SemAffiNet: Semantic-Affine Transformation for Point Cloud Segmentation

Ziyi Wang[1,2], Yongming Rao[1,2], Xumin Yu[1,2], Jie Zhou[1,2], Jiwen Lu[1,2*]
[1]Department of Automation, Tsinghua University
[2]Beijing National Research Center for Information Science and Technology
wziyi20@mails.tsinghua.edu.cn; raoyongming95@gmail.com;
yuxm20@mails.tsinghua.edu.cn; {jzhou, lujiwen}@tsinghua.edu.cn

Abstract

*Conventional point cloud semantic segmentation methods usually employ an encoder-decoder architecture, where mid-level features are locally aggregated to extract geometric information. However, the over-reliance on these class-agnostic local geometric representations may raise confusion between local parts from different categories that are similar in appearance or spatially adjacent. To address this issue, we argue that mid-level features can be further enhanced with semantic information, and propose **semantic-affine transformation** that transforms features of mid-level points belonging to different categories with class-specific affine parameters. Based on this technique, we propose **SemAffiNet** for point cloud semantic segmentation, which utilizes the attention mechanism in the Transformer module to implicitly and explicitly capture global structural knowledge within local parts for overall comprehension of each category. We conduct extensive experiments on the ScanNetV2 and NYUv2 datasets, and evaluate semantic-affine transformation on various 3D point cloud and 2D image segmentation baselines, where both qualitative and quantitative results demonstrate the superiority and generalization ability of our proposed approach. Code is available at* https://github.com/wangzy22/SemAffiNet.

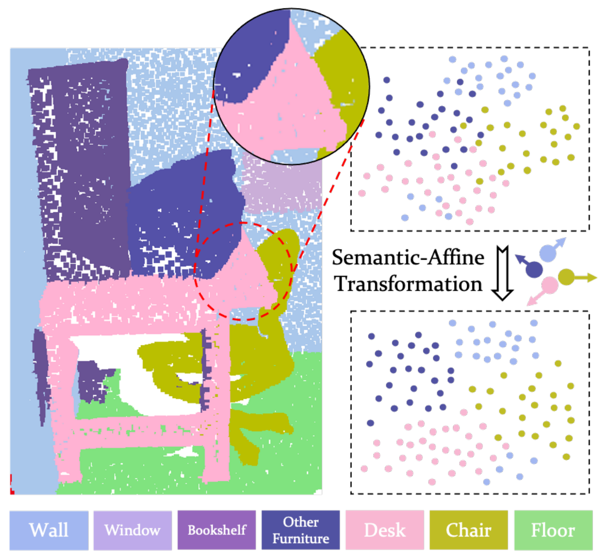


Figure 1. Illustration of Semantic-Affine Transformation. The left figure is the input point cloud, with different colors referring to different categories. We select one local part (red circle) and extract its mid-level features, resulting in the top-right figure. Some representations of the points from different categories are entangled with each other in the embedding space, which may be caused by the appearance similarities or spatial adjacency. We propose to perform the semantic-affine transformation on these mid-level features, predicting particular affine parameters for each category respectively. Therefore, once obtaining classification predictions of mid-level points, we can pull points from the same category closer while pushing points from different categories apart via semantic-affine transformation, as is shown in the bottom-right figure.

1. Introduction

Point cloud semantic segmentation is a fundamental task for both structural representation learning [11, 45, 56] and stereoscopic scene understanding [15, 27, 49] in computer vision. It aims at partitioning the scene space into semantic-meaningful regions based on conformation and geometry knowledge inherited in point cloud layouts. Its successful applications in autonomous driving, robotic manipulation, and virtual reality have been motivating researchers to develop more fine-grained and more accurate solutions.

Recent methods for point cloud segmentation usually adopt an encoder-decoder architecture as image semantic segmentation [4, 6, 7, 51, 69], ranging from voxel based-ones [11, 20, 52, 67, 71] to point-based ones [34, 46, 54, 60]. Despite the popularity of the encoder-decoder architecture, there still exists the *local confusion problem* as shown in Figure 1. On the one hand, there are local parts from different categories but with similar shapes, such as the similar

*Corresponding author

legs of chairs and desks. On the other hand, adjacent local parts are blended in the input space and may obfuscate the model during segmentation, leaving ambiguous segment boundaries. The reasons are two folds: the heavy use of the local aggregation during feature processing, and the class-agnostic nature of the mid-level features. In the commonly-used encoder-decoder architecture, the mid-level features of the decoder are locally aggregated via convolution or set abstraction. The limitation of receptive fields produces similar feature vectors for visually-similar local parts, and the aggregation operation results in entangled mid-level features for spatially-adjacent local parts. Therefore, it is insufficient to use geometric-only information and the encoder-decoder architecture demands more knowledge to separate similar and entangled local representations. One possible solution to this problem is alleviating the reliance on geometric knowledge and introducing additional semantic information to enrich mid-level features. However, most existing literature fails to fully exploit semantic knowledge in the network design of the encoder-decoder architecture, as semantic annotations are mostly used for data augmentation [8,44,65] or supervision on final prediction [45,46,56]. Therefore, the mid-level features from the intermediate layers are only implicitly or weakly supervised via gradient descent, making them almost class-agnostic.

To address the *local confusion problem*, we propose **Semantic-Affine Transformation** to transform mid-level decoder features with class-specific affine parameters that encode semantic information, which explicitly pulls features from the same category closer and pushes features from different categories apart. In this way, we enhance the semantic representation ability of mid-level features and boost semantic segmentation performance. Based on the proposed semantic-affine transformation, we design a semantic-aware network named **SemAffiNet** and introduce Transformer [55] to manage semantic information both implicitly and explicitly. The Transformer encoder implicitly communicates geometric information across modalities via the self-attention technique, while the special design of class queries in the Transformer decoder performs explicit semantic-aware reasoning to predict semantic-affine parameters via the cross-attention mechanism. We conduct extensive experiments on the ScanNetV2 [14] dataset and outperform the previous state-of-the-art BPNet [26] baselines. We also evaluate on the NYUv2 [43] dataset to verify the generalization ability of the SemAffiNet model. As the core of SemAffiNet, the proposed semantic-affine transformation is evaluated on both 3D point cloud and 2D image segmentation baselines under various settings, revealing the generalization ability of the proposed transformation.

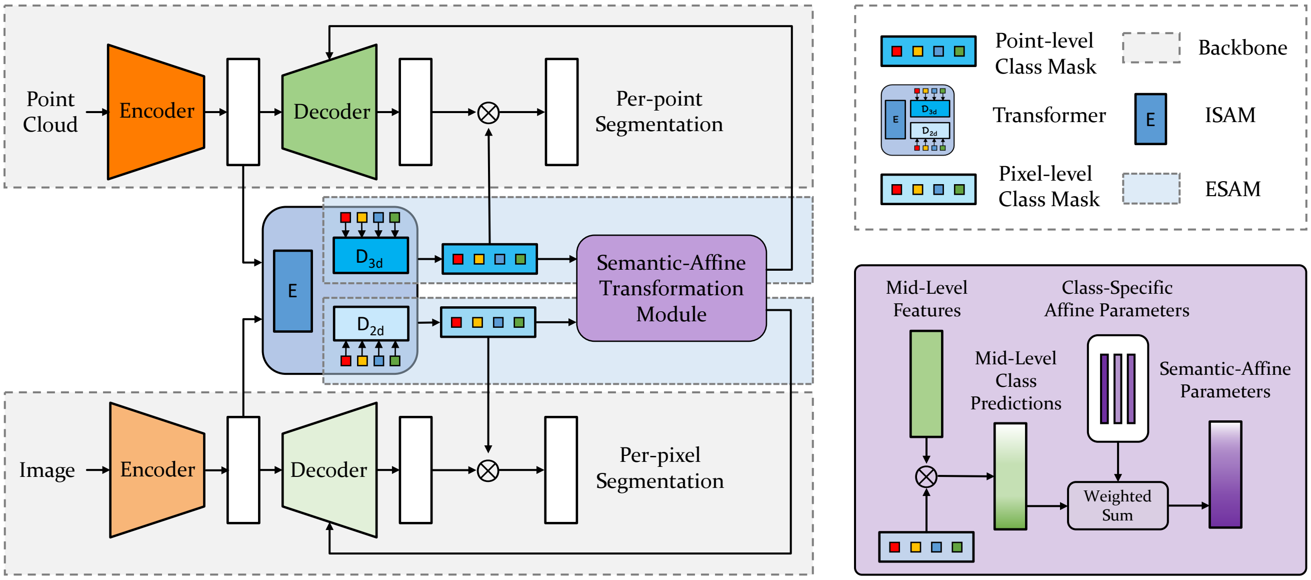
In conclusion, the contributions of our paper can be summarized as follows: (1) We propose Semantic-Affine Transformation to enhance the semantic representation ability of mid-level features in encoder-decoder segmentation architecture. (2) We propose SemAffiNet to perform semantic-aware segmentation both explicitly and implicitly via special designs of Transformer modules. (3) We conduct experiments on various datasets under different settings, revealing the superiority and generalization ability of our method.

2. Related Work

Point Cloud Semantic Segmentation. Existing point cloud semantic segmentation methods can be divided into four categories: voxel-based, point-based, projection-based, and hybrid models. Voxel-based methods aim at partitioning 3D space into ordered voxels and translating 2D convolutional encoder-decoder architectures to 3D conditions, leading by VoxNet [40]. The heavy time expense and memory cost have been addressed by later researches, including sparse convolution [11, 20], efficient data structure migration [32, 50] and novel voxelization techniques [52, 67, 71]. Point-based methods directly process points and aggregate local information instead of using conventional regular convolution kernel, leading by PointNet [45] and PointNet++ [46]. Nowadays point-based methods has become the mainstream of point cloud cognition tasks and has been developed into many branches, including MLP-based [16, 18, 25, 45, 46], convolution-based [34, 54, 60, 61] and graph-based [33, 56] posterity. Projection-based methods are mostly designed for efficient processing, including image projection [1, 13] and spherical projection [2, 42, 58, 59]. Hybrid methods are more complicated systems that combine different processing methods or fuse different modality information. Some approaches combine long-range correlations from voxel-based methods and meticulous details from point-based methods [10,37,53,63], while other approaches fuse 2D and 3D knowledge [26, 48].

Semantic-Aware Segmentation. As semantic information is critical to the segmentation task, some works elaborately design special semantic-aware approaches to boost semantic segmentation performance [19, 39, 47]. Some methods aim at reasoning context relations differently between same-category pairs and distinct-categories pairs. DependencyNet [35] for image segmentation unifies dependency reasoning at three semantic levels: intra-class, inter-class and global. CGANet [39] for point cloud segmentation utilizes different aggregation strategies between the same category and different categories. Other methods propose multi-scale supervision to realize comprehensive semantic guidance. In 2D vision, CPM [57] introduces intermediate supervision periodically while MSS-net [30] proposes layer-wise loss. In 3D vision, RFCC [19] puts forward omni-supervision on all levels of the decoder layers.

Different from the above methods, our method achieves comprehensive semantic awareness via semantic-affine



(*a*) SemAffiNet (*b*) Semantic-Affine Transformation Module

Figure 2. Illustration of the proposed network architecture. (*a*) shows the pipeline of the SemAffiNet, which consists of two backbone branches (gray), an implicit semantic-aware module (ISAM, dark blue), and two explicit semantic-aware modules (ESAM, light blue dash square). Moreover, ESAM is composed of a Transformer decoder and a Semantic-Affine Transformation module, which is further illustrated in (*b*). We calculate the weighted sum of class-specific affine parameters to obtain semantic-affine parameters for mid-level points, with mid-level classification confidences as linear combination weights.

transformation for mid-level features. Therefore, we don't need different aggregation modules that increase model scale. Additionally, our semantic guidance for intermediate layers is stronger than merely multi-level supervisions.

Transformer for Segmentation. Transformer [55] has achieved a great success in many computer vision tasks, such as classification [17, 36], detection [5], and reconstruction [64]. Recent works employ the attention mechanism in Transformer to exploit long-range correlations for deeper context comprehension and better segmentation results [21, 68, 70]. Maskformer [9] proposes a mask classification model that utilizes Transformer to predict binary masks and unifies both semantic- and instance-level segmentation. SOTR [22] proposes to dynamically generating instance segmentation masks based on Transformer attention module.

While we utilize a similar mask classification structure as Maskformer in our SemAffiNet, we migrate this idea from 2D image processing to 3D point cloud understanding, which is not trivial. Moreover, we propose semantic-affine transformation to further enhance the mask classification pipeline, which brings more progression than mask classification according to our ablation studies.

3. Approach

In the following section, we will first give an overview of the proposed SemAffiNet in Section 3.1. Then we will present details of the architecture, introducing the proposed semantic affine transformation in Section 3.2, revealing how we wrap it into a plug-and-play explicit semantic-aware module in Section 3.3, and presenting the auxiliary implicit semantic-aware module in Section 3.4. Finally, we will introduce the loss function design in Section 3.5.

3.1. Overview

We elaborately design SemAffiNet to perform the semantic-affine transformation on mid-level features from the conventional encoder-decoder model. Figure 2 shows the overall architecture, which can be divided into the following three parts: (1) Backbone, (2) the Explicit Semantic-Aware Module (ESAM), (3) the Implicit Semantic-Aware Module (ISAM).

First of all, the backbone choice of SemAffiNet is flexible and our proposed modules can be easily added to any encoder-decoder segmentation architecture. We choose BPNet [26] that consists of two encoder-decoder branches for 2D and 3D modalities to evaluate our proposed semantic-affine transformation. Please refer to the BPNet paper or our supplementary material for further details.

Most importantly, ESAM wraps our proposed semantic-affine transformation into a plug-and-play module, that explicitly exploits semantic information in mid-level features with specially designed Transformer decoders. As shown Figure 2, we employ two ESAM blocks in light

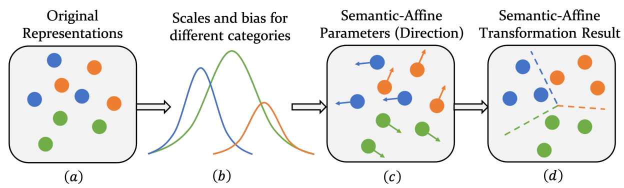


Figure 3. Illustrations of the Semantic-Affine Transformation. (a) is the original representations of mid-level points. (b) is the underlying distinctive scales and bias for different categories. The arrow directions in (c) represent the regressed class-specific affine parameters. (d) shows the semantic-affine transformation results of mid-level features, which are more classification-friendly.

blue dash square to manage semantic knowledge from different domains respectively. The Transformer decoder utilizes the cross-attention mechanism to obtain long-range dependencies for better semantic perception, while the following Semantic-Aware Transformation Module transforms mid-level features of backbone decoders with class-specific affine parameters to enlarge semantic distinctions across categories.

Last but not least, ISAM utilizes the self-attention mechanism in the Transformer encoder to enhance high-level features that are outputs from the backbone encoder and inputs to ESAM. The proposed ISAM fuses multi-modality information and realizes implicit semantic awareness.

3.2. Semantic-Affine Transformation

As semantic-affine transformation is the key contribution of our paper, we will first introduce its conceptual and technical details in this sub-section. The core idea is predicting semantic-affine parameters for each category and then determining affine parameters for each point based on its classification prediction.

Suppose there are N classes and the shape of mid-level features $\mathbf{f}^i$ at layer i is (n_i, d_i), indicating that there are n_i points at layer i and each point is represented by a d_i-dim feature vector. From Section 3.3 we can predict categories for each point p_j^i at layer i and obtain a classification confidence vector $\mathbf{a}_j^i = [a_{j1}^i, a_{j2}^i, \cdots, a_{jN}^i]$, $0 \le a_{jk}^i \le 1, 1 \le j \le n_i, 1 \le k \le N, \sum_k a_{jk}^i = 1$, where a_{jk}^i denotes confidence of point p_j at layer i belonging to class k. Simultaneously, we regress semantic-affine parameters for each class at layer i, including scale factor $\mathbf{s}^i = \{s_k^i \in \mathbb{R}^{d_i}, 1 \le k \le N, s_{kl}^i \ge 0, 1 \le l \le d_i\}$ and offset bias $\mathbf{b}^i = \{b_k^i \in \mathbb{R}^{d_i}, 1 \le k \le N\}$. Further technical details about the learning process of these semantic-aware parameters can be found in Section 3.3.

Then we can obtain the semantic-specific affine parameters S_j^i, B_j^i for each point p_j at layer i via linear combination of the affine parameters family $\mathbf{s}^i, \mathbf{b}^i$ based on per-point classification confidence vector $\mathbf{a}_j^i$:

$$S_j^i = \sum_k a_{jk}^i s_k^i, \quad B_j^i = \sum_k a_{jk}^i b_k^i \tag{1}$$

Once normalized with zero-mean and unit covariance, $\hat{f}_j^i = (f_j^i - \mu(f_j^i))/\sigma(f_j^i)$ can be further enhanced with the semantic-affine transformation to obtain semantic-aware mid-level feature $\tilde{f}_j^i$, which is used to replace f_j^i:

$$\tilde{f}_j^i = S_j^i \hat{f}_j^i + B_j^i \tag{2}$$

Note that we implement a soft semantic-affine parameters assignment that introduces a linear combination instead of a hard one that restricts the searching space within the exact value of $\mathbf{s}, \mathbf{b}$. In other words, the hard assignment strategy only considers the highest confidence scores of the category predictions and chooses the exact affine parameters accordingly. The reason is that the mid-level points are aggregation results of itself and its neighbors in the adjacent lower level, and the neighbors may have different categories from the center query point. Therefore, the mid-level points p_i may represent multiple classes when its corresponding patch in lower levels lies at the edges. We will further discuss this issue in Section 3.5.

We illustrate the principle of the semantic-affine transformation in Figure 3. The original representations of points from different categories in (a) are entangled with each other. Then we train a network to capture the underlying distinctions of scales and bias between different categories, as is shown in (b). Then we express these distinctions with semantic-affine parameters, which are demonstrated as arrow directions in (c). Finally, in (d), the semantic-affine transformation explicitly transforms mid-level features with similar category distributions towards similar scales and offsets, thus pulling them closer. On the contrary, for mid-level features with distinctive category distributions, their discrepancy in scales and offsets pushes them further apart.

According to the discussions above, the most important prerequisites for semantic-affine transformation are two-fold. The first one is a precise class predictor, predicting accurate category distributions for mid-level points. The second one is a powerful semantic-aware module, regressing representative affine transformation parameters for each class. In Section 3.3, we show that these two prerequisites can be satisfied by a multi-layer Transformer decoder.

3.3. The Explicit Semantic-Aware Module

Our goal is to wrap the learnable parameters of semantic-aware transformation introduced above into a plug-and-play module that can be implemented in most encoder-decoder-style semantic segmentation architecture. We propose an Explicit Semantic-Aware Module (ESAM) as a multi-layer Transformer decoder module to jointly and explicitly inference *semantic class masks* and *semantic affine parameters*.

Cross attention in Transformer decoder. The input to ESAM is the high-level feature $\mathbf{f}^0$, which is the output from the backbone encoder. Particularly, we design N learnable class queries $q^{(c)}$ to enquire semantic-specific knowledge. Then each layer of ESAM utilizes the attention mechanism to reason semantic information from $\mathbf{f}^0$:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where d_k is the scaling factor, and Q, K, V are queries, keys, and values matrix. We employ class queries $q^{(c)}$ as Q, while K, V are mapped embeddings of $\mathbf{f}^0$. From each layer of ESAM, we can obtain class-specific features $h \in \mathbb{R}^{N \times d_h}$ of d_h dimensions that encode both class-wise semantic knowledge from $q^{(c)}$ and geometry knowledge of the scene from $\mathbf{f}^0$.

Semantic-affine parameters. Based on the multi-layer Transformer decoder architecture of ESAM, we utilize the intermediate outputs from ESAM mid-level layers to calculate the semantic-affine parameters for mid-level layers of the backbone decoder: $\mathbf{s}_k^i = \text{MLP}(h_k^u), \mathbf{b}_k^i = \text{MLP}(h_k^u)$, where $1 \leq k \leq N$ denotes class k, and there is a one-to-one and order-preserved mapping between backbone decoder mid-layer i and ESAM mid-layer u. The principle is using output features of deeper ESAM layer u to calculate semantic-affine parameters of deeper backbone decoder layer i. Please refer to the supplementary material for detailed correspondence between i and u. The outcoming scale and bias parameters $\mathbf{s}, \mathbf{b}$ are then used to perform semantic-affine transformation introduced in Section 3.2.

Semantic class masks. The segmentation masks $M = \{m_k \in \mathbb{R}^{d_m}, 1 \leq k \leq N\}$ for N classes per-point classification are calculated via the output h^{-1} from the final layer of ESAM: $m_k = \text{MLP}(h_k^{-1})$. Then we implement dot product on class mask m_k and per-point feature $\mathbf{f}^i = \{f_j^i \in \mathbb{R}^{d_i}, 1 \leq j \leq n^i\}$ to calculate the confidence matrix $A^i = \{a_{jk}^i\}$: $a_{jk}^i = m_k \cdot f_j^i$, where a_{jk}^i indicates the confidence of point p_j at layer i of the backbone decoder belonging to class k and is utilized in Section 3.2.

Note that in conventional per-point segmentation methods such as FCN [38] or encoder-decoder-style architecture UNet [51], per-point feature $\mathbf{f}^{-1}$ from the final layer of the decoder is further processed by an MLP block as segmentation head to obtain per-point class predictions. However, instead of linearly combining channel-wise values via fully connected layers to predict class confidence, we implement the above mask prediction and dot production with per-point features to classify mid-level and final point features. The advantages of predicting class masks are two folds. On the one hand, the class masks have clearer interpretive meanings. Each mask m_k represents the comprehensive feature of the class k, and the dot product between m_k and the point feature measures the similarity between the point and the class. Therefore, the point is categorized into its most similar class. On the other hand, the class masks are more flexible to be implemented to predict categories of mid-level points. Once M is obtained, the dot-product operation is more lightweight than the MLP forward calculation. Therefore, class mask prediction is more suitable to combine with our proposed semantic-affine transformation.

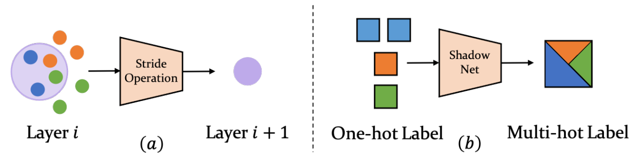


Figure 4. Illustration of ShadowNet that is designed to trace multi-hot ground truth label for mid-level points. Column (a) shows the local aggregation of stride operation in the backbone encoder. Column (b) shows how we record the class distribution of the points at the higher layer according to the class distribution of its corresponding patch at the lower layer.

In conclusion, ESAM explicitly predicts semantic class masks and semantic-affine parameters via a multi-layer Transformer decoder with learnable class queries. Each class query represents a category and inquires class-specific information in scene geometric representations. Then the semantic class masks are used to perform more flexible and lightweight multi-level per-point classification, while semantic-affine parameters are applied to transform mid-level features of the backbone decoder.

3.4. The Implicit Semantic-Aware Module

Besides ESAM that explicitly reasons semantic information via specially designed learnable class queries in the Transformer decoder, we also design ISAM to implicitly exploit and fuse semantic knowledge from multi-modalities.

In our implementation, 2D high-level feature $\mathbf{f}^{0,2d}$ and 3D high-level feature $\mathbf{f}^{0,3d}$ are concatenated together to form mixed Transformer encoder input $\mathbf{f}^{0,mix}$. Then queries, keys, and values matrix are obtained via three different linear transformations of the input mix feature $\mathbf{f}^{0,mix}$. Therefore, the self-attention calculation in Equation (3) are performed at both inter-modality and intra-modality to obtain more representative high-level features $\bar{\mathbf{f}}^{0,2d}$ and $\bar{\mathbf{f}}^{0,3d}$. On the one hand, the intra-modality self-attention reasons long-range dependencies between local parts from the same modality, attaching global information to local part features. On the other hand, the inter-modality self-attention captures similarities between parts from different modalities, creates soft correspondences, and merges knowledge from the other modal to local features. Therefore, the output features of ISAM acquire both long-range global knowledge and multi-modal information, making them more robust and more representative. Thus the *implicit* semantic-awareness is realized via intra-domain and inter-domain self-attention, where queries that are more semantic-similar to keys contribute more to updating the values matrix.

3.5. The Loss Function

Following conventional supervised segmentation approaches, we use the Cross Entropy loss for vanilla 2D per-pixel segmentation and 3D per-point segmentation.

Additionally, since we predict category labels for mid-level points in backbone decoders, we calculate the Binary Cross Entropy loss for mid-level segmentation. In order to obtain the mid-level ground truth of backbone decoders, we design ShadowNets to trace the *stride* operation in their corresponding backbone encoders. As is shown in Figure 4, stride operation in the encoder aggregates points $\{p_j^i\}$ within a local patch $P_{j'}^i$ at layer i into a meta-point $p_{j'}^{i+1}$ at higher layer $i+1$. Suppose the one-hot label for point p_j^i is l_j^i, then our ShadowNet assigns the meta-point with multi-hot label $l_{j'}^{i+1}$ that records labels of all points within its corresponding patch $P_{j'}^i$ at the lower layer i:

$$l_{j'}^{i+1} = \min(1, \sum_{p_j^i \in P_{j'}^i} l_j^i) \tag{4}$$

In this way, the multi-hot ground truth labels represent class distributions of points at mid-level layers.

4. Experiments

In this section, we conduct extensive experiments on various datasets to verify the superiority of the proposed semantic-affine transformation and the SemAffiNet architecture, calculating the class-wise mean intersection over union (mIoU) as evaluation metrics. In Section 4.1, we will present the quantitative and qualitative results of the SemAffiNet, comparing them with previous works. Then in Section 4.2, we will implement the semantic-affine transformation on both 3D point cloud and 2D image segmentation baselines under different conditions to prove its generalization ability. Furthermore, in Section 4.3, we will provide ablation studies to demonstrate the effectiveness of each proposed module. Finally, in Section 4.4, we will discuss the limitations of our proposed approach. Additionally, experiments setup including datasets introduction and implementation details can be found in supplementary materials.

4.1. Main Results

Following BPNet [26], we conduct semantic segmentation experiments on ScanNetV2, evaluating both 2cm and 5cm voxel settings on validation set. The quantitative results are shown in Table 1. Under 5cm settings, SemAffiNet exceeds BPNet baseline by 1.5% and 3.1% respectively on 3D mIoU and 2D mIoU metrics. Under 2cm settings, SemAffiNet outperforms BPNet by 2.0% and 1.5% respectively on 3D and 2D segmentation results. We also surpass other previous 3D semantic segmentation methods that take point clouds as input [11, 19, 20, 44, 46, 54, 60, 62] or use point clouds and auxiliary 2D images as input [29, 66].

Table 1. Quantitative results on the ScanNetV2 dataset. We compare both 3D and 2D mIoU with our baseline method BPNet. We also compare 3D mIoU with other works that use point cloud as input. Methods marked with * use additional 2D image input.

Method	3D mIoU(%)	2D mIoU(%)
PointNet++ [46]	53.5	–
PointConv [60]	61.0	–
PointASNL [62]	63.5	–
MVPNet* [29]	66.4	–
KPConv [54]	69.2	–
SparseConvNet [20]	69.3	–
RFCR [19]	70.2	–
FAConv* [66]	72.0	–
MinkowskiNet [11]	72.2	–
Mix3D [44]	73.6	–
BPNet* [26] (5cm)	70.6	65.1
SemAffiNet* (5cm)	**72.1**	**68.2**
BPNet* (2cm)	72.5	72.7
SemAffiNet* (2cm)	**74.5**	**74.2**

Table 2. NYUv2 2D image segmentation results (13-class task). We compare our SemAffiNet with typical RGB-D based methods and joint 2D-3D methods on dense pixel classification accuracy metric. Baseline results are from the BPNet [26] paper and the results of [14] is on the 11-class task.

Method	Accuracy(%)
SceneNet [23]	52.5
Hermans *et al.* [24]	54.3
SemanticFusion [41]	59.2
ScanNet [14]	60.7
3DMV [15]	71.2
BPNet [26]	73.5
SemAffiNet (Ours)	**78.3**

Besides ScanNetV2 which mainly focuses on 3D point cloud segmentation, we also conduct experiments on the NYUv2 [43] dataset that consists of RGB images and corresponding depth maps. We convert the depth image to pseudo point clouds according to camera pose and employ SemAffiNet. Following BPNet [26], we adopt the 13-class configuration and report the dense pixel classification accuracy. The experimental results are shown in Table 2, and our SemAffiNet outperforms those baselines by a large margin,

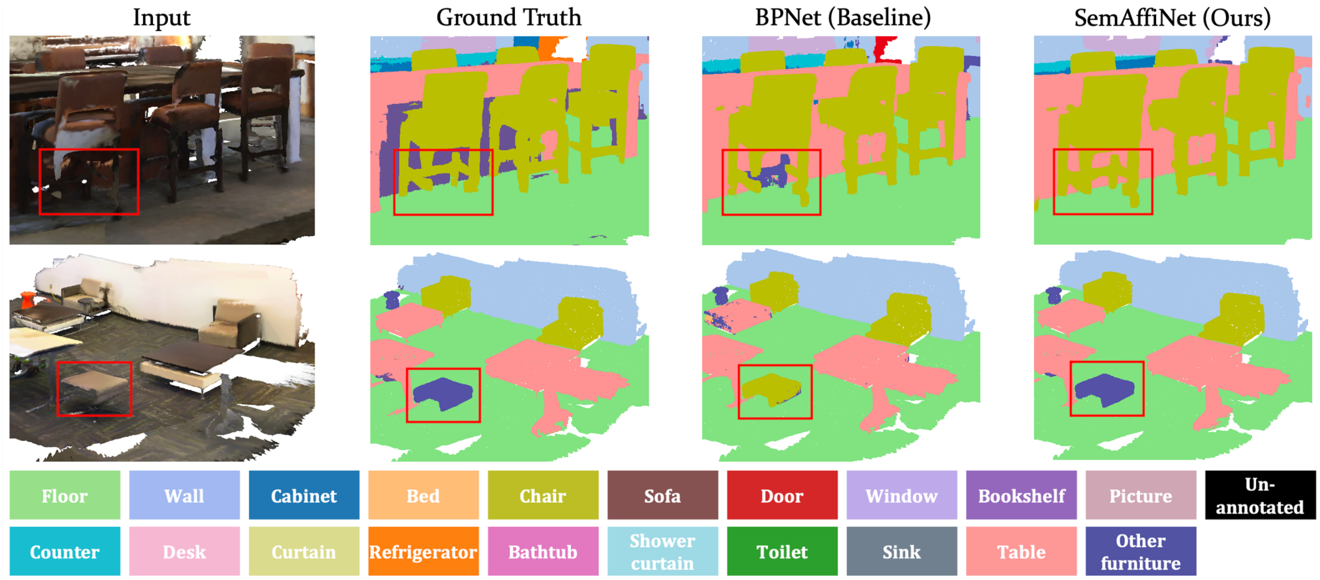


Figure 5. Qualitative results of the SemAffiNet on ScanNetV2 point cloud semantic segmentation task. The first and second columns are the input point clouds and corresponding ground truth labels. The third and fourth columns are the segmentation results of BPNet baseline and our method SemAffiNet respectively. As is shown in the red rectangle in the first line, SemAffiNet is able to recognize *chair crossbeam* while BPNet cannot. From the second line, SemAffiNet correctly identifies that the object in the red rectangle is *not chair*, while BPNet fails. The final line shows the correspondence between categories and visualization colors.

which verifies its superiority.

The quantitative results are shown in Figure 5. From the first line example, our SemAffiNet is able to recognize the subtle *chair crossbeam* that BPNet baseline fails to correctly segment. From the second line example, our SemAffiNet correctly classifies the *other furniture* object which is visually similar to *chair*. From these qualitative results, we can prove the superiority of the SemAffiNet over the BPNet baseline. On the one hand, it has the ability to recognize subtle local parts that are easy to be confused with the background. On the other hand, it can correctly classify objects that are visually similar to other categories.

4.2. Effects of Semantic-Affine Transformation

Besides SemAffiNet architecture, we also evaluate the semantic-affine transformation independently by wrapping it as a plug-and-play module ESAM and inserting it into various 3D point cloud and 2D image encoder-decoder segmentation baselines. Point cloud methods can be further divided into voxel-based and point-based ones.

For voxel-based 3D segmentation methods, we select MinkowskiNet [11] as the baseline given its efficiency and high performance. We conduct semantic segmentation on the ScanNetV2 [14] validation set under both 5cm and 2cm voxelization settings, and the results are shown in the first two lines of Table 3. By inserting ESAM into MinkowskiNet, we reach a higher performance on 3D mIoU results.

For point-based 3D segmentation methods, we choose KPConv [54] (rigid) as the baseline, since it is one of the most classical point-based approaches and has been analyzed as the baseline in many recently published papers [19, 44]. We conduct semantic segmentation on the S3DIS [3] dataset and the results are shown in the third line of Table 3. The qualitative results demonstrate the improvement made by ESAM, convincing the generalization ability of the semantic-affine transformation on both voxel-based

Table 3. Quantitative results of inserting ESAM into various semantic segmentation baselines. The first two lines show results based on 3D voxel-based MinkowskiNet on the ScanNetV2 dataset. The third line shows results based on 3D point-based KPConv on the S3DIS dataset. The last two lines show results based on SemanticFPN on the 2D Cityscapes dataset.

Method	Dataset	mIoU(%)
MinkowskiNet (5cm) [11]	ScanNetV2	67.4
+ ESAM	ScanNetV2	**68.8**
MinkowskiNet (2cm)	ScanNetV2	72.2
+ ESAM	ScanNetV2	**74.0**
KPConv [54]	S3DIS	65.8
+ ESAM	S3DIS	**66.7**
SemanticFPN (Res50) [31]	Cityscapes	76.1
+ ESAM	Cityscapes	**77.2**
SemanticFPN (Res101)	Cityscapes	77.4
+ ESAM	Cityscapes	**79.0**

Table 4. Ablation studies on individual contributions of SemAffiNet designs on the ScanNetV2 validation dataset under 5cm voxelization.

	Ablation on ↓	idx	VN	$\hat{\text{F}}$.FC	$\hat{\text{F}}$.M	$\hat{\text{M}}$.FC	$\hat{\text{M}}$.M	BN	AdaIN	SA	Fuse	2D mIoU	3D mIoU
Backbone	View Num	a	3	✓	✗	✗	✗	✓	✗	✗	✗	65.1	70.6
		b	5	✓	✗	✗	✗	✓	✗	✗	✗	65.5	70.1
ESAM	Multi-level Segmentation	c	5	✗	✓	✗	✗	✓	✗	✗	✗	65.8	70.3
		d	5	✗	✓	✓	✗	✓	✗	✗	✗	65.4	70.6
		e	5	✗	✓	✗	✓	✓	✗	✗	✗	67.2	70.8
	SemAffine	f	5	✗	✓	✗	✓	✗	✓	✗	✗	66.8	71.2
		g	5	✗	✓	✗	✓	✗	✗	✓	✗	68.0	71.8
ISAM	Multi-Modality	h	5	✗	✓	✗	✓	✗	✗	✓	✓	68.2	72.1
SemAffiNet	View Num	i	3	✗	✓	✗	✓	✗	✗	✓	✓	68.3	70.7

and point-based point cloud segmentation methods.

For 2D image segmentation methods, we choose the classical SemanticFPN [31] as the baseline, implementing ResNet-50 and ResNet-101 settings. We conduct semantic segmentation on the Cityscapes [12] dataset and the results are shown in the last two lines of Table 3. The quantitative results show that the proposed ESAM is not restricted within 3D domain and brings consistent improvements to 2D segmentation baseline under different settings.

4.3. Ablation Studies

In order to measure the contribution of each SemAffiNet module, we conduct ablation studies on the ScanNetV2 validation set under 5cm setting. ESAM can be divided into two parts: multi-level segmentation and semantic-affine transformation (SA). For multi-level segmentation, we incrementally add mid-level segmentation ($\hat{\text{M}}$.) besides final-level segmentation ($\hat{\text{F}}$.) and replace fully connected classifier (FC) with mask classifier (M). ISAM module fuses multi-modality information (Fuse). The ablation results are shown in Table 4 and we conclude that each block makes its own contribution to the overall progression, where semantic-affine transformation is the most effective one.

Besides ablations on sub-modules of SemAffiNet, we further explore the superiority of the semantic-affine transformation by replacing it with Adaptive Instance Normalization(AdaIN) [28], whose affine parameters are entirely learned with gradient descent and lack explicit semantic guidance. Comparing line e, f, g in Table 4, AdaIN (line f) brings little improvement compared with the vanilla Batch Normalization (line e). However, semantic-affine transformation (line g) yields more significant progression.

Furthermore, we also conduct ablation experiments on the number of 2D views (VN). According to line a, b, adding views to BPNet [26] baseline doesn't lead to better results but adds computation burden. The reason might be that the potential of the network is exhausted when processing 3 views. However, according to our experiments (line h and i), our SemAffiNet performs better on 3D semantic segmentation when increasing view numbers from 3 to 5. The experimental results show that our SemAffiNet reveals more semantic knowledge and has larger potential.

4.4. Limitations

Even though the proposed semantic-affine transformation is a general representation learning technique to enhance the semantic cognition ability of point clouds features, migrating and performing it on broader point cloud understanding tasks is a non-trivial problem. Since we need per-point supervision on mid-level points, we need to develop weakly-supervised or unsupervised learning techniques when the per-point annotation is inaccessible.

5. Conclusions

In this paper, we propose the semantic-affine transformation to explicitly map mid-level features of the backbone decoder to more semantic-distinct embeddings. Based on this technique, we build a semantic-aware segmentation network SemAffiNet. The ESAM explicitly predicts class masks and regresses semantic-affine parameters via the Transformer decoder with learnable query tokens, while the ISAM fuses multi-modal information via the self-attention mechanism. We conduct experiments with the SemAffiNet on the ScanNetV2 dataset and outperform the previous state-of-the-art BPNet. We also prove the generalization ability of the semantic-affine transformation by wrapping it into a plug-and-play ESAM and evaluating it on both 3D point cloud and 2D image segmentation baselines under various settings. We believe that the semantic-affine transformation will advance related works in the community, given its simple implementation and reasonable insights.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 62125603 and Grant U1813218, and in part by a grant from the Beijing Academy of Artificial Intelligence (BAAI).

References

[1] Eren Erdal Aksoy, Saimir Baci, and Selcuk Cavdar. Salsanet: Fast road and vehicle segmentation in lidar point clouds for autonomous driving. In *IEEE Intelligent Vehicles Symposium*, pages 926–932, 2020. 2

[2] Inigo Alonso, Luis Riazuelo, Luis Montesano, and Ana C Murillo. 3d-mininet: Learning a 2d representation from point clouds for fast and efficient 3d lidar semantic segmentation. *IEEE Robotics and Automation Letters*, 5(4):5432–5439, 2020. 2

[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. In *CVPR*, pages 1534–1543, 2016. 7

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *TPAMI*, 39(12):2481–2495, 2017. 1

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 3

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2017. 1

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[8] Yunlu Chen, Vincent Tao Hu, Efstratios Gavves, Thomas Mensink, Pascal Mettes, Pengwan Yang, and Cees GM Snoek. Pointmixup: Augmentation for point clouds. In *ECCV*, pages 330–345, 2020. 2

[9] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *arXiv*, 2021. 3

[10] Ran Cheng, Ryan Razani, Ehsan Taghavi, Enxu Li, and Bingbing Liu. 2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In *CVPR*, pages 12547–12556, 2021. 2

[11] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, pages 3075–3084, 2019. 1, 2, 6, 7

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 8

[13] Tiago Cortinhal, George Tzelepis, and Eren Erdal Aksoy. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In *ISVC*, pages 207–222, 2020. 2

[14] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 2, 6, 7

[15] Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *ECCV*, pages 452–468, 2018. 1, 6

[16] Shuang Deng and Qiulei Dong. Ga-net: Global attention network for point cloud semantic segmentation. *IEEE Signal Processing Letters*, 2021. 2

[17] 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. *arXiv preprint arXiv:2010.11929*, 2020. 3

[18] Siqi Fan, Qiulei Dong, Fenghua Zhu, Yisheng Lv, Peijun Ye, and Fei-Yue Wang. Scf-net: Learning spatial contextual features for large-scale point cloud segmentation. In *CVPR*, pages 14504–14513, 2021. 2

[19] Jingyu Gong, Jiachen Xu, Xin Tan, Haichuan Song, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Omni-supervised point cloud segmentation via gradual receptive field component reasoning. In *CVPR*, pages 11673–11682, 2021. 2, 6, 7

[20] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, pages 9224–9232, 2018. 1, 2, 6

[21] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, 2021. 3

[22] Ruohao Guo, Dantong Niu, Liao Qu, and Zhenbo Li. Sotr: Segmenting objects with transformers. In *ICCV*, pages 7157–7166, 2021. 3

[23] Ankur Handa, Viorica Patraucean, Vijay Badrinarayanan, Simon Stent, and Roberto Cipolla. Understanding real world indoor scenes with synthetic data. In *CVPR*, pages 4077–4085, 2016. 6

[24] Alexander Hermans, Georgios Floros, and Bastian Leibe. Dense 3d semantic mapping of indoor scenes from rgb-d images. In *ICRA*, pages 2631–2638, 2014. 6

[25] 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. In *CVPR*, pages 11108–11117, 2020. 2

[26] Wenbo Hu, Hengshuang Zhao, Li Jiang, Jiaya Jia, and Tien-Tsin Wong. Bidirectional projection network for cross dimension scene understanding. In *CVPR*, pages 14373–14382, 2021. 2, 3, 6, 8

[27] Jingwei Huang, Haotian Zhang, Li Yi, Thomas Funkhouser, Matthias Nießner, and Leonidas J Guibas. Texturenet: Consistent local parametrizations for learning from high-resolution signals on meshes. In *CVPR*, pages 4440–4449, 2019. 1

[28] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 8

[29] Maximilian Jaritz, Jiayuan Gu, and Hao Su. Multi-view pointnet for 3d scene understanding. In *ICCV Workshops*, pages 0–0, 2019. 6

[30] Lipeng Ke, Ming-Ching Chang, Honggang Qi, and Siwei Lyu. Multi-scale structure-aware network for human pose estimation. In *ECCV*, pages 713–728, 2018. 2

[31] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, pages 6399–6408, 2019. 7, 8

[32] Roman Klokov and Victor Lempitsky. Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. In *ICCV*, pages 863–872, 2017. 2

[33] Ruoyu Li, Sheng Wang, Feiyun Zhu, and Junzhou Huang. Adaptive graph convolutional neural networks. In *AAAI*, volume 32, 2018. 2

[34] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *NeurIPS*, 31:820–830, 2018. 1, 2

[35] Mingyuan Liu, Dan Schonfeld, and Wei Tang. Exploit visual dependency relations for semantic segmentation. In *CVPR*, pages 9726–9735, 2021. 2

[36] 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. *arXiv preprint arXiv:2103.14030*, 2021. 3

[37] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. *arXiv preprint arXiv:1907.03739*, 2019. 2

[38] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 5

[39] Tao Lu, Limin Wang, and Gangshan Wu. Cga-net: Category guided aggregation for point cloud semantic segmentation. In *CVPR*, pages 11693–11702, 2021. 2

[40] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IROS*, pages 922–928, 2015. 2

[41] John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *ICRA*, pages 4628–4635, 2017. 6

[42] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet++: Fast and accurate lidar semantic segmentation. In *IROS*, pages 4213–4220, 2019. 2

[43] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2, 6

[44] Alexey Nekrasov, Jonas Schult, Or Litany, Bastian Leibe, and Francis Engelmann. Mix3d: Out-of-context data augmentation for 3d scenes. *arXiv preprint arXiv:2110.02210*, 2021. 2, 6, 7

[45] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017. 1, 2

[46] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017. 1, 2, 6

[47] Shi Qiu, Saeed Anwar, and Nick Barnes. Semantic segmentation for real point cloud scenes via bilateral augmentation and adaptive fusion. In *CVPR*, pages 1757–1767, 2021. 2

[48] Tran Minh Quan, David GC Hildebrand, and Won-Ki Jeong. Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics. *arXiv preprint arXiv:1612.05360*, 2016. 2

[49] Dario Rethage, Johanna Wald, Jurgen Sturm, Nassir Navab, and Federico Tombari. Fully-convolutional point networks for large-scale point clouds. In *ECCV*, pages 596–611, 2018. 1

[50] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *CVPR*, pages 3577–3586, 2017. 2

[51] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 1, 5

[52] Hang Su, Varun Jampani, Deqing Sun, Subhransu Maji, Evangelos Kalogerakis, Ming-Hsuan Yang, and Jan Kautz. Splatnet: Sparse lattice networks for point cloud processing. In *CVPR*, pages 2530–2539, 2018. 1, 2

[53] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *ECCV*, pages 685–702, 2020. 2

[54] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *ICCV*, pages 6411–6420, 2019. 1, 2, 6, 7

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2, 3

[56] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *TOG*, 38(5):1–12, 2019. 1, 2

[57] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, pages 4724–4732, 2016. 2

[58] Bichen Wu, Alvin Wan, Xiangyu Yue, and Kurt Keutzer. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In *ICRA*, pages 1887–1893, 2018. 2

[59] Bichen Wu, Xuanyu Zhou, Sicheng Zhao, Xiangyu Yue, and Kurt Keutzer. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *ICRA*, pages 4376–4382, 2019. 2

[60] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *CVPR*, pages 9621–9630, 2019. 1, 2, 6

[61] Mutian Xu, Runyu Ding, Hengshuang Zhao, and Xiaojuan Qi. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *CVPR*, pages 3173–3182, 2021. 2

[62] Xu Yan, Chaoda Zheng, Zhen Li, Sheng Wang, and Shuguang Cui. Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In *CVPR*, pages 5589–5598, 2020. 6

[63] Maosheng Ye, Shuangjie Xu, Tongyi Cao, and Qifeng Chen. Drinet: A dual-representation iterative learning network for point cloud segmentation. In *ICCV*, pages 7447–7456, 2021. 2

[64] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *ICCV*, pages 12498–12507, 2021. 3

[65] Jinlai Zhang, Lyujie Chen, Bo Ouyang, Binbin Liu, Jihong Zhu, Yujing Chen, Yanmei Meng, and Danfeng Wu. Pointcutmix: Regularization strategy for point cloud classification. *arXiv preprint arXiv:2101.01461*, 2021. 2

[66] Jiazhao Zhang, Chenyang Zhu, Lintao Zheng, and Kai Xu. Fusion-aware point convolution for online semantic 3d scene segmentation. In *CVPR*, pages 4534–4543, 2020. 6

[67] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *CVPR*, pages 9601–9610, 2020. 1, 2

[68] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *ICCV*, pages 16259–16268, 2021. 3

[69] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 1

[70] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, pages 6881–6890, 2021. 3

[71] Hui Zhou, Xinge Zhu, Xiao Song, Yuexin Ma, Zhe Wang, Hongsheng Li, and Dahua Lin. Cylinder3d: An effective 3d framework for driving-scene lidar semantic segmentation. *arXiv preprint arXiv:2008.01550*, 2020. 1, 2