Supplementary Materials for DeepLA-Net: Very Deep Local Aggregation Networks for Point Cloud Analysis

Overview

This supplementary material is organized as follows:

- Section A provides the details of the network architecture.
- Section B presents the experimental settings.
- Section C provides additional discussion including.
- Section D shows detailed and additional experimentation results for semantic segmentation.
- Section E introduces more related works including.
- Section F shows outlook for the future work.

A. Details of the Network Architecture

We provide detailed network architectures for segmentation and classification. As illustrated in Figure 1, the encoder comprises four encoding stages, and each encoding stage consists of a down-sampling operation and multiple ResLFE blocks, with the output being supervised by HDS strategy. In the segmentation branch, DeepLA-Net follows an encoder-decoder architecture. Each decoding stage comprises an up-sampling operation and a multi-layer perceptron. In the classification branch, global average pooling is applied to the output of the encoder to obtain the global representation. Finally, fully-connected layers with a softmax are used to predict the classification scores, where the segmentation/classification results are dictated by the label with the highest score.

B. Experimental Settings

B.1. Evaluation Metrics

To quantitatively analyze the performance of the proposed architecture, overall accuracy (OA), mean Accuracy (mAcc), per-class intersection over union (IoUs), mean IoU (mIoU), are used as evaluation metrics as follows:

$$\mathrm{OA} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{TP_i}}{\mathrm{N}} \tag{1}$$

$$\mathrm{mAcc} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{Acc_i}}{\mathrm{n}} \tag{2}$$

$$\mathrm{IoU_i} = \frac{\mathrm{TP_i}}{\mathrm{TP_i} + \mathrm{FP_i} + \mathrm{FN_i}} \tag{3}$$

$$\mathrm{mIoU} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{IoU_i}}{\mathrm{n}} \tag{4}$$

where TP denotes the number of true positive samples, FP denotes the number of false positive samples, FN denotes the number of false negative samples, i denotes the i_{th} semantic class, n denotes the number of total semantic classes and N denotes the number of total points.

B.2. Implementation Details

During the training process, we use the hybrid deep supervision strategy with label smoothing to optimize our models . We adopt the AdamW optimizer [15] with an initial learning rate of 0.004, and a scheduler with weight decay of 10^{-4} using cosine learning rate decay. For data augmentation, we use random scaling, feature dropping, and color auto contrasting whenever applicable, following [28]. For semantic segmentation, we input a fixed number of 30,000 points per batch, with a batch size of 8, and train for 100 epochs. For object classification, we input a fixed number of 1,024 points per batch, with a batch size of 32, and train for 250 epochs. For part segmentation, we input a fixed number of 2,048 points per batch, with a batch size of 32, and train for 250 epochs. In the down-sampling process, for object classification and part segmentation, we employ farthest point sampling, retaining only half of the remaining points at each stage. For semantic segmentation, we employ grid sampling with linear time complexity. The initial grid size is set to 0.04m for S3DIS and 0.02m for ScanNet v2, and doubled at each stage.

B.3. Dataset Description

For semantic segmentation, we conduct experiments on S3DIS [1] and ScanNet v2 [6]. The S3DIS comprises 272 rooms from six large-scale indoor areas. Each point is annotated with a specific semantic label from 13 classes. The ScanNet v2 comprises 1,513 room scans reconstructed from RGB-D frames. The dataset is divided into 1,201 scenes for training, 312 for validation and 100 for online testing. Each point is annotated with a specific semantic label from 20 classes. We use mIoU to assess performance on both datasets: 6-fold cross-validation and Area 5 for S3DIS, and validation and online test sets for ScanNet v2.

For object classification, we conduct experiments on ScanObjectNN [37]. The ScanObjectNN contains about 15,000 real scanned objects, each annotated with a semantic label from 15 classes. Due to the existence of back-

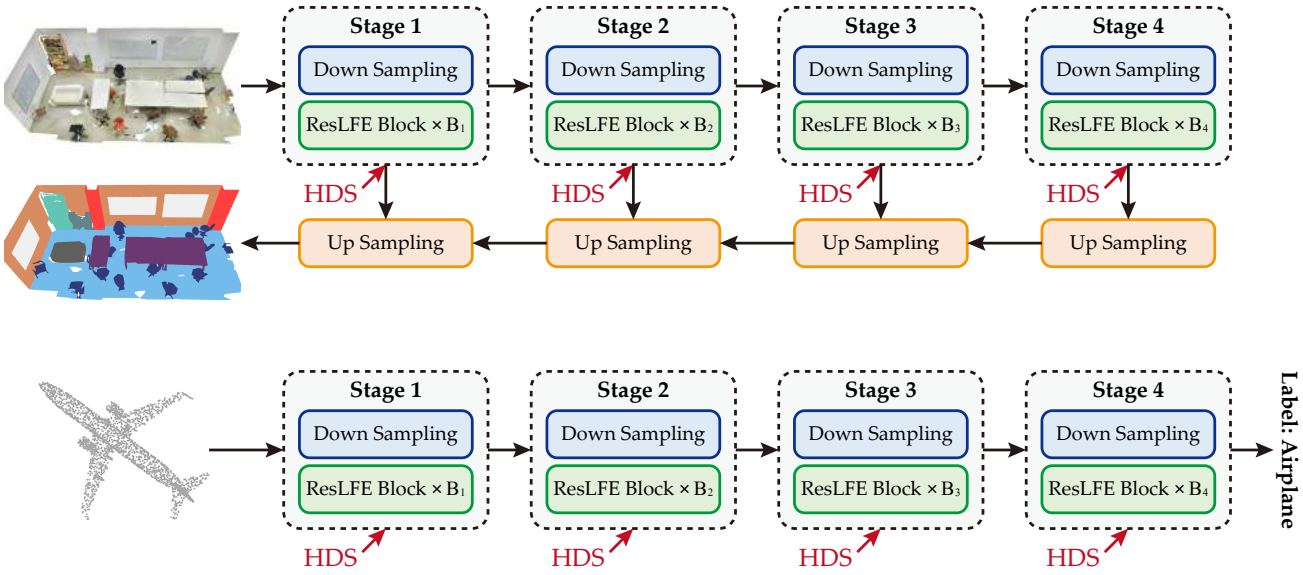

Figure 1. Architecture of the proposed DeepLA-Net for segmentation (top) and classification (bottom). **HDS** denotes hybrid deep supervision strategy. B_n denotes the number of blocks in the n-th stage.

Table 1. Ablation results of the bottleneck in ResLFE block in DeepLA-24 on S3DIS Area5.

Block Ratio	mIoU (%)	△ (%)
[1:1:3:1]	**73.2**	-
[1:1:9:1]	72.7	−0.5
[1:1:1:1]	72.5	−0.7
[5:9:5:5]	72.2	−1.0

ground elements, noise, and occlusions, ScanObjectNN poses significant challenges to the existing point cloud analysis methods. Following PointMLP [23] and PointNeXt [28], we conduct experiments on PB_T50_RS, the hardest and most commonly used variant of ScanObjectNN.

For part segmentation, we conduct experiments on ShapeNetPart [46]. The ShapeNetPart provides part-level annotation for 3D models, comprising 16,880 models across 16 distinct shape classes. Each class has 2-6 parts, amounting to a total of 50 part labels.

C. Additional discussion

C.1. Analysis on ResLFE Block Ratio

In the DeepLA-Net implementation, we set the ResLFE block ratio in encoder stages of [1:1:3:1]. As shown in Table 1, we implement different ResLFE block ratio on DeepLA-24. It is evident that the [1:1:3:1] block ratio we used achieves the best performance.

C.2. More Visual Comparison of Feature Learning with Different Network Depths

We present the visualization of the feature similarity matrix for a specific object class in 3D scenes in Figure 2. From the visualization results, it is evident that DeepLA-120 demonstrates clear segmentation boundaries. While DeepLA-24 is effective, it displays some blurred edges and occasional recognition errors. The simplest DeepLA-6 exhibits a significant number of recognition errors. These findings highlight that the feature similarity matrix of deeper LANets is more accurate and reliable, revealing more pronounced in feature differences compared to surrounding objects. This further demonstrates the enhanced capability of deep networks in feature learning, indicating that a reasonable deepening of LANets can significantly improve its ability to capture local patterns, including edge segmentation and object recognition.

C.3. Exploring to Deeper Networks

We explore aggressively deeper networks of 240 and 360 blocks. These networks are trained and tested on a single Nvidia A6000 GPU with 48GB memory, while keeping other settings consistent with DeepLA-120. As shown in Figure 3, we observe that DeepLA-240/360 exhibit better training accuracy, indicating the potential benefits of further deepening networks. However, the test results of DeepLA-240/360 are inferior to our DeepLA-120, as detailed in Table 2. We attribute this discrepancy to overfitting. Given that point cloud data can be challenging to acquire and an-

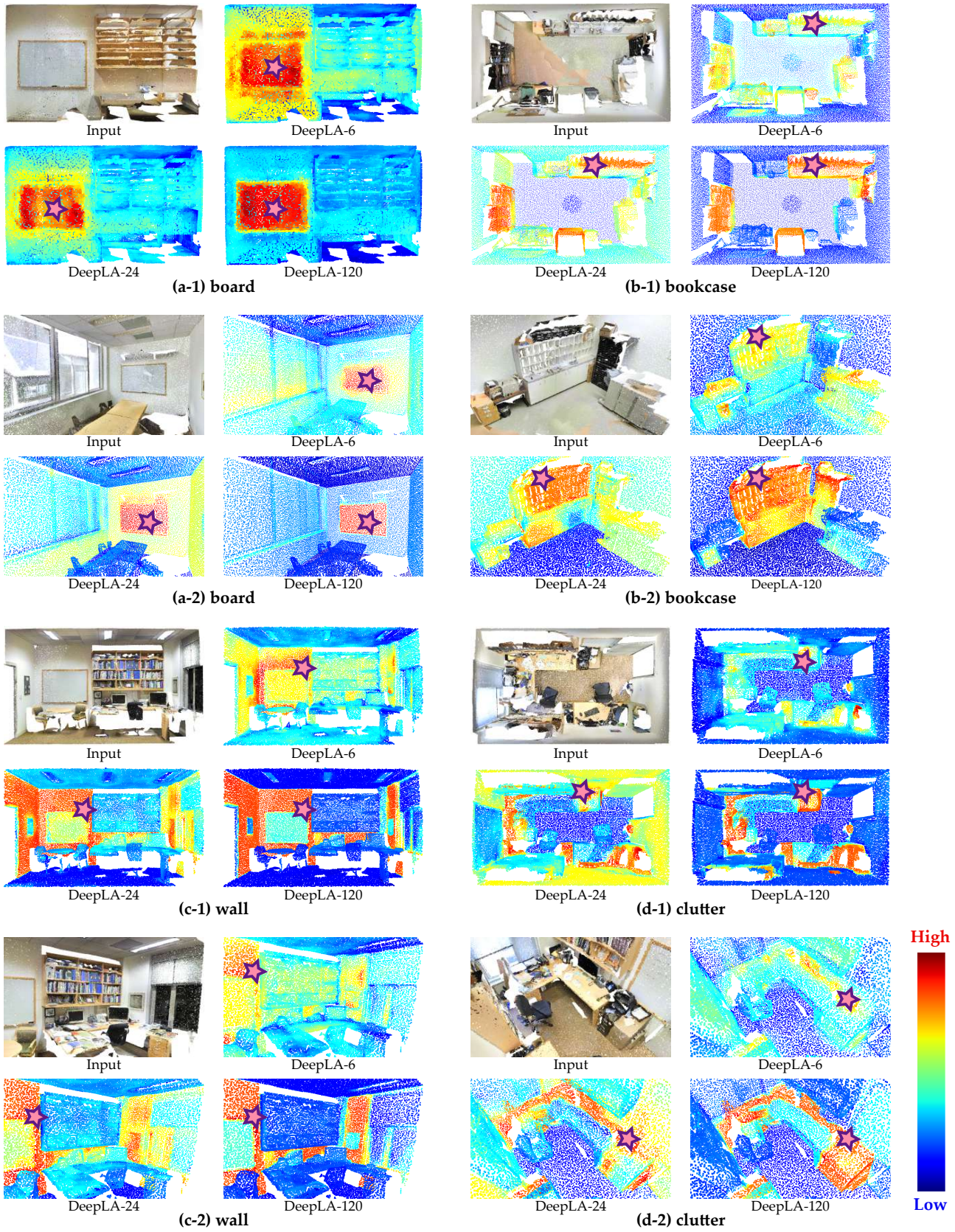

Figure 2. Visual comparison of feature similarity matrix for a specific object class predicted by DeepLA-Net of different depths. The pink stars illustrate the selected center points

notate, DeepLA-240/360 may be excessively large, potentially necessitating additional strong regularization and data augmentation methods for improved outcomes. We plan to further investigate this in future work.

Table 2. Quantitative comparisons of performance, model complexity, and latency on S3DIS Area5.

Method	mIoU (%)	Params. (M)	FLOPs (G)	Thr. Put (ins./sec.)
DeepLA-120	**75.7**	30.3	42.7	42
DeepLA-240	74.5	61.2	83.4	23
DeepLA-360	74.2	90.7	134.6	15

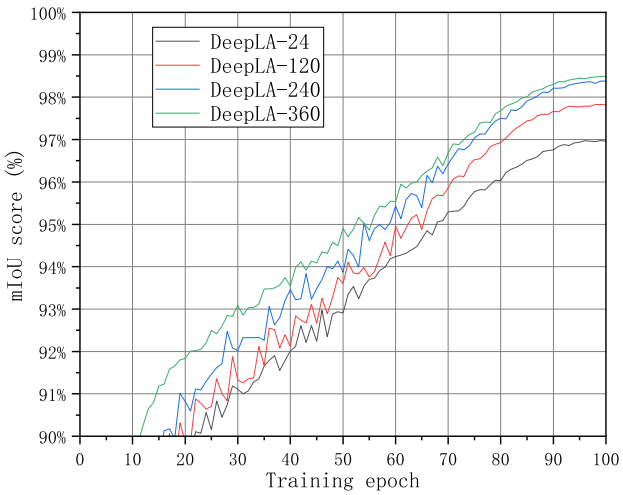

Figure 3. Training performance (mIoU score) across training epochs for deeper DeepLA-Net family on S3DIS (Area 5).

D. Additional Semantic Segmentation Results

D.1. Quantitative Comparisons

In this section, we demonstrate the per-class IoU for S3DIS Area5 (Table 5), 6-fold (Table 6), and ScanNet v2 test set (Table 7). Note that, since many methods do not show the detailed per-class IoU in the semantic segmentation task, here we only compare the methods that present per-class IoU in their papers or have released their code and model weights. For per-class IoU on the S3DIS Area5 and 6-fold, we observe that DeepLA-120 achieves the best or sub-best performance in almost all classes. This demonstrates the potential of DeepLA-Net in local pattern acquisition. Meanwhile, the DeepLA-Net family performs competitively on large-scale objects such as walls, columns, and windows. We conjecture that DeepLA-Net can obtain long-range information with further deepening of the network. Similarly, the proposed DeepLA-120 achieves best or sub-best performance in most classes in per-class IoU on the ScanNet v2 dataset, underscoring the generalizability of DeepLA-Net family.

D.2. Qualitative Comparisons

In this section, for a more perceptible comparison between various methods, we qualitatively assessed the semantic segmentation outcomes produced by PointVector-XL [7] (the best model of PointVector family) and our DeepLA-120 on S3DIS and ScanNet v2 (validation set), as illustrated in Figure 4 and Figure 5. The red boxes highlight regions where the segmentation is inaccurate or the boundary is inconspicuous in PointVector. For the S3DIS dataset, it is visually evident that our segmentation of *clutter*, *columns*, *boards*, and *bookcases* is superior to that of PointVector. These classes are challenging since they usually looks very similar to the *wall*. For example, the board, column and wall in the last row of Figure 4 have slightly different geometric shapes from one another, requiring the network to model long-range dependencies. For the ScanNet v2 dataset, the proposed DeepLA-120 can segment the boundaries more smoothly and accurately.

D.3. Outdoor Datasets

We conducted experiments on the validation sets of the outdoor datasets SemanticKITTI [2] and nuScenes [3]. Our DeepLA-120 surpasses the previous SOTA point-based method PTv3 [39] (without additional data for pre-train) in both mIoU and FPS, while DeepLA-24 achieves real-time processing (FPS>24). This demonstrates the generalizability of our approach. We will show details in the revision, and will complement more tasks including instance segmentation, and object detection in the extended version.

Table 3. Quantitative comparisons of mIoU and FPS with PTv3 on SemanticKITTI and nuScenes.

Method	SemanticKITTI	nuScenes	FPS
DeepLA-24	66.2	75.1	**31**
DeepLA-120	**71.0**	**80.8**	14
PT v3 (*w/o pretrain*)	70.8	80.4	10

D.4. Discussion with PTv3 in ScanNet

Unlike our method on ScanNet v2 [6], PTv3 [39] relied on additional data for pre-train. More importantly, PTv3's open-source code reveals the use of extensive test-time augmentation (TTA), which can significantly boost performance. As highlighted in our paper, we did not utilize TTA. To ensure a fair evaluation, we disabled TTA during the testing phase of PTv3. In this case, PTv3 achieves an mIoU of only 76.3% on the validation set **using their provided weights**, which is lower than our DeepLA-120 (77.6%).

Table 4. Quantitative comparisons with PTv3 on ScanNet v2.

Method	ScanNet val
DeepLA-120	**77.6**
PT v3 (*w/o pretrain*)	77.5
– w/o TTA	76.3

E. Additional Related Works

E.1. Deep Neural Network Architecture

In the field of 2D image processing, CNNs have been deepening since the introduction of the pioneering AlexNet [16], leading to a continuous enhancement in network fitting capability. VGG [32] builds upon AlexNet by stacking small-sized convolution filters, significantly increasing network depth and substantially improving performance. Following this, ResNet [12] introduces a simple and efficient skip connection, making it possible to further deepen the network layers. The great success of ResNet not only demonstrates the effectiveness of reasonably increasing network depth, but also inspires subsequent researches to the application and exploration of deep neural architecture [10, 11, 14, 24, 42, 43].

In the field of 3D point cloud processing, researchers have largely 'avoided' exploring network depth, primarily constrained by the historical philosophy of designing networks with more complex local representation. For example, ASSANet [27] also uses pre-linear, which is also employed in our DeepLA-Net, it has an extremely complex design for the local aggregation module with 118M parameters (ASSANet-L only with 8 blocks). In contrast, our approach avoids such complex and redundant design and thus the DeepLA-24 only has 6M parameters. Although some recent works [7, 22, 23, 28, 47] incrementally increased the depth of their networks (about 10-20 blocks), these designs essentially aim to increase the number of parameters for scale-up. In this paper, instead of deliberately following the prevailing trend in the 3D vision community of exploring sophisticated details, we pursue an empirically powerful and very deep architecture for point cloud analysis.

E.2. Deep Supervision

Deep supervision is initially proposed to address the issues of gradient vanishing and slow convergence speed during the network training [18, 34]. This effective training technique has also been applied to improve performance [5, 33, 36, 48, 53]. Lee et al. [18] demonstrate that deep supervised layers can enhance the learning capabilities of hidden layers. This encourages intermediate layers to learn discriminative features, thereby enabling faster convergence and regularization of the network. Dou et al. [8] introduce a deep supervision paradigm to address optimization challenges by supervising predictions from feature maps at varying resolutions. Deep supervision can also be employed to deepen networks. Wang et al. [38] employ a gradient-based heuristic approach to enhance gradient propagation for the training of deeper neural networks. Zhang et al. [51] employ cross-entropy loss to supervise feature maps at different scales in ResNet-50, ensuring the precise capture of context and global information in deeper neural networks. Building on these insights, we have constructed very deep LANets enhanced with deep supervision, to ensure smooth gradient backpropagation in deep networks and to mitigate training optimization challenges.

F. Future Work

Despite the encouraging results, this paper still serves as a start-up work on very deep LANets. Due to the high costs associated with acquiring and annotating point cloud data, the scale of available datasets is significantly smaller compared to 2D images. The limited scale means that deep networks might be more prone to overfitting when applied to point cloud data. In future work, we plan to delve into regularization strategies for DeepLA-Net. Additionally, recent studies are exploring pre-trained models for 3D point clouds. Integrating DeepLA-Net with 3D pre-training strategy could be a promising direction for future research. Our work has the potential to contribute significantly to the development of 3D foundation models.

G. Acknowledgments

This study is supported by the State Key Program of National Natural Science Foundation of China (52332010), the National Natural Science Foundation of China (42471480), the Major Program (JD) of Hubei Province (2023AA02604). The numerical calculations in this article have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

Table 5. Quantitative comparisons with the state-of-the-art methods on S3DIS Area5. **Bold** indicates the best result, underline indicates the best result excluding ours. We only report methods which have demonstrated per-class IoU in their papers.

Method	OA	mIoU	ceil.	floor	wall	beam	col.	wind.	door	table	chair	sofa	book.	board	clut.
PointNet [25]	-	49.0	88.8	97.3	69.8	0.0	3.9	46.3	10.8	59.0	52.6	5.9	40.3	26.4	33.2
PointCNN [20]	85.9	57.3	92.3	98.2	79.4	0.0	17.6	22.8	62.1	74.4	80.6	31.7	66.7	62.1	56.7
KPConv [35]	-	67.1	92.8	97.3	82.4	0.0	23.9	58.0	69.0	81.5	91.0	75.4	75.3	66.7	58.9
PointASNL [44]	87.7	62.6	94.3	98.4	79.1	0.0	26.7	55.2	66.2	83.3	86.8	47.6	68.3	56.4	52.1
RandLA-Net [13]	87.2	62.5	92.1	97.3	80.9	0.0	21.4	61.4	37.4	78.3	87.1	65.8	70.4	67.7	52.2
Point Trans. [52]	90.8	70.4	94.0	98.5	86.3	0.0	38.0	63.4	74.3	89.1	82.4	74.3	80.2	76.0	59.3
GSLCN [21]	90.5	68.1	94.3	98.5	82.9	0.0	20.6	59.4	69.8	83.1	91.4	76.9	75.4	72.5	60.7
PointNeXt [28]	90.6	70.5	94.2	98.5	84.4	0.0	37.7	59.3	74.0	83.1	91.6	77.4	77.2	78.8	60.6
Stra. Trans. [17]	91.5	72.0	96.2	98.7	85.6	0.0	46.1	60.0	76.8	**92.6**	84.5	77.8	75.2	78.1	64.0
PointVector [7]	91.6	72.6	95.6	98.6	85.9	0.0	40.1	61.9	76.4	84.9	**92.4**	**80.9**	78.5	**84.4**	64.6
PointMeta [22]	91.3	72.2	95.4	98.6	85.0	0.0	44.1	61.2	79.0	83.7	92.0	80.8	77.8	78.4	63.2
(Ours) DeepLA-24	91.6	73.2	94.6	98.3	86.9	0.0	48.4	65.5	79.7	88.0	91.1	78.9	77.4	78.9	64.2
(Ours) DeepLA-60	92.0	74.8	95.9	98.6	87.7	0.0	50.2	67.5	**86.0**	90.5	91.8	79.1	78.4	80.3	65.2
(Ours) DeepLA-120	**92.6**	**75.7**	**96.4**	**98.9**	**88.5**	0.0	**53.3**	**71.4**	82.7	92.1	92.2	78.0	**82.0**	81.5	**66.9**

Table 6. Quantitative comparisons with the state-of-the-art methods on S3DIS (6-fold). **Bold** indicates the best result, underline indicates the best result excluding ours. We only report methods which have demonstrated per-class IoU in their papers.

Method	OA	mIoU	ceil.	floor	wall	beam	col.	wind.	door	table	chair	sofa	book.	board	clut.
PointNet [25]	78.6	47.6	88.0	88.7	69.3	42.4	23.1	47.5	51.6	54.1	42.0	9.6	38.2	29.4	35.2
PointCNN [20]	88.1	65.4	94.2	97.3	75.8	63.3	51.7	58.4	57.2	71.6	69.1	39.1	61.2	52.2	58.6
KPConv [35]	-	70.6	93.6	92.4	83.1	63.9	54.3	66.1	76.6	57.8	64.0	69.3	74.9	61.3	60.3
RandLA-Net [13]	88.0	70.0	93.1	96.1	80.6	62.4	48.0	64.4	69.4	69.4	76.4	60.0	64.2	65.9	60.1
BAAF-Net [29]	88.9	72.2	93.3	96.8	81.6	61.9	49.5	65.4	73.3	72.0	83.7	67.5	64.3	67.0	62.4
LEARD-Net [49]	89.1	72.5	94.2	96.9	81.8	65.1	50.9	69.9	72.5	70.6	78.2	68.6	67.2	66.1	60.3
LACV-Net [50]	89.7	72.7	94.5	96.7	82.1	65.2	48.6	69.3	71.2	72.7	78.1	67.3	67.2	70.9	61.6
PointTrans. [52]	90.2	73.5	94.3	97.5	84.7	55.6	58.1	66.1	78.2	77.6	74.1	67.3	71.2	65.7	64.8
U-Next [48]	89.5	73.2	93.6	96.9	84.2	66.1	54.6	67.6	75.5	73.6	74.5	62.9	66.2	74.0	61.7
DeepViewAgg. [31]	-	74.7	90.0	96.1	85.1	66.9	56.3	71.9	78.9	79.7	73.9	69.4	61.1	75.0	65.9
PointNeXt [28]	90.3	74.8	94.2	96.8	85.0	61.5	64.2	68.5	78.7	77.0	70.1	72.4	70.9	70.3	63.3
SPTrans. [30]	-	76.0	93.9	96.3	84.3	71.4	61.3	70.1	78.2	**84.6**	74.1	67.8	**77.1**	63.6	65.0
PointVector[7]	91.8	78.4	95.3	97.5	86.2	64.8	65.2	69.5	81.6	77.8	**89.3**	75.6	72.2	73.9	**70.2**
PointMeta [22]	91.4	77.0	94.9	97.6	85.6	64.4	62.8	68.2	82.1	77.1	83.8	75.4	71.1	70.1	68.5
(Ours) DeepLA-24	91.4	77.9	94.2	96.9	87.0	74.5	68.5	72.5	80.4	76.4	76.9	77.0	71.3	71.3	65.7
(Ours) DeepLA-60	91.9	79.0	94.8	97.6	88.2	**76.2**	69.9	73.6	**82.7**	78.0	77.6	**78.1**	72.1	71.8	66.8
(Ours) DeepLA-120	**92.3**	**79.8**	**95.5**	**97.8**	**89.5**	75.0	**70.3**	**74.8**	82.3	77.2	78.1	77.3	75.1	**75.7**	69.2

Table 7. Quantitative comparisons with the state-of-the-art methods on ScanNet v2 (test set). **Bold** indicates the best result, underline indicates the best result excluding ours. We only report methods which have demonstrated per-class IoU in their papers.

Method	mIoU	bathtub	bed	bookshelf	cabinet	chair	counter	curtain	desk	door	floor	other furniture	picture	refrigerator	shower curtain	sink	sofa	table	toilet	wall	window
PointNet++ [26]	55.7	73.5	66.1	68.6	49.1	74.4	39.2	53.9	45.1	37.5	94.6	37.6	20.5	40.3	35.6	55.3	64.3	49.7	82.4	75.6	51.5
KPConv [35]	68.4	84.7	75.8	78.4	64.7	81.4	47.3	77.2	60.5	59.4	93.5	45.0	18.1	58.7	80.5	69.0	78.5	61.4	88.2	81.9	63.2
PointASNL [44]	66.6	70.3	78.1	75.1	65.5	83.0	47.1	76.9	47.4	53.7	95.1	47.5	27.9	63.5	69.8	67.5	75.1	55.3	81.6	80.6	70.3
RandLA-Net [13]	64.5	77.8	73.1	69.9	57.7	82.9	44.6	73.6	47.7	52.3	94.5	45.4	26.9	48.4	74.9	61.8	73.8	59.9	82.7	79.2	62.1
Stra. Trans. [17]	74.7	90.1	80.3	84.5	75.7	84.6	51.2	82.5	69.6	64.5	95.6	57.6	26.2	74.4	86.1	74.2	77.0	70.5	89.9	86.0	73.4
Point Trans. v2 [40]	75.2	74.2	80.9	**87.2**	75.8	**86.0**	55.2	89.1	61.0	**68.7**	96.0	55.9	30.4	76.6	**92.6**	76.7	79.7	64.4	**94.2**	87.6	72.2
PointMeta [22]	71.4	83.5	78.5	82.1	68.4	84.6	53.1	86.5	61.4	59.6	95.3	50.0	24.6	67.4	88.8	69.2	76.4	62.4	84.9	84.4	67.5
LargeKernel3D [4]	73.9	90.9	82.0	80.6	74.0	85.2	54.5	82.6	59.4	64.3	95.5	54.1	26.3	72.3	85.8	77.5	76.7	67.8	93.3	84.8	69.4
LRPNet [19]	74.2	81.6	80.6	80.7	75.2	82.8	**57.5**	83.9	**69.9**	63.7	95.4	52.0	32.0	75.5	83.4	76.0	77.2	67.6	91.5	86.2	71.7
Retro-FPN [41]	74.4	84.2	80.0	76.7	74.0	83.6	54.1	**91.4**	67.2	62.6	95.8	55.2	27.2	77.7	88.6	69.6	80.1	67.4	94.1	85.8	71.7
DMF-Net[45]	75.2	90.6	79.3	80.2	68.9	82.5	55.6	86.7	68.1	60.2	96.0	55.5	**36.5**	**77.9**	85.9	74.7	79.5	**71.7**	91.7	85.6	76.4
CondaFormer [9]	75.5	92.7	82.2	83.6	**80.1**	84.9	51.6	86.4	65.1	68.0	95.8	**58.4**	28.2	75.9	85.5	72.8	80.2	67.8	88.0	87.3	75.6
(Ours) DeepLA-120	**77.2**	**93.9**	**82.4**	85.4	77.1	84.0	56.4	90.0	68.6	67.7	**96.1**	53.7	34.8	76.9	90.3	**78.5**	**81.5**	67.6	93.9	**88.0**	**77.2**

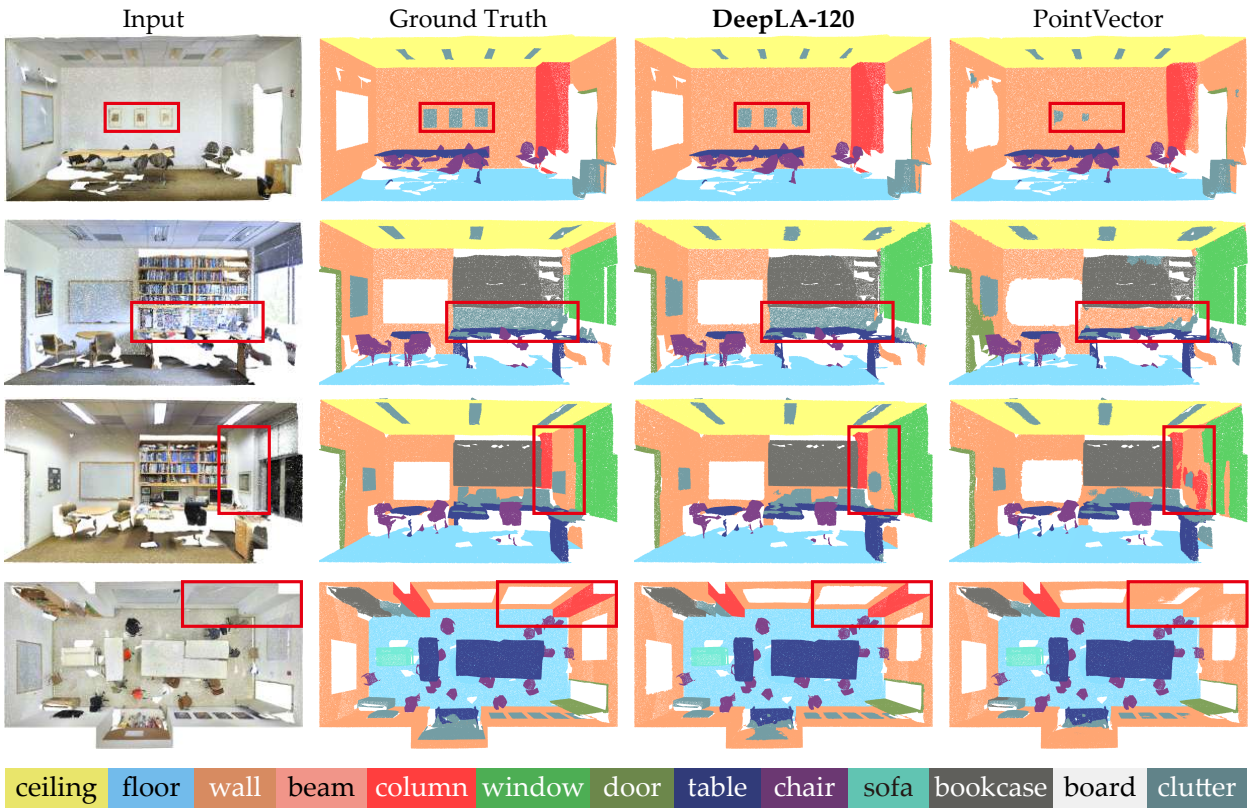

Figure 4. Visual comparison of semantic segmentation results on S3DIS dataset.

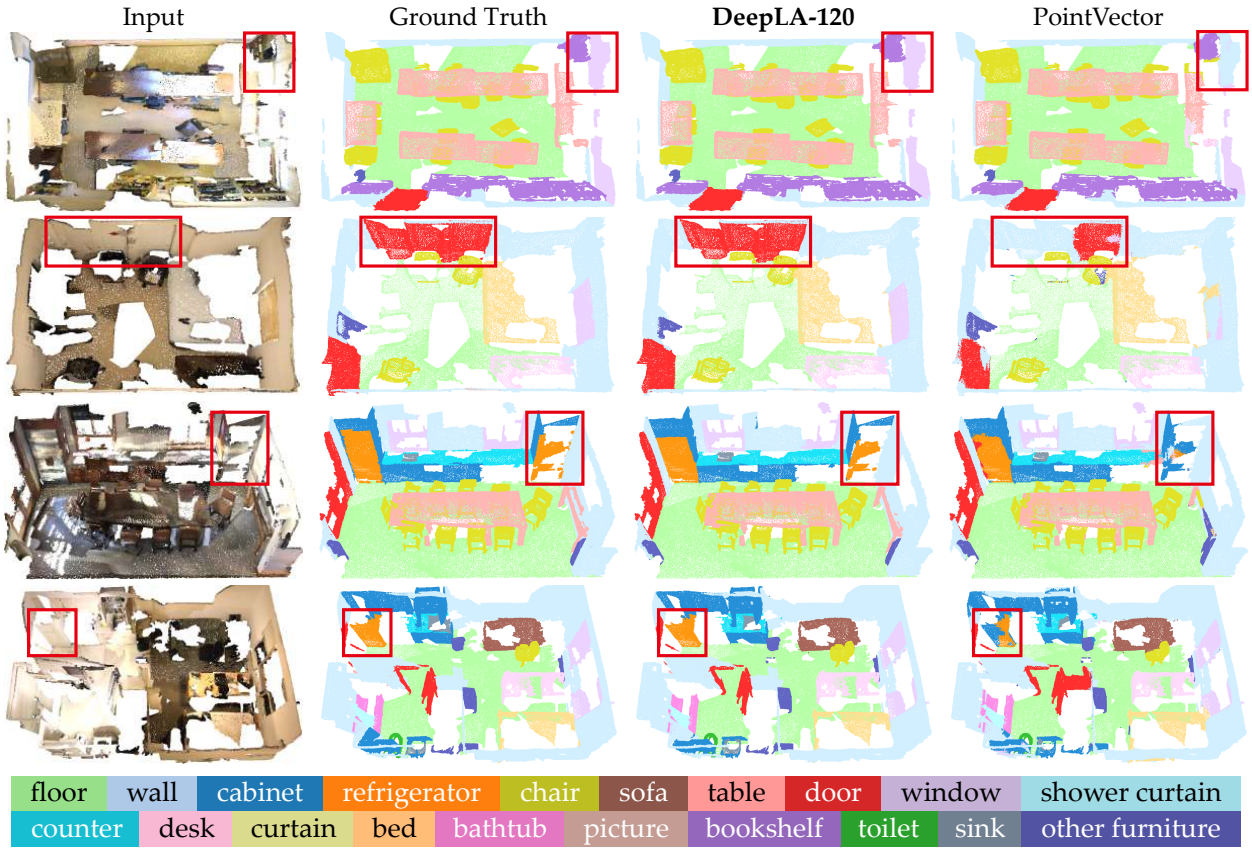

Figure 5. Visual comparison of semantic segmentation results on ScanNet v2 dataset.

References

[1] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 13

[2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In IEEE/CVF International Conference on Computer Vision (ICCV), 2019. 16

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 16

[4] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Largekernel3d: Scaling up kernels in 3d sparse cnns. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 18

[5] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In IEEE/CVF International Conference on Computer Vision (ICCV), 2017. 17

[6] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 13, 16

[7] Xin Deng, WenYu Zhang, Qing Ding, and XinMing Zhang. Pointvector: A vector representation in point cloud analysis. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 16, 17, 18

[8] Qi Dou, Hao Chen, Yueming Jin, Lequan Yu, Jing Qin, and Pheng-Ann Heng. 3D deeply supervised network for automatic liver segmentation from ct volumes. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 2016. 17

[9] Lunhao Duan, Shanshan Zhao, Nan Xue, Mingming Gong, Xia Gui-Song, and Dacheng Tao. Condaformer: Disassembled transformer with local structure enhancement for 3d point cloud understanding. In Neural Information Processing Systems (NeurIPS), 2023. 18

[10] Steven Guan, Amir A Khan, Siddhartha Sikdar, and Parag V Chitnis. Fully dense unet for 2-D sparse photoacoustic tomography artifact removal. IEEE Journal of Biomedical and Health Informatics, 24(2):568–576, 2019. 17

[11] Dongyoon Han, Sangdoo Yun, Byeongho Heo, and YoungJoon Yoo. Rethinking channel dimensions for efficient model design. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 17

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 17

[13] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 18

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 17

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations, 2015. 13

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Neural Information Processing Systems (NeurIPS), 2012. 17

[17] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 18

[18] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In Artificial Intelligence and Statistics, 2015. 17

[19] Xiang-Li Li, Meng-Hao Guo, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Long range pooling for 3d large-scale scene understanding. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 18

[20] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. Neural Information Processing Systems (NeurIPS), 31, 2018. 18

[21] Jiye Liang, Zijin Du, Jianqing Liang, Kaixuan Yao, and Feilong Cao. Long and short-range dependency graph structure learning framework on point cloud. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(12):14975–14989, 2023. 18

[22] Haojia Lin, Xiawu Zheng, Lijiang Li, Fei Chao, Shanshan Wang, Yan Wang, Yonghong Tian, and Rongrong Ji. Meta architecture for point cloud analysis. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 17, 18

[23] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. In International Conference on Learning Representations (ICLR), 2022. 14, 17

[24] Gregory Naitzat, Andrey Zhitnikov, and Lek-Heng Lim. Topology of deep neural networks. The Journal of Machine Learning Research, 21(1):7503–7542, 2020. 17

[25] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3D classification and segmentation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 18

[26] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In Neural Information Processing Systems (NeurIPS), 2018. 18

[27] Guocheng Qian, Hasan Hammoud, Guohao Li, Ali Thabet,

and Bernard Ghanem. Assanet: An anisotropical separable set abstraction for efficient point cloud representation learning. In Neural Information Processing Systems (NeurIPS), 2021. 17

[28] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Abed Al Kader Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. Neural Information Processing Systems (NeurIPS), 35:23192–23204, 2022. 13, 14, 17, 18

[29] Shi Qiu, Saeed Anwar, and Nick Barnes. Semantic segmentation for real point cloud scenes via bilateral augmentation and adaptive fusion. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 18

[30] Damien Robert, Hugo Raguet, and Loic Landrieu. Efficient 3d semantic segmentation with superpoint transformer. In IEEE/CVF International Conference on Computer Vision (ICCV), 2023. 18

[31] Damien Robert, Bruno Vallet, and Loic Landrieu. Learning multi-view aggregation in the wild for large-scale 3d semantic segmentation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 18

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In IAPR Asian Conference on Pattern Recognition, 2015. 17

[33] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 17

[34] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 17

[35] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas Guibas. Kpconv: Flexible and deformable convolution for point clouds. In IEEE/CVF International Conference on Computer Vision (ICCV), 2019. 18

[36] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In IEEE/CVF International Conference on Computer Vision (ICCV), 2015. 17

[37] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In IEEE/CVF International Conference on Computer Vision (ICCV), 2019. 13

[38] Liwei Wang, Chen-Yu Lee, Zhuowen Tu, and Svetlana Lazebnik. Training deeper convolutional networks with deep supervision. arXiv preprint arXiv:1505.02496, 2015. 17

[39] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 4840–4851, 2024. 16

[40] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In Neural Information Processing Systems (NeurIPS), 2022. 18

[41] Peng Xiang, Xin Wen, Yu-Shen Liu, Hui Zhang, Yi Fang, and Zhizhong Han. Retro-fpn: Retrospective feature pyramid network for point cloud semantic segmentation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 18

[42] Xiao Xiao, Shen Lian, Zhiming Luo, and Shaozi Li. Weighted res-unet for high-quality retina vessel segmentation. In International Conference on Information Technology in Medicine and Education, 2018. 17

[43] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 17

[44] Xu Yan, Chaoda Zheng, Zhen Li, Sheng Wang, and Shuguang Cui. Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 18

[45] Chaolong Yang, Yuyao Yan, Weiguang Zhao, Jianan Ye, Xi Yang, Amir Hussain, Bin Dong, and Kaizhu Huang. Towards deeper and better multi-view feature fusion for 3d semantic segmentation. In International Conference on Neural Information Processing, 2023. 18

[46] Li Yi, Vladimir G. Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3d shape collections. ACM Transactions on Graphics, 35(6):1–12, 2016. 14

[47] Xingyilang Yin, Xi Yang, Liangchen Liu, Nannan Wang, and Xinbo Gao. Point deformable network with enhanced normal embedding for point cloud analysis. In AAAI Conference on Artificial Intelligence, 2024. 17

[48] Ziyin Zeng, Qingyong Hu, Zhong Xie, Bijun Li, Jian Zhou, and Yongyang Xu. Small but mighty: Enhancing 3d point clouds semantic segmentation with u-next framework. International Journal of Applied Earth Observation and Geoinformation, 2025. 17, 18

[49] Ziyin Zeng, Yongyang Xu, Zhong Xie, Wei Tang, Jie Wan, and Weichao Wu. Leard-net: Semantic segmentation for large-scale point cloud scene. International Journal of Applied Earth Observation and Geoinformation, 112:102953, 2022. 18

[50] Ziyin Zeng, Yongyang Xu, Zhong Xie, Wei Tang, Jie Wan, and Weichao Wu. Large-scale point cloud semantic segmentation via local perception and global descriptor vector. Expert Systems with Applications, 2024. 18

[51] Chao Zhang, Zhiguo Cao, Xin Xiong, Ke Xian, and Xinyuan Qi. Salient object detection via deep hierarchical context aggregation and multi-layer supervision. In IEEE International Conference on Image Processing, 2019. 17

[52] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H.S. Torr, and Vladlen Koltun. Point transformer. In IEEE/CVF International Conference on Computer Vision (ICCV), 2021. 18

[53] Qikui Zhu, Bo Du, Baris Turkbey, Peter L Choyke, and Pingkun Yan. Deeply-supervised cnn for prostate segmentation. In International Joint Conference on Neural Networks, 2017. 17