Efficient 3D Semantic Segmentation with Superpoint Transformer

Damien Robert[1, 2] damien.robert@ign.fr
Hugo Raguet[3] hugo.raguet@insa-cvl.fr
Loic Landrieu[2,4] loic.landrieu@enpc.fr

[1]CSAI, ENGIE Lab CRIGEN, France
[2] LASTIG, IGN, ENSG, Univ Gustave Eiffel, France
[3]INSA Centre Val-de-Loire Univ de Tours, LIFAT, France
[4]LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS, France

Abstract

We introduce a novel superpoint-based transformer architecture for efficient semantic segmentation of large-scale 3D scenes. Our method incorporates a fast algorithm to partition point clouds into a hierarchical superpoint structure, which makes our preprocessing 7 times faster than existing superpoint-based approaches. Additionally, we leverage a self-attention mechanism to capture the relationships between superpoints at multiple scales, leading to state-of-the-art performance on three challenging benchmark datasets: S3DIS (76.0% mIoU 6-fold validation), KITTI-360 (63.5% on Val), and DALES (79.6%). With only 212k parameters, our approach is up to 200 times more compact than other state-of-the-art models while maintaining similar performance. Furthermore, our model can be trained on a single GPU in 3 hours for a fold of the S3DIS dataset, which is 7× to 70× fewer GPU-hours than the best-performing methods. Our code and models are accessible at `github.com/drprojects/superpoint_transformer`.

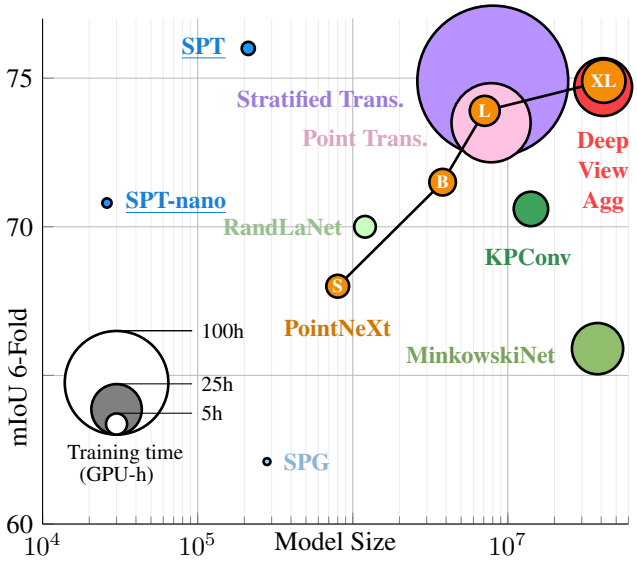


Figure 1: **Model Size vs. Performance.** We visualize the performance of different methods on the S3DIS dataset (6-fold validation) in relation to their model size in log-scale. The area of the markers indicates the GPU-time to train on a single fold. Our proposed method Superpoint Transformer (SPT) achieves state-of-the-art with a reduction of up to 200-fold in model size and 70-fold in training time (in GPU-h) compared to recent methods. The even smaller SPT-nano model achieves a fair performance with 26k parameters only.

1. Introduction

As the expressivity of deep learning models increases rapidly, so do their complexity and resource requirements [13]. In particular, vision transformers have demonstrated remarkable results for 3D point cloud semantic segmentation [56, 37, 16, 23, 32], but their high computational requirements make them challenging to train effectively. Additionally, these models rely on regular grids or point samplings, which do not adapt to the varying complexity of 3D data: the same computational effort is allocated everywhere, regardless of the local geometry or radiometry of the point cloud. This issue leads to needlessly high memory consumption, limits the number of points that can be processed simultaneously, and hinders the modeling of long-range interactions.

Superpoint-based methods [26, 24, 21, 40] address the limitation of regular grids by partitioning large point clouds into sets of points— superpoints—which adapt to the local complexity. By directly learning the interaction between superpoints instead of individual points, these methods enable the analysis of large scenes with compact and parsimonious models that can be trained faster than standard approaches. However, superpoint-based methods often require a costly preprocessing step, and their range and expressivity are lim-

ited by their use of local graph-convolution schemes [46].

In this paper, we propose a novel superpoint-based transformer architecture that overcomes the limitations of both approaches, see Figure 1. Our method starts by partitioning a 3D point cloud into a hierarchical superpoint structure that adapts to the local properties of the acquisition at multiple scales simultaneously. To compute this partition efficiently, we propose a new algorithm that is an order of magnitude faster than existing superpoint preprocessing algorithms. Next, we introduce the Superpoint Transformer (SPT) architecture, which uses a sparse self-attention scheme to learn relationships between superpoints at multiple scales. By viewing the semantic segmentation of large point clouds as the classification of a small number of superpoints, our model can accurately classify millions of 3D points simultaneously without relying on sliding windows. SPT achieves near state-of-the-art accuracy on various open benchmarks while being significantly more compact and able to train much quicker than common approaches. The main contributions of this paper are as follows:

- **Efficient Superpoint Computation:** We propose a new method to compute a hierarchical superpoint structure for large point clouds, which is more than 7 times faster than existing superpoint-based methods. Our preprocessing time is also comparable or faster than standard approaches, addressing a significant drawback of superpoint methods.
- **State-of-the-Art Performance:** Our model reaches performance at or close to the state-of-the-art for three open benchmarks with distinct settings: S3DIS for indoor scanning [3], KITTI-360 for outdoor mobile acquisitions [28], and DALES for city-scale aerial LiDAR [50].
- **Resource-Efficient Models:** SPT is particularly resource-efficient as it only has 212k parameters for S3DIS and DALES, a 200-fold reduction compared to other state-of-the-art models such as PointNeXt [39] and takes 70 times fewer GPU-h to train than Stratified Transformer [23]. The even more compact SPT-nano reaches 70.8% 6-Fold mIoU on S3DIS with only 26k parameters, making it the smallest model to reach above 70% by a factor of almost 300.

2. Related Work

This section provides an overview of the main inspirations for this paper, which include 3D vision transformers, partition-based methods, and efficient learning for 3D data.

3D Vision Transformers. Following their adoption for image processing [9, 30], Transformer architectures [51] designed explicitly for 3D analysis have shown promising results in terms of performance [56, 16] and speed [37, 32]. In particular, the Stratified Transformer of Lai *et al*. uses a specific sampling scheme [23] to model long-range interactions. However, the reliance of 3D vision transformers on arbitrary K-nearest or voxel neighborhoods leads to high memory consumption, which hinders the processing of large scenes and the ability to leverage global context cues.

Partition-Based Methods. Partitioning images into superpixels has been studied extensively to simplify image analysis, both before and after the widespread use of deep learning [1, 49]. Similarly, superpoints are used for 3D point cloud segmentation [36, 29] and object detection [17, 10]. SuperPointGraph [26] proposed to learn the relationship between superpoints using graph convolutions [46] for semantic segmentation. While this method trains fast, its preprocessing is slow and its expressivity and range are limited, as it operates on a single partition. Recent works have proposed ways of learning the superpoints themselves [24, 21, 48], which yields improved results but at the cost of an extra training step or a large point-based backbone [22].

Hierarchical partitions are used for image processing [2, 54, 55] and 3D analysis tasks such as point cloud compression [11] and object detection [7, 27]. Hierarchical approaches for semantic segmentation use Octrees with fixed grids [35, 43]. On the contrary, SPT uses a multi-scale hierarchical structure that adapts to the local geometry of the data. This leads to partitions that conform more closely to semantic boundaries, enabling the network to model the interactions between objects or object parts.

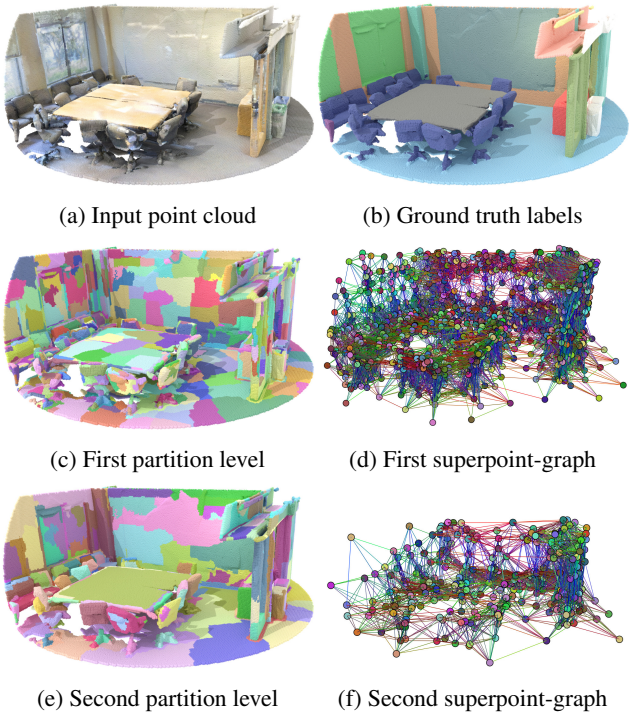
Efficient 3D Learning. As 3D scans of real-world scenes can contain hundreds of millions of points, optimizing the efficiency of 3D analysis is an essential area of research. PointNeXt [39] proposes several effective techniques that allow simple and efficient methods [38] to achieve state-of-the-art performance. RandLANet [20] demonstrates that efficient sampling strategies can yield excellent results. Sparse [14] or hybrid [31] point cloud representations have also helped reduce memory usage. However, by leveraging the local similarity of dense point clouds, superpoint-based methods can achieve an input reduction of several orders of magnitude, resulting in unparalleled efficiency.

3. Method

Our method has two key components. First, we use an efficient algorithm to segment an input point cloud into a compact multi-scale hierarchical structure. Second, a transformer-based network leverages this structure to classify the elements of the finest scale.

3.1. Efficient Hierarchical Superpoint Partition

We consider a point cloud $\mathcal{C}$ with positional and radiometric information. To learn multiscale interactions, we compute a hierarchical partition of $\mathcal{C}$ into geometrically-homogeneous superpoints of increasing coarseness; see Figure 2. We first define the concept of hierarchical partitions.



(a) Input point cloud (b) Ground truth labels

(c) First partition level (d) First superpoint-graph

(e) Second partition level (f) Second superpoint-graph

Figure 2: **Superpoint Transformer.** Our method takes as input a point cloud a) and computes its hierarchical partition into geometrically homogeneous superpoints at multiple scales: c) and e). For all partition levels, we construct superpoint adjacency graphs d) and f), which are used by an attention-based network to classify the finest superpoints.

Definition 1 Hierarchical Partitions. *A partition of a set $\mathcal{X}$ is a collection of subsets of $\mathcal{X}$ such that each element of $\mathcal{X}$ is in one and only one of such subsets. $\mathcal{P} := [\mathcal{P}_0, \cdots, \mathcal{P}_I]$ is a hierarchical partition of $\mathcal{X}$ if $\mathcal{P}_0 = \mathcal{X}$, and $\mathcal{P}_{i+1}$ is a partition of $\mathcal{P}_i$ for $i \in [0, I-1]$.*

Throughout this paper, all functions or tensors related to a specific partition level i are denoted with an exponent i.

Hierarchical Superpoint Partitions. We propose an efficient approach for constructing hierarchical partitions of large point clouds. First, we associate each point c of $\mathcal{C}$ with features f_c representing its local geometric and radiometric information. These features can be handcrafted [15] or learned [24, 21]. See the Appendix for more details on point features. We also define a graph $\mathcal{G}$ encoding the adjacency between points usually based on spatial proximity, *e.g.* k-nearest neighbors.

We view the features f_c for all c of $\mathcal{C}$ as a signal f defined on the nodes of the graph $\mathcal{G}$. Following the ideas of SuperPoint Graph [26], we compute an approximation of f into constant components by solving an energy minimization problem penalized with a graph-based notion of *simplicity*. The resulting constant components form a partition whose granularity is determined by a regularization strength $\lambda > 0$: higher values yield fewer and coarser components.

For each component of the partition, we can compute the mean position (centroid) and feature of its elements, defining a coarser point cloud on which we can repeat the partitioning process. We can now compute a hierarchical partition $\mathcal{P} := [\mathcal{P}_0, \cdots, \mathcal{P}_I]$ of $\mathcal{C}$ from a list of regularization strengths $\lambda_1, \cdots, \lambda_I$. First, we set $\mathcal{P}_0$ as the point cloud $\mathcal{C}$ and f^0 as the point features f. Then, for $i = 1$ to I, we compute (i) a partition $\mathcal{P}_i$ of f^{i-1} penalized with λ_i; (ii) the mean signal f^i for all components of $\mathcal{P}_i$. The coarseness of the resulting partitions $[\mathcal{P}_0, \cdots, \mathcal{P}_I]$ is thus strictly increasing. See the Appendix for a more detailed description of this process.

Hierarchical Graph Structure. A hierarchical partition defines a polytree structure across the different levels. Let p be an element of $\mathcal{P}_i$. If $i \in [0, I-1]$, $\text{parent}(p)$ is the component of $\mathcal{P}_{i+1}$ which contains p. If $i \in [1, I]$, $\text{children}(p)$ is the set of components of $\mathcal{P}_{i-1}$ whose parent is p.

Superpoints also share adjacency relationships with superpoints *of the same partition level*. For each level $i \geq 1$, we build a *superpoint-graph* $\mathcal{G}_i$ by connecting adjacent components of $\mathcal{P}_i$, *i.e.* superpoints whose closest points are within a distance gap $\epsilon_i > 0$. For $p \in \mathcal{P}_i$, we denote $\mathcal{N}(p) \subset \mathcal{P}_i$ the set of neighbours of p in the graph $\mathcal{G}_i$. More details on the superpoint-graph construction can be found in the Appendix.

Hierarchical Parallel ℓ_0-Cut Pursuit. Computing the hierarchical components involves solving a recursive sequence of non-convex, non-differentiable optimization problems on large graphs. We propose an adaptation of the ℓ_0-cut pursuit algorithm [25] to solve this problem. To improve efficiency, we adapt the graph-cut parallelization strategy initially introduced by Raguet *et al.* [41] in the convex setting.

3.2. Superpoint Transformer

Our proposed SPT architecture draws inspiration from the popular U-Net [45, 12]. However, instead of using grid, point, or graph subsampling, our approach derives its different resolution levels from the hierarchical partition $\mathcal{P}$.

General Architecture. As represented in Figure 3, SPT comprises an encoder with I stages and a decoder with $I-1$ stages: the prediction takes place at the level $\mathcal{P}_1$ and not on individual points. We start by computing the relative positions x of all points and superpoints with respect to their parent. For a superpoint $p \in \mathcal{P}_i$, we define x_p^i as the position of the centroid of p relative to its parent's. The coarsest superpoints of $\mathcal{P}_I$ have no parent and use the center of the scene as a reference centroid. We then normalize these values so that the sets $\{x_p^i | p \in \text{children}(q)\}$ have a radius of 1 for all $q \in \mathcal{P}_{i+1}$. We compute features for each 3D point by

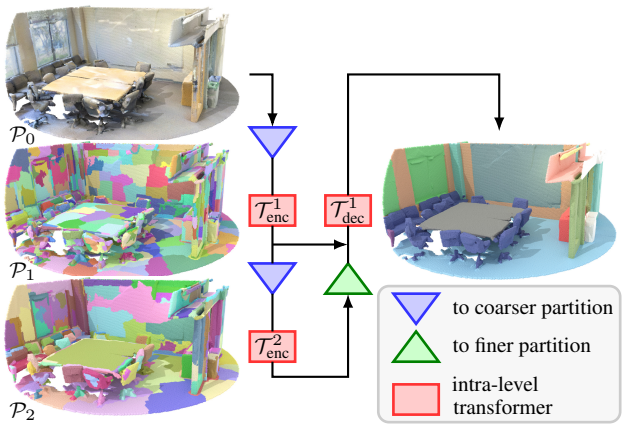


Figure 3: **Superpoint Transformer.** We represent our proposed architecture with two partitions levels $\mathcal{P}_1$ and $\mathcal{P}_2$. We use a transformer-based module to leverage the context at different scales, leading to large receptive fields. We only classify the superpoints of the partition $\mathcal{P}_1$ and not individual 3D points, allowing fast training and inference.

using a multi-layer perceptron (MLP) to mix their relative positions and handcrafted features: $g^0 := \phi^0_{\text{enc}}([x^0, f^0])$, with $[\cdot,\cdot]$ the channelwise concatenation operator.

Each level $i \geq 1$ of the encoder maxpools the features of the finer partition level $i-1$, adds relative positions x^i and propagates information between neighboring superpoints in $\mathcal{G}_i$. For a superpoint p in $\mathcal{P}_i$, this translates as:

$$g^i_p = \mathcal{T}^i_{\text{enc}} \circ \phi^i_{\text{enc}} \left(\left[x^i_p, \max_{q \in \text{children}(p)} (g^{i-1}_q) \right] \right) \quad (1)$$

with ϕ^i_{enc} an MLP and $\mathcal{T}^i_{\text{enc}}$ a transformer module explained below. By avoiding communication between the 3D points of $\mathcal{P}_0$, we bypass a potential computational bottleneck.

The decoder passes information from the coarser partition level $i+1$ to the finer level i. It uses the relative positions x^i and the encoder features g^i to improve the spatial resolution of its feature maps h^i [45]. For a superpoint p in partition $\mathcal{P}_i$ with $1 \leq i < I-1$, this can be expressed as:

$$h^i_p = \mathcal{T}^i_{\text{dec}} \circ \phi^i_{\text{dec}} \left(\left[x^i_p, g^i_p, h^{i+1}_{\text{parent}(p)} \right] \right) \quad (2)$$

with $h^I = g^I$, ϕ^i_{dec} an MLP, and $\mathcal{T}^i_{\text{dec}}$ an attention-based module similar to $\mathcal{T}^i_{\text{enc}}$.

Self-Attention Between Superpoints. We propose a variation of graph-attention networks [52] to propagate information between neighboring superpoints of the same partition level. For each level of the encoder and decoder, we associate to superpoint $p \in \mathcal{P}_i$ a triplet of key, query, value vectors K_p, Q_p, V_p of size D_{key}, D_{key} and D_{val}. These values are obtained by applying a linear layer to the corresponding feature map m after GraphNorm normalization [5].

We then characterize the relationship between two superpoints p, q of $\mathcal{P}_i$ adjacent in $\mathcal{G}_i$ by a triplet of features $a^{\text{key}}_{p,q}, a^{\text{que}}_{p,q}, a^{\text{val}}_{p,q}$ of dimensions $D_{\text{key}}, D_{\text{key}}$ and D_{val}, and whose computation is detailed in the next section. Given a superpoint p, we stack the vectors $a^{\text{key}}_{p,q}, a^{\text{que}}_{p,q}, a^{\text{val}}_{p,q}$ for $q \in \mathcal{N}(p)$ in matrices $A^{\text{key}}_p, A^{\text{que}}_p, A^{\text{val}}_p$ of dimensions $|\mathcal{N}(p)| \times D_{\text{key}}$ or $|\mathcal{N}(p)| \times D_{\text{val}}$. The modules $\mathcal{T}^i_{\text{enc}}$ and $\mathcal{T}^i_{\text{dec}}$ gather contextual information as follows:

$$[\mathcal{T}(m)]_p \overset{+}{=} \text{att}(Q^\intercal_p \oplus A^{\text{que}}_p, K_{\mathcal{N}(p)} + A^{\text{key}}_p, V_{\mathcal{N}(p)} + A^{\text{val}}_p) \ , \quad (3)$$

with $\overset{+}{=}$ a residual connection [18], $\oplus$ the addition operator with broadcasting on the first dimension, and $K_{\mathcal{N}(p)}$ the matrix of stacked vectors K_q for $q \in \mathcal{N}(p)$. The attention mechanism writes as follows:

$$\text{att}(Q, K, V) := V^\intercal \, \text{softmax}\left(\frac{Q \odot K\mathbf{1}}{\sqrt{|\mathcal{N}(p)|}} \right) \ , \quad (4)$$

with $\odot$ the Hadamard termwise product and $\mathbf{1}$ a column-vector with D_{key} ones. Our proposed scheme is similar to classic attention schemes with two differences: (i) the queries adapt to each neighbor, and (ii) we normalize the softmax with the neighborhood size instead of the key dimension. In practice, we use multiple independent attention modules in parallel (multi-head attention) and several consecutive attention blocks.

3.3. Leveraging the Hierarchical Graph Structure

The hierarchical superpoint partition $\mathcal{P}$ can be used for more than guidance for graph pooling operations. Indeed, we can learn expressive adjacency encodings capturing the complex adjacency relationships between superpoints and employ powerful supervision and augmentation strategies based on the hierarchical partitions.

Adjacency Encoding. While the adjacency between two 3D points is entirely defined by their distance vector, the relationships between superpoints are governed by additional factors such as their alignment, proximity, and difference in sizes or shapes. We characterize the adjacency of pairs of adjacent superpoints of the same partition level using a set of handcrafted features based on: (i) the relative positions of centroids, (ii) position of paired points in each superpoints, (iii) the superpoint principal directions, and (iv) the ratio between the superpoints' length, volume, surface, and point count. These features are efficiently computed only once during preprocessing.

For each pair of superpoints (p, q) adjacent in $\mathcal{G}_i$, we jointly compute the concatenated $a^{\text{key}}_{p,q}, a^{\text{que}}_{p,q}, a^{\text{val}}_{p,q}$ by applying an MLP ϕ^i_{adj} to the handcrafted adjacency features defined above. Further details on the superpoint-graph construction and specific adjacency features are provided in the Appendix.

Hierarchical Supervision. We propose to take advantage of the nested structure of the hierarchical partition $\mathcal{P}$ into the supervision of our model. We can naturally associate the superpoints of any level $i \geq 1$ with a set of 3D points in $\mathcal{P}_0$. The superpoints at the finest level $i = 1$ are almost semantically pure (see Figure 6), while the superpoints at coarser levels $i > 1$ typically encompass multiple objects. Therefore, we use a dual learning objective: (i) we predict the most frequent label within the superpoints of $\mathcal{P}_1$, and (ii) we predict the label distribution for the superpoints of $\mathcal{P}_i$ with $i > 1$. We supervise both predictions with the cross-entropy loss.

Let y_p^i denote the true label distribution of the 3D points within a superpoint $p \in \mathcal{P}_i$, and $\hat{y}_p^i$ a one-hot-encoding of its most frequent label. We use a dedicated linear layer at each partition level to map the decoder feature g_p^i to a predicted label distribution z_p^i. Our objective function can be formulated as follows:

$$\mathcal{L} = \sum_{p \in \mathcal{P}_1} \frac{-N_p^1}{|\mathcal{C}|} H(\hat{y}_p^1, z_p^1) + \sum_{i=2}^{I} \sum_{p \in \mathcal{P}_i} \frac{\mu^i N_p^i}{|\mathcal{C}|} H(y_p^i, z_p^i) \ , \quad (5)$$

where $\mu^2, \cdots, \mu^I$ are positive weights, N_p^i represents the number of points within a superpoint $p \in \mathcal{P}_i$, and $|\mathcal{C}|$ is the total number of points in the point cloud, and $H(y, z) = -\sum_{k \in \mathcal{K}} y_k \log(z_k)$ and $\mathcal{K}$ the class set.

Superpoint-Based Augmentations. Although our approach classifies superpoints rather than individual 3D points, we still need to load the points of $\mathcal{P}_0$ in memory to embed the superpoints from $\mathcal{P}_1$. However, since superpoints are designed to be geometrically simple, only a subset of their points is needed to characterize their shape. Therefore, when computing the feature g_p^1 of a superpoint p of $\mathcal{P}_1$ containing n points with Eq. (1), we sample only a portion $\tanh(n/n_{\text{max}})$ of its points, with a minimum of n_{min}. This sampling strategy reduces the memory load and acts as a powerful data augmentation. The lightweight version of our model SPT-nano goes even further. It ignores the points entirely and only use handcrafted features to embed the superpoints of $\mathcal{P}_1$, thus avoiding entirely the complexity associated with the size of the input point cloud $\mathcal{P}_0$.

To further augment the data, we exploit the geometric consistency of superpoints and their hierarchical arrangement. During the batch construction, we randomly drop each superpoint with a given probability at all levels. Dropping superpoints at the fine levels removes random objects or object parts, while dropping superpoints at the coarser levels removes entire structures such as walls, buildings, or portions of roads, for example.

Table 1: **Partition Configuration.** We report the point count of different datasets before and after subsampling, as well as the size of the partitions.

Dataset	Points	Subsampled	$\lvert \mathcal{P}_1 \rvert$	$\lvert \mathcal{P}_2 \rvert$
S3DIS [3]	273m	32m	979k	292k
DALES [50]	492m	449m	14.8m	2.56m
KITTI-360 [28]	919m	432m	16.2m	2.98m

4. Experiments

We evaluate our model on three diverse datasets described in Section 4.1. In Section 4.2, we evaluate our approach in terms of precision, but also quantify the gains in terms of pre-processing, training, and inference times. Finally, we propose an extensive ablation study in Section 4.3.

4.1. Datasets and Models

Datasets. To demonstrate its versatility, we evaluate SPT on three large-scale datasets of different natures.

S3DIS [3]. This indoor dataset of office buildings contains over 274 million points across 6 building floors—or areas. The dataset is organized by individual rooms, but can also be processed by considering entire areas at once.

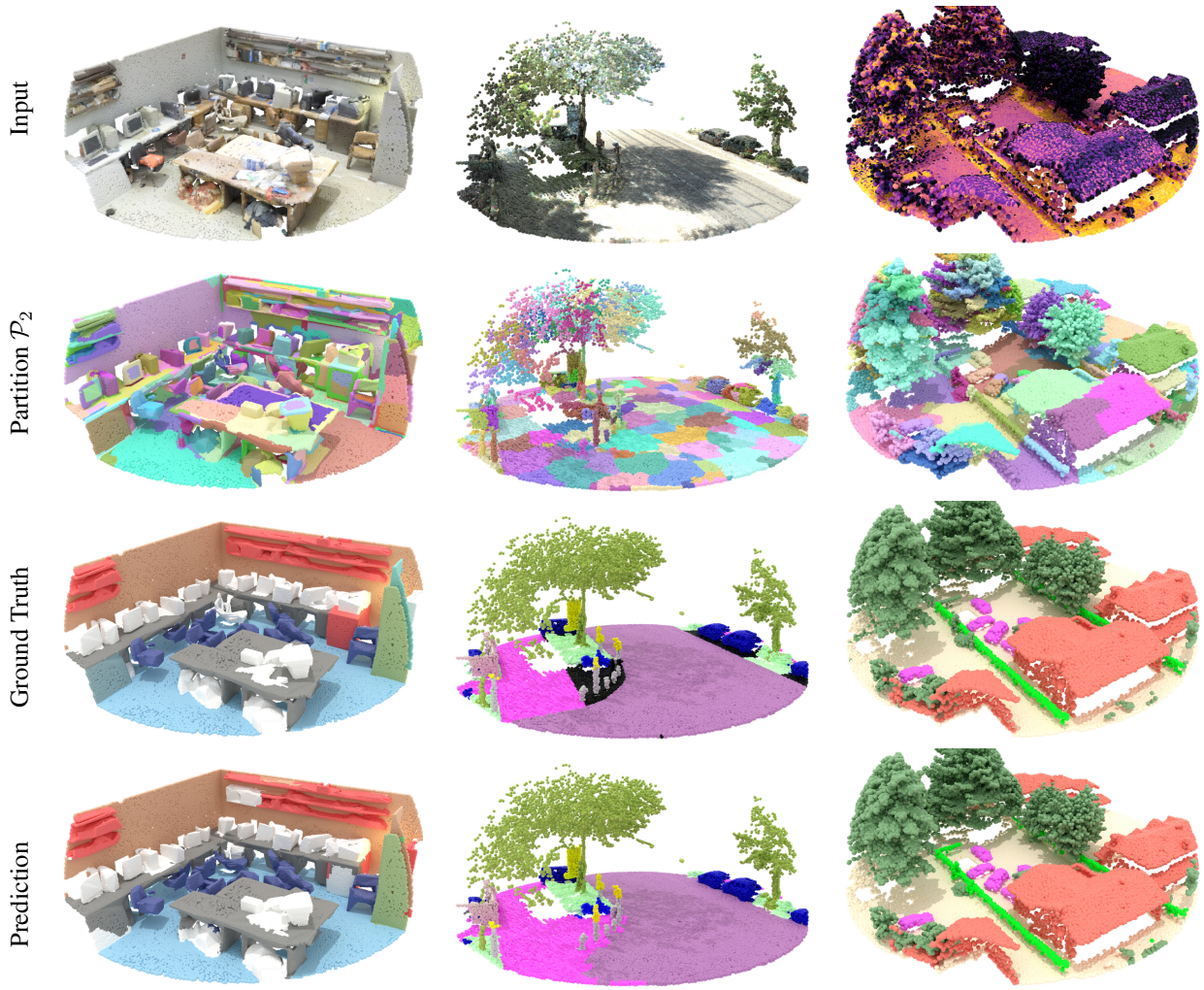
KITTI-360 [28]. This outdoor dataset contains more than 100 k laser scans acquired in various urban settings on a mobile platform. We use the *accumulated point clouds* format, which consists of large scenes with around 3 million points. There are 239 training scenes and 61 for validation.

DALES [50]. This 10 km^2 aerial LiDAR dataset contains 500 millions of points across 40 urban and rural scenes, including 12 for evaluation.

We subsample the datasets using a 3cm grid for S3DIS, and 10cm for KITTI-360 and DALES. All accuracy metrics are reported for the full, unsampled point clouds. We use a two-level partition ($I = 2$) with $\mu^2 = 50$ for all datasets and select the partition parameters to obtain a 30-fold reduction between $\mathcal{P}_1$ and $\mathcal{P}_0$ and a further 5-fold reduction for $\mathcal{P}_2$. See Table 1 for more details.

Models. We use the same model configuration for all three datasets with minimal adaptations. All transformer modules have a shared width D_{val}, a small key space of dimension $D_{\text{key}} = 4$, 16 heads, with 3 blocks in the encoder and 1 in the decoder. We set $D_{\text{val}} = 64$ for S3DIS and DALES (210k parameters), and $D_{\text{val}} = 128$ (777k parameters) for KITTI360. See the Appendix and our open repository for the detailed configuration of all modules.

We also propose SPT-nano, a lightweight version of our model that does not compute point-level features but operates directly on the first partition level $\mathcal{P}_1$. The value of the maxpool over points in Eq. (1) for $i = 1$ is replaced by f^1, the aggregated handcrafted point features at the level 1 of



(a) S3DIS (b) KITTI-360 (c) DALES

Figure 4: **Qualitative Results.** We represent input samples (with color or intensity) of our approach and its predictions for all three datasets. Additionally, we show the coarsest partition level and demonstrate how superpoints can accurately capture the contours of complex objects and classify them accordingly. Black points are unlabeled in the ground truth.

the partition. This model never considers the full point cloud $\mathcal{P}_0$ but only operates on the partitions. For this model, we set $D_{\text{val}} = 16$ for S3DIS and DALES (26k parameters), and $D_{\text{val}} = 32$ for KITTI360 (70k parameters).

Batch Construction. Batches are sampled from large *tiles*: entire building floors for S3DIS, and full scenes for KITTI-360 or DALES. Each batch is composed of 4 randomly sampled portions of the partition with a radius of 7m for S3DIS and 50m for KITTI and DALES, allowing us to model long-range interactions. During training, we apply a superpoint dropout rate of 0.2 for each superpoint at all hierarchy levels, as well as random rotation, tilting, point jitter and hand-crafted features dropout. When sampling points within each superpoint, we set $n_{\text{min}} = 32$ and $n_{\text{max}} = 128$.

Optimization. We use the ADAMW optimizer [34] with default parameters, a weight decay of 10^{-4}, a learning rate of 10^{-2} for DALES and KITTI-360 on and 10^{-1} for S3DIS. The learning rate for the attention modules is 10 times smaller than for other weights. Learning rates are warmed up from 10^{-6} for 20 epochs and progressively reduced to 10^{-6} with cosine annealing [33].

4.2. Quantitative Evaluation

Performance Evaluation. As seen in Table 2, SPT performs at the state-of-the-art on two of three datasets despite being a significantly smaller model. On S3DIS, SPT beats

Table 2: **Performance Evaluation.** We report the Mean Intersection-over-Union of different methods on three different datasets. SPT performs on par or better than recent methods with significantly fewer parameters. † superpoint-based. ⋆/∗ model with 777k/70k parameters.

Model	Size $\times 10^6$	S3DIS 6-Fold	S3DIS Area 5	KITTI 360 val	DALES
PointNet++ [38]	3.0	56.7	-	-	68.3
† SPG [26]	0.28	62.1	58.0	-	60.6
ConvPoint [4]	4.7	68.2	-	-	67.4
† SPG + SSP [24]	0.29	68.4	61.7	-	-
† SPNet [21]	0.32	68.7	-	-	-
MinkowskiNet [8, 6]	37.9	69.1	65.4	58.3	-
RandLANet [20]	1.2	70.0	-	-	-
KPConv [47]	14.1	70.6	67.1	-	**81.1**
Point Trans.[56]	7.8	73.5	70.4	-	-
RepSurf-U [42]	0.97	74.3	68.9	-	-
DeepViewAgg [44]	41.2	74.7	67.2	62.1	-
Strat. Trans. [23, 53]	8.0	74.9	**72.0**	-	-
PointNeXt-XL [39]	41.6	74.9	71.1	-	-
† **SPT** (ours)	0.21	**76.0**	68.9	**63.5**⋆	79.6
† **SPT-nano** (ours)	**0.026**	70.8	64.9	57.2∗	75.2

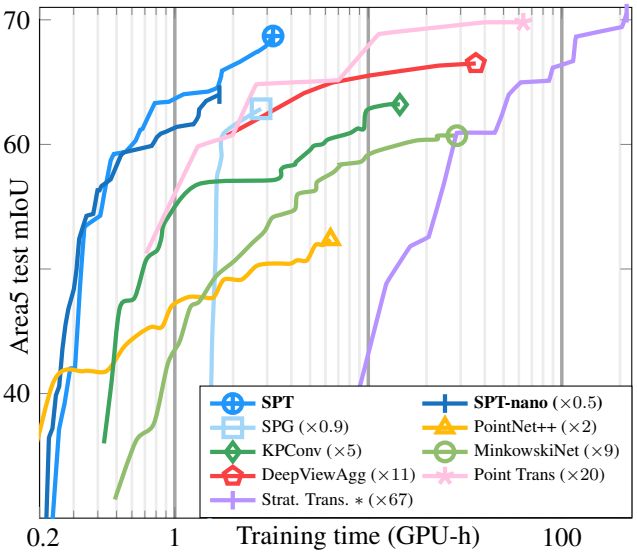


Figure 5: **Training Speed.** We report the evolution of the test mIoU for S3DIS Area 5 for different methods *until the best epoch is reached*. The curves are shifted right according to the preprocessing time. We report in parenthesis the time ratio compared to SPT.

PointNeXt-XL with 196× fewer parameters. On KITTI-360, SPT outperforms MinkowskiNet despite a size ratio of 49, and surpasses the performance of the even larger multimodal point-image model DeepViewAgg. On DALES, SPT outperforms ConvPoint by more than 12 points with over 21 times fewer parameters. Although SPT is 1.5 points behind KPConv on this dataset, it achieves these results with 67 times fewer parameters. SPT achieves significant performance improvements over all superpoint-based methods on all datasets, ranging from 7 to 14 points. SPT overtakes the SSP and SPNet superpoint methods that *learn* the partition in a two-stage training setup, leading to pre-processing times of several hours.

Interestingly, the lightweight SPT-nano model matches KPConv and MinkowskiNet with only 26k parameters.

See Figure 4 for qualitative illustrations.

Preprocessing Speed. As reported in Table 3, our implementation of the preprocessing step is highly efficient. We can compute partitions, superpoint-graphs, and handcrafted features, and perform I/O operations quickly: 12.4min for S3DIS, 117 for KITTI-360, and 148 for DALES using a server with a 48-core CPU. An 8-core workstation can preprocess S3DIS in 26.6min. Our preprocessing time is as fast or faster than point-level methods and 7× faster than SuperPoint Graph's, thus alleviating one of the main drawbacks of superpoint-based methods.

Training Speed. We trained several state-of-the-art methods from scratch and report in Figure 5 the evolution of test performance as a function of training time. We used the official training logs for the multi-GPU Point Transformer and Stratified Transformer. SPT can train much faster than all methods not based on superpoints while attaining similar performance. Although Superpoint Graph trains even faster, its performance saturates earlier, 6.0 mIoU points below SPT . We also report the inference time of our method in Table 3, which is significantly lower than competing approaches, with a speed-up factor ranging from 8 to 80. All speed measurements were conducted on a single-GPU server (48 cores, 512Go RAM, A40 GPU). Nevertheless, our model can be trained on a standard workstation (8 cores, 64Go, 2080Ti) with smaller batches, taking only 1.5 times longer and with comparable results.

SPT performs on par or better than complex models with up to two orders of magnitude more parameters and significantly longer training times. Such efficiency and compactness have many benefits for real-world scenarios where hardware, time, or energy may be limited.

4.3. Ablation Study

We evaluate the impact of several design choices in Table 4 and reports here our observations.

a) Handcrafted features. Without handcrafted point features, our model perform worse on all datasets. This observation is in line with other works which also remarked the

Table 3: **Efficiency Analysis.** We report the preprocessing time for the entire S3DIS dataset and the training and inference time for Area 5. SPT and SPT-nano shows significant speedups in pre-processing, training, and inference times.

	Preprocessing in min	Training in GPU-h	Inference in s
PointNet++ [38]	8.0	6.3	42
KPConv [47]	23.1	14.1	162
MinkowskiNet [8]	20.7	28.8	83
Stratified Trans. [23]	8.0	216.4	30
Superpoint Graph [26]	89.9	1.3	16
SPT (ours)	12.4	3.0	2
SPT-nano (ours)	12.4	1.9	1

Table 4: **Ablation Study.** Impact of some of our design choices on the mIoU for all tested datasets.

Experiment	S3DIS 6-Fold	KITTI 360 Val	DALES
Best Model	76.0	63.5	79.6
a) No handcrafted features	-0.7	-4.1	-1.4
b) No adjacency encoding	-6.3	-5.4	-3.0
b) Fewer edges	-3.5	-1.1	-0.3
c) No point sampling	-1.3	-0.9	-0.5
c) No superpoint sampling	-2.7	-2.5	-0.7
c) Only 1 partition level	-8.4	-5.1	-0.9

positive impact of well-designed handcrafted features on the performance of smaller models [19, 42].

b) Influence of Edges. Removing the adjacency encoding between superpoints leads to a significant drop of 6.3 points on S3DIS; characterizing the relative position and relationship between superpoints appears crucial to exploiting their context. We also find that pruning the 50% longest edges of each superpoint results in a systematic performance drop, highlighting the importance of modeling long relationships.

c) Partition-Based Improvements. We assess the impact of several improvements made possible by using hierarchical superpoints. First, we find that using all available points when embedding the superpoints of $\mathcal{P}_1$ instead of random sampling resulted in a small performance drop. Second, setting the superpoint dropout rate to 0 worsens the performance by over 2.5 points on S3DIS and KITTI-360.

While we did not observe better results with three or more partition levels, only using one level leads to a significant loss of performance for all datasets.

d) Influence of Partition Purity. In Figure 6, we plot the performance of the "oracle" model which associates

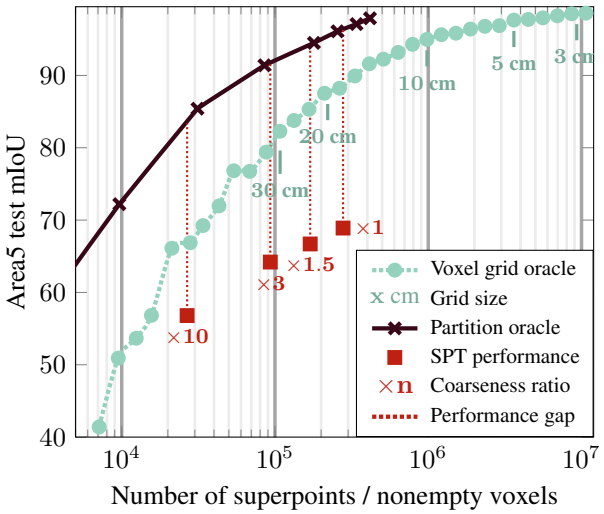


Figure 6: **Partition Purity.** We plot the highest achievable "oracle" prediction for our partitions and a regular voxel grid. We also show the performance of SPT for 4 partitions with a coarseness ratio from $\times 1$ to $\times 10$.

to each superpoint of $\mathcal{P}_1$ with its most frequent true label. This model acts as an upper bound on the achievable performance with a given partition. Our proposed partition has significantly higher semantic purity than a regular voxel grid with as many nonempty voxels as superpoints. This implies that our superpoints adhere better to semantic boundaries between objects.

We also report the performance of our model for different partitions of varying coarseness, measured as the number of superpoints in $\mathcal{P}_1$. Using, respectively, $\times 1.5$ ($\times 3$) fewer superpoints leads to a performance drop of 2.2 (4.7) mIoU points, but reduce the training time to 2.4 (1.6) hours. The performance of SPT is more than 20 points below the oracle, suggesting that the partition does not strongly limit its performance.

Limitations. See the Appendix.

5. Conclusion

We have introduced the Superpoint Transformer approach for semantic segmentation of large point clouds, combining superpoints and transformers to achieve state-of-the-art results with significantly reduced training time, inference time, and model size. This approach particularly benefits large-scale applications and computing with limited resources. More broadly, we argue that small, tailored models can offer a more flexible and sustainable alternative to large, generic models for 3D learning. With training times of a few hours on a single GPU, our approach allows practitioners to easily customize the models to their specific needs, enhancing the overall usability and accessibility of 3D learning.

Acknowledgements. This work was funded by ENGIE Lab CRIGEN. This work was supported by ANR project READY3D ANR-19-CE23-0007, and was granted access to the HPC resources of IDRIS under the allocation AD011013388R1 made by GENCI. We thank Bruno Vallet, Romain Loiseau and Ewelina Rupnik for inspiring discussions and valuable feedback.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 2012.

[2] Pablo Arbelaez. Boundary extraction in natural images using ultrametric contour maps. *CVPR Workshop*, 2006.

[3] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3D semantic parsing of large-scale indoor spaces. *CVPR*, 2016.

[4] Alexandre Boulch. ConvPoint: Continuous convolutions for point cloud processing. *Computers & Graphics*, 2020.

[5] Tianle Cai, Shengjie Luo, Keyulu Xu, Di He, Tie-yan Liu, and Liwei Wang. GraphNorm: A principled approach to accelerating graph neural network training. *ICML*, 2021.

[6] Thomas Chaton, Nicolas Chaulet, Sofiane Horache, and Loic Landrieu. Torch-Points3D: A modular multi-task framework for reproducible deep learning on 3D point clouds. *3DV*, 2020.

[7] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3D instance segmentation. *CVPR*, 2021.

[8] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal ConvNets: Minkowski convolutional neural networks. *CVPR*, 2019.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2020.

[10] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3D-MPA: Multi-proposal aggregation for 3D semantic instance segmentation. *CVPR*, 2020.

[11] Yuxue Fan, Yan Huang, and Jingliang Peng. Point cloud compression based on hierarchical point clustering. In *APSIPA ASC*, 2013.

[12] Hongyang Gao and Shuiwang Ji. Graph U-Nets. *ICML*, 2019.

[13] Charlie Giattino, Edouard Mathieu, Julia Broden, and Max Roser. Artificial intelligence. *Our World in Data*, 2022. https://ourworldindata.org/artificial-intelligence.

[14] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3D semantic segmentation with submanifold sparse convolutional networks. *CVPR*, 2018.

[15] Stéphane Guinard and Loic Landrieu. Weakly supervised segmentation-aided classification of urban scenes from 3D LiDAR point clouds. *ISPRS Workshop*, 2017.

[16] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. PCT: Point cloud transformer. *CVM*, 2021.

[17] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3D instance segmentation. *CVPR*, 2020.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[19] Pai-Hui Hsu and Zong-Yi Zhuang. Incorporating handcrafted features into deep learning for point cloud classification. *Remote Sensing*, 2020.

[20] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. RandLA-Net: Efficient semantic segmentation of large-scale point clouds. *CVPR*, 2020.

[21] Le Hui, Jia Yuan, Mingmei Cheng, Jin Xie, Xiaoya Zhang, and Jian Yang. Superpoint network for point cloud oversegmentation. *ICCV*, 2021.

[22] Xin Kang, Chaoqun Wang, and Xuejin Chen. Region-enhanced feature learning for scene semantic segmentation. *arXiv preprint arXiv:2304.07486*, 2023.

[23] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3D point cloud segmentation. *CVPR*, 2022.

[24] Loic Landrieu and Mohamed Boussaha. Point cloud oversegmentation with graph-structured deep metric learning. *CVPR*, 2019.

[25] Loic Landrieu and Guillaume Obozinski. Cut pursuit: Fast algorithms to learn piecewise constant functions on general weighted graphs. In *SIAM Journal on Imaging Sciences*, 2017.

[26] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. *CVPR*, 2018.

[27] Zhihao Liang, Zhihao Li, Songcen Xu, Mingkui Tan, and Kui Jia. Instance segmentation in 3D scenes using semantic superpoint tree networks. *CVPR*, 2021.

[28] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. *TPAMI*, 2022.

[29] Yangbin Lin, Cheng Wang, Dawei Zhai, Wei Li, and Jonathan Li. Toward better boundary preserved supervoxel segmentation for 3D point clouds. *ISPRS journal of photogrammetry and remote sensing*, 2018.

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CVPR*, 2021.

[31] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel CNN for efficient 3D deep learning. *NeurIPS*, 2019.

[32] Romain Loiseau, Mathieu Aubry, and Loïc Landrieu. Online segmentation of LiDAR sequences: Dataset and algorithm. *ECCV*, 2022.

[33] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *ICLR*, 2017.

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019.

[35] J Narasimhamurthy, Karthikeyan Vaiapury, Ramanathan Muthuganapathy, and Balamuralidhar Purushothaman. Hierarchical-based semantic segmentation of 3D point cloud using deep learning. *Smart Computer Vision*, 2023.

[36] Jeremie Papon, Alexey Abramov, Markus Schoeler, and Florentin Worgotter. Voxel cloud connectivity segmentation-supervoxels for point clouds. *CVPR*, 2013.

[37] Chunghyun Park, Yoonwoo Jeong, Minsu Cho, and Jaesik Park. Fast point transformer. *CVPR*, 2022.

[38] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 2017.

[39] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. PointNeXt: Revisiting PoinNet++ with improved training and scaling strategies. *NeurIPS*, 2022.

[40] Xingwen Quana, Binbin Hea, Marta Yebrab, Changmin Yina, Zhanmang Liaoa, Xueting Zhanga, and Xing Lia. Hierarchical semantic segmentation of urban scene point clouds via group proposal and graph attention network. *International Journal of Applied Earth Observations and Geoinformation*, 2016.

[41] Hugo Raguet and Loic Landrieu. Parallel cut pursuit for minimization of the graph total variation. *ICML Workshop on Graph Reasoning*, 2019.

[42] Haoxi Ran, Jun Liu, and Chengjie Wang. Surface representation for point clouds. *CVPR*, 2022.

[43] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. OctNet: Learning deep 3D representations at high resolutions. *CVPR*, 2017.

[44] Damien Robert, Bruno Vallet, and Loic Landrieu. Learning multi-view aggregation in the wild for large-scale 3D semantic segmentation. *CVPR*, 2022.

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. *MICCAI*, 2015.

[46] Martin Simonovsky and Nikos Komodakis. Dynamic edge-conditioned filters in convolutional neural networks on graphs. *CVPR*, 2017.

[47] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. KPConv: Flexible and deformable convolution for point clouds. *ICCV*, 2019.

[48] Anirud Thyagharajan, Benjamin Ummenhofer, Prashant Laddha, Om Ji Omer, and Sreenivas Subramoney. Segment-fusion: Hierarchical context fusion for robust 3D semantic segmentation. *CVPR*, 2022.

[49] Wei-Chih Tu, Ming-Yu Liu, Varun Jampani, Deqing Sun, Shao-Yi Chien, Ming-Hsuan Yang, and Jan Kautz. Learning superpixels with segmentation-aware affinity loss. *CVPR*, 2018.

[50] Nina Varney, Vijayan K Asari, and Quinn Graehling. DALES: A large-scale aerial LiDAR data set for semantic segmentation. *CVPR Workshops*, 2020.

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

[52] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *ICLR*, 2018.

[53] Qi Wang, Shengge Shi, Jiahui Li, Wuming Jiang, and Xiangde Zhang. Window normalization: Enhancing point cloud understanding by unifying inconsistent point densities. 2022.

[54] Yongchao Xu, Thierry Géraud, and Laurent Najman. Hierarchical image simplification and segmentation based on mumford–shah-salient level line selection. *Pattern Recognition Letters*, 2016.

[55] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, Sercan Ö Arik, and Tomas Pfister. Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding. *AAAI*, 2022.

[56] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. *ICCV*, 2021.