity to deep learning formulations of CRFs [51], which we discuss more in Section 3.4.

## 3. Method

The main obstacle that our framework tries to overcome is the size of LiDAR scans. Indeed, they can reach hundreds of millions of points, making direct deep learning approaches intractable. The proposed SPG representation allows us to split the semantic segmentation problem into three distinct problems of different scales, shown in Figure 2, which can in turn be solved by methods of corresponding complexity:

- 1 **Geometrically homogeneous partition:** The first step of our algorithm is to partition the point cloud into geometrically simple yet meaningful shapes, called superpoints. This unsupervised step takes the whole point cloud as input, and therefore must be computationally very efficient. The SPG can be easily computed from this partition.
- 2 **Superpoint embedding:** Each node of the SPG corresponds to a small part of the point cloud correspond-

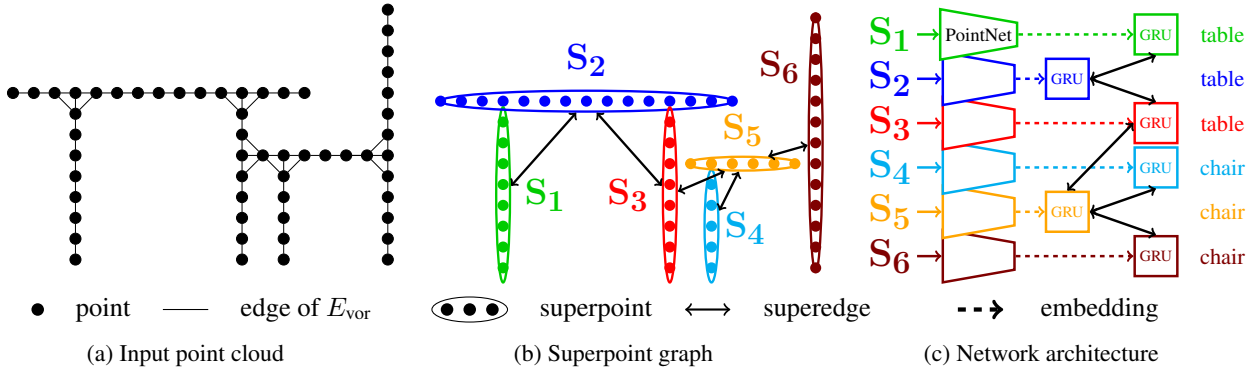


Figure 2: Illustration of our framework on a toy scan of a table and a chair. We perform geometric partitioning on the point cloud (a), which allows us to build the superpoint graph (b). Each superpoint is embedded by a PointNet network. The embeddings are then refined in GRUs by message passing along superedges to produce the final labeling (c).

ing to a geometrically simple primitive, which we assume to be semantically homogeneous. Such primitives can be reliably represented by downsampling small point clouds to at most hundreds of points. This small size allows us to utilize recent point cloud embedding methods such as PointNet [39].

**3 Contextual segmentation:** The graph of superpoints is by orders of magnitude smaller than any graph built on the original point cloud. Deep learning algorithms based on graph convolutions can then be used to classify its nodes using rich edge features facilitating long-range interactions.

The SPG representation allows us to perform end-to-end learning of the trainable two last steps. We will describe each step of our pipeline in the following subsections.

### 3.1. Geometric Partition with a Global Energy

In this subsection, we describe our method for partitioning the input point cloud into parts of simple shape. Our objective is not to retrieve individual objects such as cars or chairs, but rather to break down the objects into simple parts, as seen in Figure 3. However, the clusters being geometrically simple, one can expect them to be semantically homogeneous as well, *i.e.* not to cover objects of different classes. Note that this step of the pipeline is purely unsupervised and makes no use of class labels beyond validation.

We follow the global energy model described by [13] for its computational efficiency. Another advantage is that the segmentation is adaptive to the local geometric complexity. In other words, the segments obtained can be large simple shapes such as roads or walls, as well as much smaller components such as parts of a car or a chair.

Let us consider the input point cloud  $C$  as a set of  $n$  3D points. Each point  $i \in C$  is defined by its 3D position

$p_i$ , and, if available, other observations  $o_i$  such as color or intensity. For each point, we compute a set of  $d_g$  geometric features  $f_i \in \mathbb{R}^{d_g}$  characterizing the shape of its local neighborhood. In this paper, we use three dimensionality values proposed by [9]: linearity, planarity and scattering, as well as the verticality feature introduced by [13]. We also compute the elevation of each point, defined as the  $z$  coordinate of  $p_i$  normalized over the whole input cloud.

The global energy proposed by [13] is defined with respect to the 10-nearest neighbor *adjacency graph*  $G_{nn} = (C, E_{nn})$  of the point cloud (note that this is *not* the SPG). The geometrically homogeneous partition is defined as the constant connected components of the solution of the following optimization problem:

$$\arg \min_{g \in \mathbb{R}^{d_g}} \sum_{i \in C} \|g_i - f_i\|^2 + \mu \sum_{(i,j) \in E_{nn}} w_{i,j} [g_i - g_j \neq 0], \quad (1)$$

where  $[\cdot]$  is the Iverson bracket. The edge weight  $w \in \mathbb{R}_+^{|E|}$  is chosen to be linearly decreasing with respect to the edge length. The factor  $\mu$  is the regularization strength and determines the coarseness of the resulting partition.

The problem defined in Equation 1 is known as *generalized minimal partition problem*, and can be seen as a continuous-space version of the Potts energy model, or an  $\ell_0$  variant of the graph total variation. The minimized functional being nonconvex and noncontinuous implies that the problem cannot realistically be solved exactly for large point clouds. However, the  $\ell_0$ -cut pursuit algorithm introduced by [26] is able to quickly find an approximate solution with a few graph-cut iterations. In contrast to other optimization methods such as  $\alpha$ -expansion [6], the  $\ell_0$ -cut pursuit algorithm does not require selecting the size of the partition in advance. The constant connected components  $S = \{S_1, \dots, S_k\}$  of the solution of Equation 1 define our geometrically simple elements, and are referred as *super-*

| Feature name      | Size | Description   |
|-------------------|------|---|
| mean offset       | 3    | $\text{mean}_{m \in \delta(S,T)} \delta_m$              |
| offset deviation  | 3    | $\text{std}_{m \in \delta(S,T)} \delta_m$               |
| centroid offset   | 3    | $\text{mean}_{i \in S} p_i - \text{mean}_{j \in T} p_j$ |
| length ratio      | 1    | $\log \text{length}(S) / \text{length}(T)$              |
| surface ratio     | 1    | $\log \text{surface}(S) / \text{surface}(T)$            |
| volume ratio      | 1    | $\log \text{volume}(S) / \text{volume}(T)$              |
| point count ratio | 1    | $\log  S  /  T $  |

Table 1: List of  $d_f = 13$  superedge features characterizing the adjacency between two superpoints  $S$  and  $T$ .

points (i.e. set of points) in the rest of this paper.

### 3.2. Superpoint Graph Construction

In this subsection, we describe how we compute the SPG as well as its key features. The SPG is a structured representation of the point cloud, defined as an oriented attributed graph  $\mathcal{G} = (S, \mathcal{E}, F)$  whose nodes are the set of superpoints  $S$  and edges  $\mathcal{E}$  (referred to as *superedges*) represent the adjacency between superpoints. The superedges are annotated by a set of  $d_f$  features:  $F \in \mathbb{R}^{\mathcal{E} \times d_f}$  characterizing the adjacency relationship between superpoints.

We define  $G_{\text{vor}} = (C, E_{\text{vor}})$  as the symmetric Voronoi adjacency graph of the complete input point cloud as defined by [20]. Two superpoints  $S$  and  $T$  are adjacent if there is at least one edge in  $E_{\text{vor}}$  with one end in  $S$  and one end in  $T$ :

$$\mathcal{E} = \{(S, T) \in \mathcal{S}^2 \mid \exists (i, j) \in E_{\text{vor}} \cap (S \times T)\}. \quad (2)$$

Important spatial features associated with a superedge  $(S, T)$  are obtained from the set of offsets  $\delta(S, T)$  for edges in  $E_{\text{vor}}$  linking both superpoints:

$$\delta(S, T) = \{(p_i - p_j) \mid (i, j) \in E_{\text{vor}} \cap (S \times T)\}. \quad (3)$$

Superedge features can also be derived by comparing the shape and size of the adjacent superpoints. To this end, we compute  $|S|$  as the number of points comprised in a superpoint  $S$ , as well as shape features  $\text{length}(S) = \lambda_1$ ,  $\text{surface}(S) = \lambda_1 \lambda_2$ ,  $\text{volume}(S) = \lambda_1 \lambda_2 \lambda_3$  derived from the eigenvalues  $\lambda_1, \lambda_2, \lambda_3$  of the covariance of the positions of the points comprised in each superpoint, sorted by decreasing value. In Table 1, we describe a list of the different superedge features used in this paper. Note that the break of symmetry in the edge features makes the SPG a directed graph.

### 3.3. Superpoint Embedding

The goal of this stage is to compute a descriptor for every superpoint  $S_i$  by embedding it into a vector  $\mathbf{z}_i$  of fixed-size dimensionality  $d_z$ . Note that each superpoint is embedded

in isolation; contextual information required for its reliable classification is provided only in the following stage by the means of graph convolutions.

Several deep learning-based methods have been proposed for this purpose recently. We choose PointNet [39] for its remarkable simplicity, efficiency, and robustness. In PointNet, input points are first aligned by a Spatial Transformer Network [19], independently processed by multi-layer perceptrons (MLPs), and finally max-pooled to summarize the shape.

In our case, input shapes are geometrically simple objects, which can be reliably represented by a small amount of points and embedded by a rather compact PointNet. This is important to limit the memory needed when evaluating many superpoints on current GPUs. In particular, we subsample superpoints on-the-fly down to  $n_p = 128$  points to maintain efficient computation in batches and facilitate data augmentation. Superpoints of less than  $n_p$  points are sampled with replacement, which in principle does not affect the evaluation of PointNet due to its max-pooling. However, we observed that including very small superpoints of less than  $n_{\text{minp}} = 40$  points in training harms the overall performance. Thus, embedding of such superpoints is set to zero so that their classification relies solely on contextual information.

In order for PointNet to learn spatial distribution of different shapes, each superpoint is rescaled to unit sphere before embedding. Points are represented by their normalized position  $p'_i$ , observations  $o_i$ , and geometric features  $f_i$  (since these are already available precomputed from the partitioning step). Furthermore, the original metric diameter of the superpoint is concatenated as an additional feature after PointNet max-pooling in order to stay covariant with shape sizes.

### 3.4. Contextual Segmentation

The final stage of the pipeline is to classify each superpoint  $S_i$  based on its embedding  $\mathbf{z}_i$  and its local surroundings within the SPG. Graph convolutions are naturally suited to this task. In this section, we explain the propagation model of our system.

Our approach builds on the ideas from Gated Graph Neural Networks [30] and Edge-Conditioned Convolutions (ECC) [45]. The general idea is that superpoints refine their embedding according to pieces of information passed along superedges. Concretely, each superpoint  $S_i$  maintains its state hidden in a Gated Recurrent Unit (GRU) [8]. The hidden state is initialized with embedding  $\mathbf{z}_i$  and is then processed over several iterations (time steps)  $t = 1 \dots T$ . At each iteration  $t$ , a GRU takes its hidden state  $\mathbf{h}_i^{(t)}$  and an incoming message  $\mathbf{m}_i^{(t)}$  as input, and computes its new hidden state  $\mathbf{h}_i^{(t+1)}$ . The incoming message  $\mathbf{m}_i^{(t)}$  to superpoint

$i$  is computed as a weighted sum of hidden states  $\mathbf{h}_j^{(t)}$  of neighboring superpoints  $j$ . The actual weighting for a superedge  $(j, i)$  depends on its attributes  $F_{ji,\cdot}$ , listed in Table 1. In particular, it is computed from the attributes by a multi-layer perceptron  $\Theta$ , so-called Filter Generating Network. Formally:

$$\begin{aligned} \mathbf{h}_i^{(t+1)} &= (1 - \mathbf{u}_i^{(t)}) \odot \mathbf{q}_i^{(t)} + \mathbf{u}_i^{(t)} \odot \mathbf{h}_i^{(t)} \\ \mathbf{q}_i^{(t)} &= \tanh(\mathbf{x}_{1,i}^{(t)} + \mathbf{r}_i^{(t)} \odot \mathbf{h}_{1,i}^{(t)}) \\ \mathbf{u}_i^{(t)} &= \sigma(\mathbf{x}_{2,i}^{(t)} + \mathbf{h}_{2,i}^{(t)}), \quad \mathbf{r}_i^{(t)} = \sigma(\mathbf{x}_{3,i}^{(t)} + \mathbf{h}_{3,i}^{(t)}) \end{aligned} \quad (4)$$

$$\begin{aligned} (\mathbf{h}_{1,i}^{(t)}, \mathbf{h}_{2,i}^{(t)}, \mathbf{h}_{3,i}^{(t)})^T &= \rho(W_h \mathbf{h}_i^{(t)} + b_h) \\ (\mathbf{x}_{1,i}^{(t)}, \mathbf{x}_{2,i}^{(t)}, \mathbf{x}_{3,i}^{(t)})^T &= \rho(W_x \mathbf{x}_i^{(t)} + b_x) \end{aligned} \quad (5)$$

$$\mathbf{x}_i^{(t)} = \sigma(W_g \mathbf{h}_i^{(t)} + b_g) \odot \mathbf{m}_i^{(t)} \quad (6)$$

$$\mathbf{m}_i^{(t)} = \text{mean}_{j|(j,i) \in \mathcal{E}} \Theta(F_{ji,\cdot}; W_e) \odot \mathbf{h}_j^{(t)} \quad (7)$$

$$\mathbf{h}_i^{(1)} = \mathbf{z}_i, \quad \mathbf{y}_i = W_o(\mathbf{h}_i^{(1)}, \dots, \mathbf{h}_i^{(T+1)})^T, \quad (8)$$

where  $\odot$  is element-wise multiplication,  $\sigma(\cdot)$  sigmoid function, and  $W$  and  $b$  are trainable parameters shared among all GRUs. Equation 4 lists the standard GRU rules [8] with its update gate  $\mathbf{u}_i^{(t)}$  and reset gate  $\mathbf{r}_i^{(t)}$ . To improve stability during training, in Equation 5 we apply Layer Normalization [4] defined as  $\rho(\mathbf{a}) := (\mathbf{a} - \text{mean}(\mathbf{a})) / (\text{std}(\mathbf{a}) + \epsilon)$  separately to linearly transformed input  $\mathbf{x}_i^{(t)}$  and transformed hidden state  $\mathbf{h}_i^{(t)}$ , with  $\epsilon$  being a small constant. Finally, the model includes three interesting extensions in Equations 6–8, which we detail below.

**Input Gating.** We argue that GRU should possess the ability to down-weight (parts of) an input vector based on its hidden state. For example, GRU might learn to ignore its context if its class state is highly certain or to direct its attention to only specific feature channels. Equation 6 achieves this by gating message  $\mathbf{m}_i^{(t)}$  by the hidden state before using it as input  $\mathbf{x}_i^{(t)}$ .

**Edge-Conditioned Convolution.** ECC plays a crucial role in our model as it can dynamically generate filtering weights for any value of continuous attributes  $F_{ji,\cdot}$  by processing them with a multi-layer perceptron  $\Theta$ . In the original formulation [45] (ECC-MV),  $\Theta$  regresses a weight matrix to perform matrix-vector multiplication  $\Theta(F_{ji,\cdot}; W_e) \mathbf{h}_j^{(t)}$  for each edge. In this work, we propose a lightweight variant with lower memory requirements and fewer parameters, which is beneficial for datasets with few but large point clouds. Specifically, we regress only an

edge-specific weight vector and perform element-wise multiplication as in Equation 7 (ECC-VV). Channel mixing, albeit in an edge-unspecific fashion, is postponed to Equation 5. Finally, let us remark that  $\Theta$  is shared over time iterations and that self-loops as proposed in [45] are not necessary due to the existence of hidden states in GRUs.

**State Concatenation.** Inspired by DenseNet [17], we concatenate hidden states over all time steps and linearly transform them to produce segmentation logits  $\mathbf{y}_i$  in Equation 8. This allows to exploit the dynamics of hidden states due to increasing receptive field for the final classification.

**Relation to CRFs.** In image segmentation, post-processing of convolutional outputs using Conditional Random Fields (CRFs) is widely popular. Several inference algorithms can be formulated as (recurrent) network layers amendable to end-to-end learning [51, 43], possibly with general pairwise potentials [33, 7, 28]. While our method of information propagation shares both these characteristics, our GRUs operate on  $d_z$ -dimensional intermediate feature space, which is richer and less constrained than low-dimensional vectors representing beliefs over classes, as also discussed in [11]. Such enhanced access to information is motivated by the desire to learn a powerful representation of context, which goes beyond belief compatibilities, as well as the desire to be able to discriminate our often relatively weak unaries (superpixel embeddings). We empirically evaluate these claims in Section 4.3.

### 3.5. Further Details

**Adjacency Graphs.** In this paper, we use two different adjacency graphs between points of the input clouds:  $G_{\text{nn}}$  in Section 3.1 and  $G_{\text{vor}}$  in Section 3.2. Indeed, different definitions of adjacency have different advantages. Voronoi adjacency is more suited to capture long-range relationships between superpoints, which is beneficial for the SPG. Nearest neighbors adjacency tends not to connect objects separated by a small gap. This is desirable for the global energy but tends to produce a SPG with many small connected components, decreasing embedding quality. Fixed radius adjacency should be avoided in general as it handles the variable density of LiDAR scans poorly.

**Training.** While the geometric partitioning step is unsupervised, superpoint embedding and contextual segmentation are trained jointly in a supervised way with cross entropy loss. Superpoints are assumed to be semantically homogeneous and, consequently, assigned a hard ground truth label corresponding to the majority label among their contained points. We also considered using soft labels corresponding to normalized histograms of point labels and

training with Kullback-Leibler [25] divergence loss. It performed slightly worse in our initial experiments, though.

Naive training on large SPGs may approach memory limits of current GPUs. We circumvent this issue by randomly subsampling the sets of superpoints at each iteration and training on induced subgraphs, *i.e.* graphs composed of subsets of nodes and the original edges connecting them. Specifically, graph neighborhoods of order 3 are sampled to select at most 512 superpoints per SPG with more than  $n_{\min p}$  points, as smaller superpoints are not embedded. Note that as the induced graph is a union of small neighborhoods, relationships over many hops may still be formed and learned. This strategy also doubles as data augmentation and a strong regularization, together with randomized sampling of point clouds described in Section 3.3. Additional data augmentation is performed by randomly rotating superpoints around the vertical axis and jittering point features by Gaussian noise  $\mathcal{N}(0, 0.01)$  truncated to  $[-0.05, 0.05]$ .

**Testing.** In modern deep learning frameworks, testing can be made very memory-efficient by discarding layer activations as soon as the follow-up layers have been computed. In practice, we were able to label full SPGs at once. To compensate for randomness due to subsampling of point clouds in PointNets, we average logits obtained over 10 runs with different seeds.

## 4. Experiments

We evaluate our pipeline on the two currently largest point cloud segmentation benchmarks, Semantic3D [14] and Stanford Large-Scale 3D Indoor Spaces (S3DIS) [3], on both of which we set the new state of the art. Furthermore, we perform an ablation study of our pipeline in Section 4.3.

Even though the two data sets are quite different in nature (large outdoor scenes for Semantic3D, smaller indoor scanning for S3DIS), we use nearly the same model for both. The deep model is rather compact and 6 GB of GPU memory is enough for both testing and training. We refer to Appendix A for precise details on hyperparameter selection, architecture configuration, and training procedure.

Performance is evaluated using three metrics: per-class intersection over union (IoU), per-class accuracy (Acc), and overall accuracy (OA), defined as the proportion of correctly classified points. We stress that the metrics are computed on the original point clouds, not on superpoints.

### 4.1. Semantic3D

Semantic3D [14] is the largest available LiDAR dataset with over 3 billion points from a variety of urban and rural scenes. Each point has RGB and intensity values (the latter of which we do not use). The dataset consists of 15 training

scans and 15 test scans with withheld labels. We also evaluate on the reduced set of 4 subsampled scans, as common in past work.

In Table 2, we provide the results of our algorithm compared to other state of the art recent algorithms and in Figure 3, we provide qualitative results of our framework. Our framework improves significantly on the state of the art of semantic segmentation for this data set, *i.e.* by nearly 12 mIoU points on the reduced set and by nearly 9 mIoU points on the full set. In particular, we observe a steep gain on the "artefact" class. This can be explained by the ability of the partitioning algorithm to detect artifacts due to their singular shape, while they are hard to capture using snapshots, as suggested by [5]. Furthermore, these small object are often merged with the road when performing spatial regularization.

### 4.2. Stanford Large-Scale 3D Indoor Spaces

The S3DIS dataset [3] consists of 3D RGB point clouds of six floors from three different buildings split into individual rooms. We evaluate our framework following two dominant strategies found in previous works. As advocated by [39, 10], we perform 6-fold cross validation with micro-averaging, *i.e.* computing metrics once over the merged predictions of all test folds. Following [46], we also report the performance on the fifth fold only (Area 5), corresponding to a building not present in the other folds. Since some classes in this data set cannot be partitioned purely using geometric features (such as boards or paintings on walls), we concatenate the color information  $o$  to the geometric features  $f$  for the partitioning step.

The quantitative results are displayed in Table 3, with qualitative results in Figure 3 and in Appendix D. S3DIS is a difficult dataset with hard to retrieve classes such as white boards on white walls and columns within walls. From the quantitative results we can see that our framework performs better than other methods on average. Notably, doors are able to be correctly classified at a higher rate than other approaches, as long as they are open, as illustrated in Figure 3. Indeed, doors are geometrically similar to walls, but their position with respect to the door frame allows our network to retrieve them correctly. On the other hand, the partition merges white boards with walls, depriving the network from the opportunity to even learn to classify them: the IoU of boards for theoretical perfect classification of superpoints (as in Section 4.3) is only 51.3.

**Computation Time.** In Table 4, we report computation time over the different steps of our pipeline for the inference on Area 5 measured on a 4 GHz CPU and GTX 1080 Ti GPU. While the bulk of time is spent on the CPU for partitioning and SPG computation, we show that voxelization as pre-processing, detailed in Appendix A, leads to a significant speed-up as well as improved accuracy.

| Method                              | OA          | mIoU        | man-made terrain | natural terrain | high vegetation | low vegetation | buildings   | hard-scape  | scanning artefact | cars        |
|-------------------------------------|-------------|-------------|------------------|-----------------|-----------------|----------------|-------------|-------------|-------------------|-------------|
| reduced test set: 78 699 329 points |             |             |                  |                 |                 |                |             |             |                   |             |
| TMLC-MSR [15]                       | 86.2        | 54.2        | 89.8             | 74.5            | 53.7            | 26.8           | 88.8        | 18.9        | 36.4              | 44.7        |
| DeePr3SS [29]                       | 88.9        | 58.5        | 85.6             | 83.2            | 74.2            | 32.4           | 89.7        | 18.5        | 25.1              | 59.2        |
| SnapNet [5]                         | 88.6        | 59.1        | 82.0             | 77.3            | 79.7            | 22.9           | 91.1        | 18.4        | 37.3              | 64.4        |
| SegCloud [46]                       | 88.1        | 61.3        | 83.9             | 66.0            | 86.0            | 40.5           | 91.1        | 30.9        | 27.5              | 64.3        |
| SPG (Ours)                          | <b>94.0</b> | <b>73.2</b> | <b>97.4</b>      | <b>92.6</b>     | <b>87.9</b>     | <b>44.0</b>    | <b>93.2</b> | <b>31.0</b> | <b>63.5</b>       | <b>76.2</b> |
| full test set: 2 091 952 018 points |             |             |                  |                 |                 |                |             |             |                   |             |
| TMLC-MS [15]                        | 85.0        | 49.4        | 91.1             | 69.5            | 32.8            | 21.6           | 87.6        | 25.9        | 11.3              | 55.3        |
| SnapNet [5]                         | 91.0        | 67.4        | 89.6             | <b>79.5</b>     | 74.8            | 56.1           | 90.9        | 36.5        | 34.3              | 77.2        |
| SPG (Ours)                          | <b>92.9</b> | <b>76.2</b> | <b>91.5</b>      | 75.6            | <b>78.3</b>     | <b>71.7</b>    | <b>94.4</b> | <b>56.8</b> | <b>52.9</b>       | <b>88.4</b> |

Table 2: Intersection over union metric for the different classes of the Semantic3D dataset. OA is the global accuracy, while mIoU refers to the unweighted average of IoU of each class.

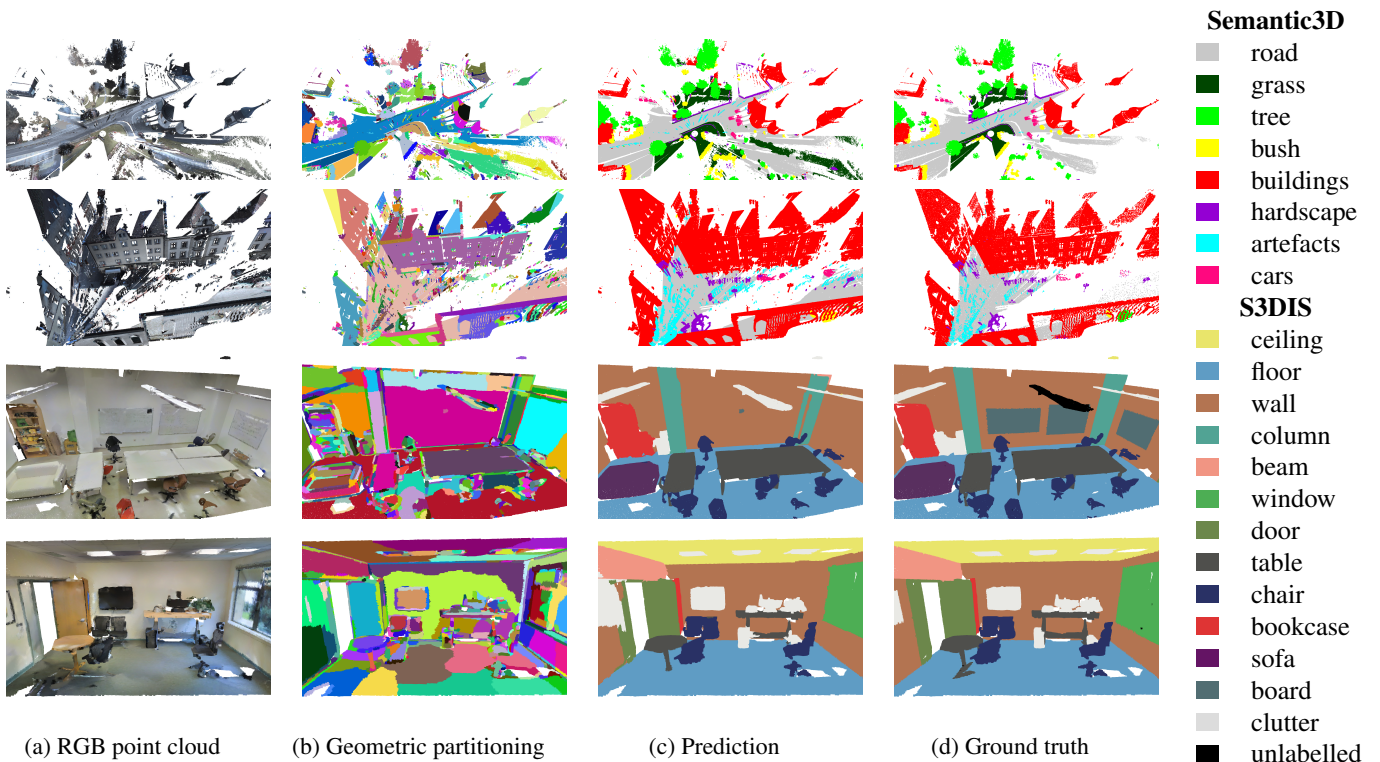


Figure 3: Example visualizations on both datasets. The colors in (b) are chosen randomly for each element of the partition.

### 4.3. Ablation Studies

To better understand the influence of various design choices made in our framework, we compare it to several baselines and perform an ablation study. Due to the lack of public ground truth for test sets of Semantic3D, we evaluate on S3DIS with 6-fold cross validation and show comparison of different models to our Best model in Table 5.

**Performance Limits.** The contribution of contextual segmentation can be bounded both from below and above.

The lower bound (Unary) is estimated by training PointNet with  $d_z = 13$  but otherwise the same architecture, denoted as PointNet13, to directly predict class logits, without SPG and GRUs. The upper bound (Perfect) corresponds to assigning each superpoint its ground truth label, and thus sets the limit of performance due to the geometric partition. We can see that contextual segmentation is able to win roughly 22 mIoU points over unaries, confirming its importance. Nevertheless, the learned model still has room of up to 26



| Method                       | OA          | mAcc         | mIoU         | ceiling      | floor        | wall         | beam        | column       | window       | door         | chair        | table        | bookcase     | sofa         | board        | clutter      |
|------------------------------|-------------|--------------|--------------|--------------|--------------|--------------|-------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| A5 PointNet [39]             | –           | 48.98        | 41.09        | 88.80        | <b>97.33</b> | 69.80        | <b>0.05</b> | 3.92         | 46.26        | 10.76        | 52.61        | 58.93        | 40.28        | 5.85         | <b>26.38</b> | 33.22        |
| A5 SEGCloud [46]             | –           | 57.35        | 48.92        | <b>90.06</b> | 96.05        | 69.86        | 0.00        | 18.37        | 38.35        | 23.12        | 75.89        | 70.40        | 58.42        | 40.88        | 12.96        | 41.60        |
| A5 SPG (Ours)                | 86.38       | <b>66.50</b> | <b>58.04</b> | 89.35        | 96.87        | <b>78.12</b> | 0.0         | <b>42.81</b> | <b>48.93</b> | <b>61.58</b> | <b>84.66</b> | <b>75.41</b> | <b>69.84</b> | <b>52.60</b> | 2.10         | <b>52.22</b> |
| PointNet [39] in [10]        | 78.5        | 66.2         | 47.6         | 88.0         | 88.7         | 69.3         | 42.4        | 23.1         | 47.5         | 51.6         | 42.0         | 54.1         | 38.2         | 9.6          | 29.4         | 35.2         |
| Engelmann <i>et al.</i> [10] | 81.1        | 66.4         | 49.7         | <b>90.3</b>  | 92.1         | 67.9         | 44.7        | 24.2         | 52.3         | 51.2         | 47.4         | 58.1         | 39.0         | 6.9          | <b>30.0</b>  | 41.9         |
| SPG (Ours)                   | <b>85.5</b> | <b>73.0</b>  | <b>62.1</b>  | 89.9         | <b>95.1</b>  | <b>76.4</b>  | <b>62.8</b> | <b>47.1</b>  | <b>55.3</b>  | <b>68.4</b>  | <b>73.5</b>  | <b>69.2</b>  | <b>63.2</b>  | <b>45.9</b>  | 8.7          | <b>52.9</b>  |

Table 3: Results on the S3DIS dataset on fold “Area 5” (top) and micro-averaged over all 6 folds (bottom). Intersection over union is shown split per class.

| Step                | Full cloud | 2 cm    | 3 cm   | 4 cm   |
|---------------------|------------|---------|--------|--------|
| Voxelization        | 0          | 40      | 24     | 16     |
| Feature computation | 439        | 194     | 88     | 43     |
| Geometric partition | 3428       | 1013    | 447    | 238    |
| SPG computation     | 3800       | 958     | 436    | 252    |
| Inference           | 10 × 24    | 10 × 11 | 10 × 6 | 10 × 5 |
| Total               | 7907       | 2315    | 1055   | 599    |
| mIoU 6-fold         | 54.1       | 60.2    | 62.1   | 57.1   |

Table 4: Computation time in seconds for the inference on S3DIS Area 5 (68 rooms, 78 649 682 points) for different voxel sizes.

mIoU points for improvement, while about 12 mIoU points are forfeited to the semantic inhomogeneity of superpoints.

**CRFs.** We compare the effect of our GRU+ECC-based network to CRF-based regularization. As a baseline (iCRF), we post-process Unary outputs by CRF inference over SPG connectivity with scalar transition matrix, as described by [13]. Next (CRF – ECC), we adapt CRF-RNN framework of Zheng *et al.* [51] to general graphs with edge-conditioned convolutions (see Appendix B for details) and train it with PointNet13 end-to-end. Finally (GRU13), we modify Best to use PointNet13. We observe that iCRF barely improves accuracy (+1 mIoU), which is to be expected, since the partitioning step already encourages spatial regularity. CRF – ECC does better (+15 mIoU) due to end-to-end learning and use of edge attributes, though it is still below GRU13 (+18 mIoU), which performs more complex operations and does not enforce normalization of the embedding. Nevertheless, the 32 channels used in Best instead of the 13 used in GRU13 provide even more freedom for feature representation (+22 mIoU).

**Ablation.** We explore the advantages of several design choices by individually removing them from Best in order to compare the framework’s performance with and without them. In NoInputGate we remove input gating in GRU; in NoConcat we only consider the last hidden state in GRU for output as  $\mathbf{y}_i = W_o \mathbf{h}_i^{(T+1)}$  instead of concatenation of all steps; in NoEdgeFeat we perform homogeneous regularization by setting all superedge features to scalar 1; and

| Model       | mAcc | mIoU |
|-------------|------|------|
| Best        | 73.0 | 62.1 |
| Perfect     | 92.7 | 88.2 |
| Unary       | 50.8 | 40.0 |
| iCRF        | 51.5 | 40.7 |
| CRF – ECC   | 65.6 | 55.3 |
| GRU13       | 69.1 | 58.5 |
| NoInputGate | 68.6 | 57.5 |
| NoConcat    | 69.3 | 57.7 |
| NoEdgeFeat  | 50.1 | 39.9 |
| ECC – VV    | 70.2 | 59.4 |

Table 5: Ablation study and comparison to various baselines on S3DIS (6-fold cross validation).

in ECC – VV we use the proposed lightweight formulation of ECC. We can see that each of the first two choices accounts for about 5 mIoU points. Next, without edge features our method falls back even below iCRF to the level of Unary, which validates their design and overall motivation for SPG. ECC – VV decreases the performance on the S3DIS dataset by 3 mIoU points, whereas it has improved the performance on Semantic3D by 2 mIoU. Finally, we invite the reader to Appendix C for further ablations.

## 5. Conclusion

We presented a deep learning framework for performing semantic segmentation of large point clouds based on a partition into simple shapes. We showed that SPGs allow us to use effective deep learning tools, which would not be able to handle the data volume otherwise. Our method significantly improves on the state of the art on two publicly available datasets. Our experimental analysis suggested that future improvements can be made in both partitioning and learning deep contextual classifiers.

The source code in PyTorch as well as the trained models are available at [https://github.com/loicland/superpoint\\_graph](https://github.com/loicland/superpoint_graph).

## References

- [1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012. **1**
- [2] A. Anand, H. S. Koppula, T. Joachims, and A. Saxena. Contextually guided semantic labeling and search for three-dimensional point clouds. *The International Journal of Robotics Research*, 32(1):19–34, 2013. **2**
- [3] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese. 3D semantic parsing of large-scale indoor spaces. In *CVPR*, 2016. **1, 6**
- [4] L. J. Ba, R. Kiros, and G. E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016. **5**
- [5] A. Boulch, B. L. Saux, and N. Audebert. Unstructured point cloud semantic labeling using deep segmentation networks. In *Eurographics Workshop on 3D Object Retrieval*, volume 2, 2017. **1, 6, 7**
- [6] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001. **3**
- [7] S. Chandra and I. Kokkinos. Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs. In *ECCV*, 2016. **5**
- [8] K. Cho, B. van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *EMNLP*, 2014. **1, 4, 5**
- [9] J. Demantk, C. Mallet, N. David, and B. Vallet. Dimensionality based scale selection in 3D lidar point clouds. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXVIII-5/W12:97–102, 2011. **3**
- [10] F. Engelmann, T. Kontogianni, A. Hermans, and B. Leibe. Exploring spatial context for 3d semantic segmentation of point clouds. In *ICCV, 3DRMS Workshop*, 2017. **1, 2, 6, 8**
- [11] R. Gadde, V. Jampani, M. Kiefel, D. Kappler, and P. Gehler. Superpixel convolutional networks using bilateral inceptions. In *ECCV*, 2016. **5**
- [12] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In *ICML*, pages 1263–1272, 2017. **2**
- [13] S. Guinard and L. Landrieu. Weakly supervised segmentation-aided classification of urban scenes from 3d LiDAR point clouds. In *ISPRS 2017*, 2017. **2, 3, 8**
- [14] T. Hackel, N. Savinov, L. Ladicky, J. D. Wegner, K. Schindler, and M. Pollefeys. Semantic3d. net: A new large-scale point cloud classification benchmark. *arXiv preprint arXiv:1704.03847*, 2017. **1, 6**
- [15] T. Hackel, J. D. Wegner, and K. Schindler. Fast semantic segmentation of 3D point clouds with strongly varying density. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 3(3), 2016. **7**
- [16] H. Hu, D. Munoz, J. A. Bagnell, and M. Hebert. Efficient 3-d scene analysis from streaming data. In *ICRA*, 2013. **2**
- [17] G. Huang, Z. Liu, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. **5**
- [18] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. **10**
- [19] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *NIPS*, pages 2017–2025, 2015. **4**
- [20] J. W. Jaromczyk and G. T. Toussaint. Relative neighborhood graphs and their relatives. *Proceedings of the IEEE*, 80(9):1502–1517, 1992. **4**
- [21] B.-S. Kim, P. Kohli, and S. Savarese. 3D scene understanding by voxel-CRF. In *ICCV*, 2013. **2**
- [22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. **11**
- [23] R. Klokov and V. S. Lempitsky. Escape from cells: Deep Kd-networks for the recognition of 3D point cloud models. *CoRR*, abs/1704.01222, 2017. **2**
- [24] H. S. Koppula, A. Anand, T. Joachims, and A. Saxena. Semantic labeling of 3D point clouds for indoor scenes. In *NIPS*, pages 244–252, 2011. **2**
- [25] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22 (1):79–86, 1951. **6**
- [26] L. Landrieu and G. Obozinski. Cut pursuit: Fast algorithms to learn piecewise constant functions on general weighted graphs. *SIAM Journal on Imaging Sciences*, 10(4):1724–1766, 2017. **3**
- [27] L. Landrieu, H. Raguét, B. Vallet, C. Mallet, and M. Weinmann. A structured regularization framework for spatially smoothing semantic labelings of 3D point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 132:102 – 118, 2017. **2**
- [28] M. Larsson, F. Kahl, S. Zheng, A. Arnab, P. H. S. Torr, and R. I. Hartley. Learning arbitrary potentials in CRFs with gradient descent. *CoRR*, abs/1701.06805, 2017. **5**
- [29] F. J. Lawin, M. Danelljan, P. Tosteberg, G. Bhat, F. S. Khan, and M. Felsberg. Deep projective 3D semantic segmentation. *arXiv preprint arXiv:1705.03428*, 2017. **7**
- [30] Y. Li, D. Tarlow, M. Brockschmidt, and R. S. Zemel. Gated graph sequence neural networks. In *ICLR*, 2016. **4**
- [31] X. Liang, L. Lin, X. Shen, J. Feng, S. Yan, and E. P. Xing. Interpretable structure-evolving LSTM. In *CVPR*, pages 2175–2184, 2017. **2**
- [32] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan. Semantic object parsing with graph LSTM. In *ECCV*, pages 125–143, 2016. **2**
- [33] G. Lin, C. Shen, A. van den Hengel, and I. D. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, 2016. **5**
- [34] Y. Lu and C. Rasmussen. Simplified Markov random fields for efficient semantic labeling of 3d point clouds. In *IROS*, pages 2690–2697, 2012. **2**
- [35] A. Martinovic, J. Knopp, H. Riemenschneider, and L. Van Gool. 3D all the way: Semantic segmentation of urban scenes from start to end in 3D. In *CVPR*, 2015. **2**

- [36] F. Monti, D. Boscaini, J. Masci, E. Rodolà, J. Svoboda, and M. M. Bronstein. Geometric deep learning on graphs and manifolds using mixture model CNNs. In *CVPR*, pages 5425–5434, 2017. 2
- [37] D. Munoz, J. A. Bagnell, N. Vandapel, and M. Hebert. Contextual classification with functional max-margin Markov networks. In *CVPR*, 2009. 2
- [38] J. Niemeyer, F. Rottensteiner, and U. Soergel. Contextual classification of lidar data and building object detection in urban areas. *ISPRS Journal of Photogrammetry and Remote Sensing*, 87:152–165, 2014. 2
- [39] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*, 2017. 1, 2, 3, 4, 6, 8, 11
- [40] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, 2017. 1, 2
- [41] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun. 3D graph neural networks for RGBD semantic segmentation. In *ICCV*, pages 5209–5218, 2017. 2
- [42] G. Riegler, A. O. Ulusoy, and A. Geiger. OctNet: Learning deep 3D representations at high resolutions. In *CVPR*, 2017. 1, 2
- [43] A. G. Schwing and R. Urtasun. Fully connected deep structured networks. *CoRR*, abs/1503.02351, 2015. 5
- [44] R. Shapovalov, D. Vetrov, and P. Kohli. Spatial inference machines. In *CVPR*, 2013. 2
- [45] M. Simonovsky and N. Komodakis. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *CVPR*, 2017. 1, 2, 4, 5, 11
- [46] L. P. Tchammi, C. B. Choy, I. Armeni, J. Gwak, and S. Savarese. SEGCloud: Semantic segmentation of 3D point clouds. *arXiv preprint arXiv:1710.07563*, 2017. 1, 2, 6, 7, 8, 11
- [47] M. Weinmann, S. Hinz, and M. Weinmann. A hybrid semantic point cloud classification-segmentation framework based on geometric features and semantic rules. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 85(3):183–194, 2017. 2
- [48] M. Weinmann, A. Schmidt, C. Mallet, S. Hinz, F. Rottensteiner, and B. Jutzi. Contextual classification of point cloud data by exploiting individual 3D neighborhoods. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3/W4:271–278, 2015. 2
- [49] D. Wolf, J. Prankl, and M. Vincze. Fast semantic segmentation of 3D point clouds using a dense CRF with learned parameters. In *ICRA*, 2015. 2
- [50] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, pages 3097–3106, 2017. 2
- [51] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015. 2, 5, 8, 11

## Appendix

### A. Model Details

**Voxelization.** We pre-process input point clouds with voxelization subsampling by computing per-voxel mean positions and observations over a regular 3D grid (5 cm bins for Semantic3D and 3 cm bins for S3DIS dataset). The resulting semantic segmentation is interpolated back to the original point cloud in a nearest neighbor fashion. Voxelization helps decreasing the computation time and memory requirement, and improves the accuracy of the semantic segmentation by acting as a form of geometric and radiometric denoising as well (Table 4 in the main paper). The quality of further steps is practically not affected, as superpoints are usually strongly subsampled for embedding during learning and inference anyway (Section 3.3 in the main paper).

**Geometric Partition.** We set regularization strength  $\mu = 0.8$  for Semantic3D and  $\mu = 0.03$  for S3DIS, which strikes a balance between semantic homogeneity of superpoints and the potential for their successful discrimination (S3DIS is composed of smaller semantic parts than Semantic3D). In addition to five geometric features  $f$  (linearity, planarity, scattering, verticality, elevation), we use color information  $o$  for clustering in S3DIS due to some classes being geometrically indistinguishable, such as boards or doors.

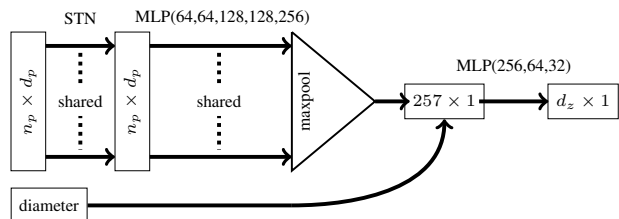


Figure 4: The PointNet embedding  $n_p$   $d_p$ -dimensional samples of a superpoint to a  $d_z$ -dimensional vector.

**PointNet.** We use a simplified shallow and narrow PointNet architecture with just a single Spatial Transformer Network (STN), see Figure 4. We set  $n_p = 128$  and  $n_{\min p} = 40$ . Input points are processed by a sequence of MLPs (widths 64, 64, 128, 128, 256) and max pooled to a single vector of 256 features. The scalar metric diameter is appended and the result further processed by a sequence of MLPs (widths 256, 64,  $d_z=32$ ). A residual matrix  $\Phi \in \mathbb{R}^{2 \times 2}$  is regressed by STN and  $(I + \Phi)$  is used to transform XY coordinates of input points as the first step. The architecture of STN is a ”small PointNet” with 3 MLPs (widths 64, 64, 128) before max pooling and 3 MLPs after (widths 128, 64, 4). Batch Normalization [18] and ReLUs are used

everywhere. Input points have  $d_p=11$  dimensional features for Semantic3D (position  $p_i$ , color  $o_i$ , geometric features  $f_i$ ), with 3 additional ones for S3DIS (room-normalized spatial coordinates, as in past work [39]).

**Segmentation Network.** We use embedding dimensionality  $d_z = 32$  and  $T = 10$  iterations. ECC-VV is used for Semantic3D (there are only 15 point clouds even though the amount of points is large), while ECC-MV is used for S3DIS (large number of point clouds). Filter-generating network  $\Theta$  is a MLP with 4 layers (widths 32, 128, 64, and 32 or  $32^2$  for ECC-VV or ECC-MV) with ReLUs. Batch Normalization is used only after the third parametric layer. No bias is used in the last layer. Superedges have  $d_f = 13$  dimensional features, normalized by mean subtraction and scaling to unit variance based on the whole training set.

**Training.** We train using Adam [22] with initial learning rate 0.01 and batch size 2, *i.e.* effectively up to 1024 superpoints per batch. For Semantic3D, we train for 500 epochs with stepwise learning rate decay of 0.7 at epochs 350, 400, and 450. For S3DIS, we train for 250 epochs with steps at 200 and 230. We clip gradients within  $[-1, 1]$ .

## B. CRF-ECC

In this section, we describe our adaptation of CRF-RNN mean field inference by Zheng *et al.* [51] for post-processing PointNet embeddings in SPG, denoted as unary potentials  $U_i$  here.

The original work proposed a dense CRF with pairwise potentials  $\Psi$  defined to be a mixture of  $m$  Gaussian kernels as  $\Psi_{ij} = \mu \sum_m w_m K_m(F_{ij})$ , where  $\mu$  is label compatibility matrix,  $w$  are parameters, and  $K$  are fixed Gaussian kernels applied on edge features.

We replace this definition of the pairwise term with a Filter generating network  $\Theta$  [45] parameterized with weights  $W_e$ , which generalizes the message passing and compatibility transform steps of Zheng *et al.*. Furthermore, we use superedge connectivity  $\mathcal{E}$  instead of assuming a complete graph. The pseudo-code is listed in Algorithm 1. Its output are marginal probability distributions  $Q$ . In practice we run the inference for  $T = 10$  iterations.

---

### Algorithm 1 CRF-ECC

---

```

 $Q_i \leftarrow \text{softmax}(U_i)$ 
while not converged do
   $\hat{Q}_i \leftarrow \sum_{j|(j,i) \in \mathcal{E}} \Theta(F_{ji}, W_e) Q_j$ 
   $\check{Q}_i \leftarrow U_i - \hat{Q}_i$ 
   $Q_i \leftarrow \text{softmax}(\check{Q}_i)$ 
end while

```

---

## C. Extended Ablation Studies

In this section, we present additional set of experiments to validate our design choices and present their results in Table 6.

**a) Spatial Transformer Network.** While STN makes superpoint embedding orientation invariant, the relationship with surrounding objects are still captured by superedges, which are orientation variant. In practice, STN helps by 4 mIoU points.

**b) Geometric Features.** Geometric features  $f_i$  are computed in the geometric partition step and can therefore be used in the following learning step for free. While PointNets could be expected to learn similar features from the data, this is hampered by superpoint subsampling, and therefore their explicit use helps (+4 mIoU).

**c) Sampling Superpoints.** The main effect of subsampling SPG is regularization by data augmentation. Too small a sample size leads to disregarding contextual information (-4 mIoU) while too large a size leads to overfitting (-2 mIoU). Lower memory requirements at training is an extra benefit. There is no subsampling at test time.

**d) Long-range Context.** We observe that limiting the range of context information in SPG harms the performance. Specifically, capping distances in  $G_{\text{vor}}$  to 1 m (as used in PointNet [39]) or 5 m (as used in SegCloud [46]) worsens the performance of our method (even more on our Semantic 3D validation set).

**e) Input Gate.** We evaluate the effect of input gating (IG) for GRUs as well as LSTM units. While a LSTM unit achieves higher score than a GRU (-3 mIoU), the proposed IG reverses this situation in favor of GRU (+1 mIoU). Unlike the standard input gate of LSTM, which controls the information flow from the hidden state and input to the cell, our IG controls the input even before it is used to compute all other gates.

**f) Regularization Strength  $\mu$ .** We investigate the balance between superpoints' discriminative potential and their homogeneity controlled by parameter  $\mu$ . We observe that the system is able to perform reasonably over a range of SPG sizes.

**g) Superpoint Sizes.** We include a breakdown of superpoint sizes for  $\mu = 0.03$  in relation to hyperparameters  $n_{\text{minp}} = 40$  and  $n_p = 128$ , showing that 93% of points are in embedded superpoints, and 79% in superpoints that are subsampled.

**Superedge Features.** Finally, in Table 7 we evaluate empirical importance of individual superedge features by removing them from Best. Although no single feature is crucial, the most being offset deviation (+3 mIoU), we remind the reader than without any superedge features the net-

---

Furthermore, SegCloud divides the inference into cubes without overlap, possibly causing inconsistencies across boundaries.

|  |                 |                    |                        |                       |
|--|-----------------|--------------------|------------------------|-----------------------|
| a) <i>Spatial transf.</i><br>mIoU                          | no<br>58.1      | <b>yes</b><br>62.1 |                        |                       |
| b) <i>Geometric features</i><br>mIoU                       | no<br>58.4      | <b>yes</b><br>62.1 |                        |                       |
| c) <i>Max superpoints</i><br>mIoU                          | 256<br>57.9     | <b>512</b><br>62.1 | 1024<br>60.4           |                       |
| d) <i>Superedge limit</i><br>mIoU                          | 1 m<br>61.0     | 5 m<br>61.3        | $\infty$<br>62.1       |                       |
| e) <i>Input gate</i><br>mIoU                               | LSTM<br>61.0    | LSTM+IG<br>61.0    | GRU<br>57.5            | <b>GRU+IG</b><br>62.1 |
| f) <i>Regularization <math>\mu</math></i><br># superpoints | 0.01<br>785 010 | 0.02<br>385 091    | <b>0.03</b><br>251 266 | 0.04<br>186 108       |
| perfect mIoU   | 90.6            | 88.2               | 86.6                   | 85.2                  |
| mIoU   | 59.1            | 59.2               | 62.1                   | 58.8                  |
| g) <i>Superpoint size</i><br>proportion of points          | 1-40<br>7%      | 40-128<br>14%      | 128-1000<br>27%        | $\geq 1000$<br>52%    |

Table 6: Ablation study of design decisions on S3DIS (6-fold cross validation). Our choices in bold.

| Model                  | mAcc | mIoU |
|------------------------|------|------|
| Best                   | 73.0 | 62.1 |
| no mean offset         | 72.5 | 61.8 |
| no offset deviation    | 71.7 | 59.3 |
| no centroid offset     | 74.5 | 61.2 |
| no len/surf/vol ratios | 71.2 | 60.7 |
| no point count ratio   | 72.7 | 61.7 |

Table 7: Ablation study of superedge features on S3DIS (6-fold cross validation).

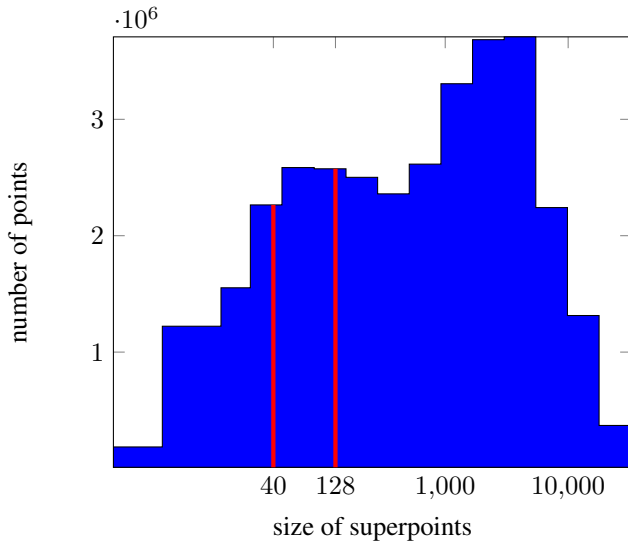


Figure 5: Histogram of points contained in superpoints of different size (in log scale) on the full S3DIS dataset. The embedding threshold  $n_{\min p}$  and subsampling threshold  $n_p$  are marked in red.

work performs distinctly worse (NoEdgeFeat, -22 mIoU).

## D. Video Illustration

We provide a video illustrating our method and qualitative results on S3DIS dataset, which can be viewed at [https://youtu.be/Ijr3kGSU\\_tU](https://youtu.be/Ijr3kGSU_tU).