remote sensing

MDPI

Article

PIIE-DSA-Net for 3D Semantic Segmentation of Urban Indoor and Outdoor Datasets

Fengjiao Gao [1], Yiming Yan [2,*], Hemin Lin [2,3] and Ruiyao Shi [2]

[1] Intelligent Manufacturing Research Institute, Heilongjiang Academy of Sciences, Harbin 150001, China; gaofengjiao@haai.org.cn
[2] College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China; hemin.lin@sophgo.com (H.L.); shiruiyao@hrbeu.edu.cn (R.S.)
[3] SOPHGO, Beijing 100080, China
* Correspondence: yanyiming@hrbeu.edu.cn; Tel.: +86-13936513116

Abstract: In this paper, a 3D semantic segmentation method is proposed, in which a novel feature extraction framework is introduced assembling point initial information embedding (PIIE) and dynamic self-attention (DSA)—named PIIE-DSA-net. Ideal segmentation accuracy is a challenging task, since the sparse, irregular and disordered structure of point cloud. Currently, taking into account both low-level features and deep features of the point cloud is the more reliable and widely used feature extraction method. Since the asymmetry between the length of the low-level features and deep features, most methods cannot reliably extract and fuse the features as expected and obtain ideal segmentation results. Our PIIE-DSA-net first introduced the PIIE module to maintain the low-level initial point-cloud position and RGB information (optional), and we combined them with deep features extracted by the PAConv backbone. Secondly, we proposed a DSA module by using a learnable weight transformation tensor to transform the combined PIIE features and following a self-attention structure. In this way, we obtain optimized fused low-level and deep features, which is more efficient for segmentation. Experiments show that our PIIE-DSA-net is ranked at least in the top seventh among the most recent published state-of-art methods on the indoor dataset and also made a great improvement than original PAConv on outdoor datasets.

Keywords: 3D semantic segmentation; point cloud; feature extraction; self-attention

check for updates

Citation: Gao, F.; Yan, Y.; Lin, H.; Shi, R. PIIE-DSA-Net for 3D Semantic Segmentation of Urban Indoor and Outdoor Datasets. *Remote Sens.* **2022**, *14*, 3583. https://doi.org/10.3390/rs14153583

Academic Editors: Martin Weinmann, Florent Poux and Eleonora Grilli

Received: 16 May 2022
Accepted: 21 July 2022
Published: 26 July 2022

1. Introduction

In recent years, the development of 3D point-cloud-processing technology has been greatly promoted for its wide urban applications, such as urban 3D modeling [1], power line inspection [2], simultaneous positioning and mapping [3] and self-driving cars [4]. 3D semantic segmentation is to classify each 3D point to one specific category [5], which is one of the most important 3D point-cloud-processing tasks. Airborne laser scanner (ALS) [6], mobile laser scanning (MLS) [7], terrestrial laser scanning (TLS) [8,9] and unmanned aerial vehicle (UAV) photogrammetry [10] are the most popular methods to collect urban 3D point clouds from indoor and outdoor scenes.

Irregular and disordered structure of 3D point clouds is one of the greatest challenges for 3D feature extraction and further semantic segmentation [11–15]. Therefore, more efficient feature extraction methods are needed. At present, most of point cloud feature extraction methods and their corresponding semantic segmentation methods can be grouped into three kinds: point cloud projection-based methods [16–18], voxel-based methods [19–23] and point-based methods [24–26]. Since both the projection-based methods and the voxel-based methods may lose information during projection or voxelization, most researchers focus on point-based methods.

To solve the disorder problem, point-grouping methods and point-representation methods are widely researched for optimized local and global feature extraction. Pointnet [25] is one of the milestones on point-based methods. It has a framework that uses the spherical space with a set radius to search for the neighbor points of each specific point. Convolution-based methods are used to further obtain the local and global features of different positions. Many methods have followed this framework and that of the improved version Pointnet++ [5,26]. RSnet [27] uses a data slicing operation to cut the point cloud into different parts, extract the local features respectively and then aggregate these local features to obtain the global features.

The feature extraction method of RSnet has less computational complexity and can be trained end-to-end. In Pointweb [28], an adaptive feature adjustment module for adjusting local features is proposed. For local point clusters in spherical space, the network is used to learn the influence of each point on the other points to thus improve the local features. An efficient feature sampling structure Shellnet [29] is proposed to optimize the sampling of the point cloud. It divides spherical space with different radii and performs corresponding feature extraction and pooling operations on the features of spherical space within the radius.

In [30], an anisotropic separable set abstraction (ASSA) module was proposed to improve PointNet++. Triangular representation was proposed in RepSurf-U [31], and a high-efficiency plug-and-play module for point cloud was constructed. PointNeXt [32] is an improved training strategy that can be widely used in the point cloud domain; it is considered as the next generation version of PointNets. In PointASNL [33], a processing method based on nonlocal neural networks with adaptive sampling was proposed and obtained state-of-art (SOTA) results. Although points can be grouped and represented in different ways, no methods can be accepted as the most robust and efficient yet.

Simultaneously, many researchers attempt to find better ways of convolution and feature encoding. PointCNN [34] is a framework for dealing with point-cloud problems from a convolution perspective, and a feature integration for the point features around each representative point was proposed to replace the conventional convolution operation, and good results have been achieved on the 3D segmentation task. In KPConv [35], the authors proposed a spatially adaptive deformable convolution kernel suitable for point clouds, which learns the spatial position offset of each node of the convolution kernel while learning the parameters of the convolution kernel, and thus effective features can still be extracted when facing different spatial locations of the point cloud.

RandLA-Net [36] employed an efficient point cloud downsampling strategy and local spatial location encoding, which can achieve high segmentation accuracy and processing speed. Continuous convolutions for point-cloud processing proposed by ConvPoint [37] is also an efficient method. PAConv [38] introduced a general convolution operation position adaptive convolution and obtained SOTA performance. The key idea of PAConv is to build convolution kernels by dynamically combining basic weight matrices stored in the weight library, where the coefficients of these weight matrices are adaptively learned from point locations via ScoreNet. In this way, the kernel is built in a data-driven manner, giving PAConv greater flexibility than 2D convolution to handle irregular and disordered point-cloud data. Novel methods of convolution and feature encoding are various, and none can be recognized as the best.

Moreover, most researchers focus on new structures of backbone networks, and different kinds of attention-mechanism-based methods are also employed in different backbones [39–43], including channel attention, spatial attention, self-attention and multi-attention. Most of these methods can enhance the feature extraction and achieve higher accuracy. Self-organizing mapping was proposed in SO-net [44] and explicitly uses the spatial distribution of point cloud to extract the features of different layer structures of a single point and self-organizing mapping node. PointTransformer [45] attempted to prove that self-attention can completely replace convolutions in point-cloud-processing.

PatchFormer [46] proposed a linear attention mechanism in the point-cloud analysis paradigm: Patch ATtention (PAT), which is faster than PointTransformer. Stratified Transformer [47] builds a strong transformer tailored for 3D point cloud segmentation by enlarging the effective receptive field and building direct long-range dependency. Most X-Transformer-based methods have a good performance on accuracy but also have high computational costs.

Furthermore, some new works are dedicated to enhancing features of point cloud in different ways. First, graph-based methods were proposed for feature augmentation. In [48], an edge convolution based on Pointnet was proposed. It attempts to extract edge features using graph convolution to optimize the problem of insufficient local features of Pointnet. A local spectral convolution [49] was proposed to learn the structure information of each point. The local spectral convolution layer is realized by constructing a dynamic graph, dynamically calculating the Laplace operator and pooling hierarchy, and the features at the graph nodes are aggregated by recursive clustering spectral coordinates.

A graph-structured method based on deep metric learning [50] also obtains high-ranking performance in different datasets. In [51], a multi-resolution graph neural network was proposed, which focuses on large-scale segmentation. Graph-based feature extraction methods enhance the relationship between points; however, these methods still need to be further developed for more robust results. CGA-Net [52] has a two-path feature augmentation architecture based on category information.

BAAF-Net [53] has an adaptive feature fusion module and a bilateral block to augment the local context of the points. Indeed, most recent published works, which obtained SOTA results, benefited from different novel feature augmentation strategies. However, current SOTA methods pay insufficient attention to low-level features, most of the SOTA backbones extract a long feature vector, and the low-level feature vector is always short. Since the asymmetry between the length of the low-level features and deep features, it is difficult to properly fuse the features as expected and obtain ideal segmentation results.

Although many SOTA methods have made great progress using different strategies [54–57], the accuracy of 3D semantic segmentation is still low on most of the new datasets. Our work is also based on the idea of feature augmentation. The contributions of this paper are as follows: PIIE-DSA-net is proposed for 3D semantic segmentation based on PAConv. (1) Point initial information embedding (PIIE) module was employed to keep the low-level initial point-cloud position and RGB information (optional) and combine them with deep features extracted by PAConv encoder. (2) A dynamic self-attention (DSA) module was proposed by using a learnable weights transformation tensor to transform the combined features and following a self-attention structure to generate more effective fused features for 3D segmentation.

The following of the paper is organized as follows: In Section 2, the detailed methodology of PIIE-DSA-net is introduced. In Section 3, we test the performance of PIIE-DSA-net on both an indoor and an outdoor datasets. Additionally, the ablation experiments and module analysis are given. We discuss the results and summarize the work in the final section.

2. Methodology

2.1. Framework of PIIE-DSA-Net

The framework of our PIIE-DSA-net is shown in Figure 1. Our PIIE-DSA-net can be divided into four main modules: (1) Pre-processing. (2) Point initial information embedding (PIIE). (3) Dynamic self-attention (DSA). (4) Segmentation decoder. Both training and testing data must follow all the calculation processes of the four modules. First, pre-processing of the point-cloud data is needed before they are input into the framework. The same pre-processing method in PAConv [38] is used for the point grouping, color mapping and normalization of coordinates.

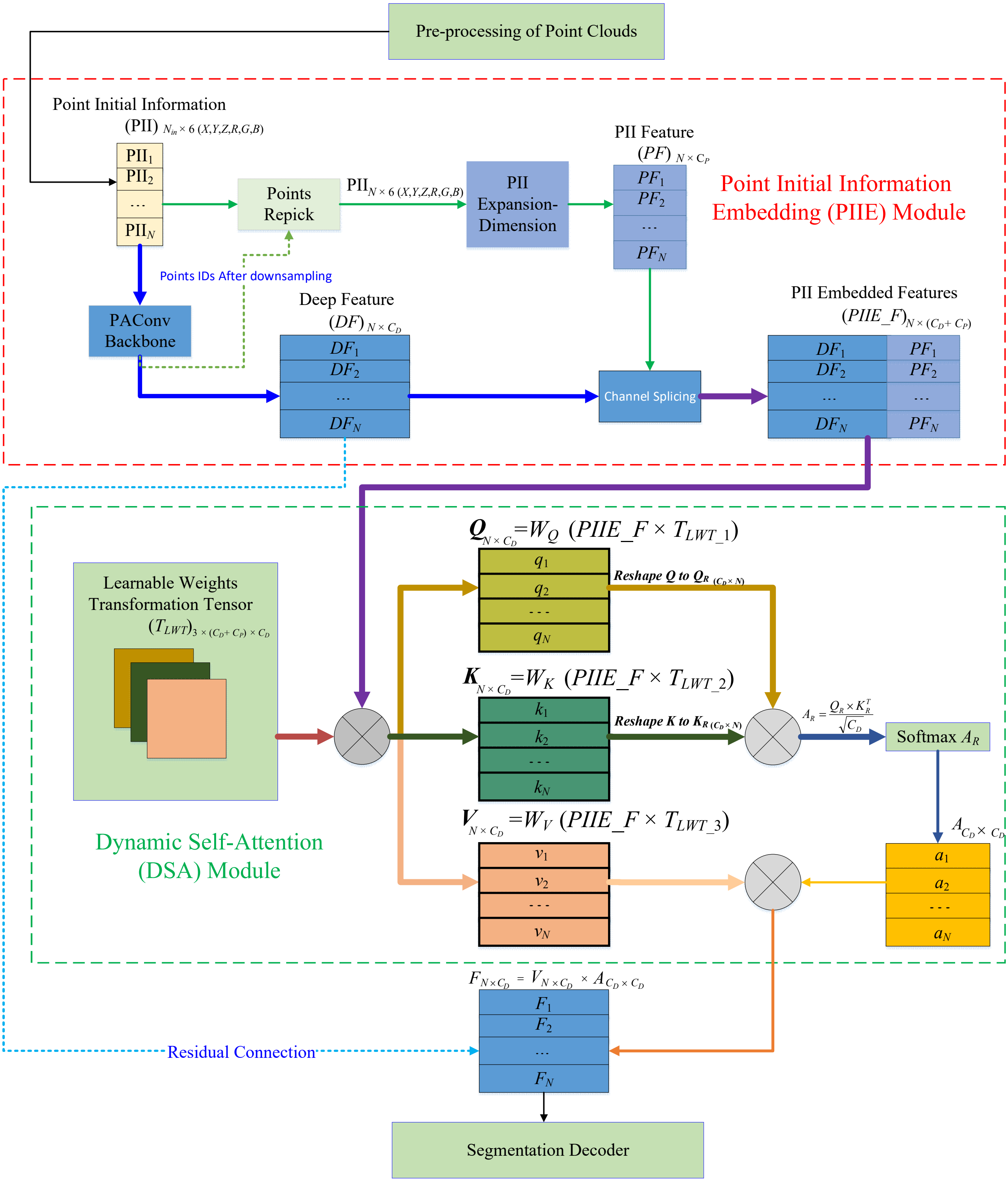
Pre-processing of Point Clouds
Point Initial Information
(PII) $N_{in} \times 6\ (X,Y,Z,R,G,B)$
PII$_1$
PII$_2$
…
PII$_N$
Points Repick
PII$_{N \times 6\ (X,Y,Z,R,G,B)}$
PII Expansion-Dimension
PII Feature
$(PF)_{N \times C_P}$
PF_1
PF_2
…
PF_N
Point Initial Information Embedding (PIIE) Module
Points IDs After downsampling
PAConv Backbone
Deep Feature
$(DF)_{N \times C_D}$
DF_1
DF_2
…
DF_N
Channel Splicing
PII Embedded Features
$(PIIE\_F)_{N \times (C_D + C_P)}$
$Q_{N \times C_D} = W_Q\ (PIIE\_F \times T_{LWT\_1})$
q_1
q_2
q_N
Reshape Q to Q_R $(C_D \times N)$
Learnable Weights Transformation Tensor
$(T_{LWT})_{3 \times (C_D + C_P) \times C_D}$
$K_{N \times C_D} = W_K\ (PIIE\_F \times T_{LWT\_2})$
k_1
k_2
k_N
Reshape K to K_R $(C_D \times N)$
$A_R = \frac{Q_R \times K_R^T}{\sqrt{C_D}}$
Softmax A_R
$V_{N \times C_D} = W_V\ (PIIE\_F \times T_{LWT\_3})$
v_1
v_2
v_N
$A_{C_D \times C_D}$
a_1
a_2
a_N
Dynamic Self-Attention (DSA) Module
$F_{N \times C_D} = V_{N \times C_D} \times A_{C_D \times C_D}$
F_1
F_2
…
F_N
Residual Connection
Segmentation Decoder

Figure 1. The framework of PIIE-DSA-net.

Secondly, PIIE is introduced to efficiently extract and assemble features extracted in different ways. Thirdly, DSA is proposed for optimized organizing the extracted features by PIIE to generate the optimized features. A residual connection is used between the features from the backbone in PIIE and the features from DSA for more reliable training.

Finally, the semantic segmentation decoder [38] decodes the features, up-samples layer by layer and outputs the predicted category point by point. The PIIE and DSA modules are introduced as follows.

2.2. Point Initial Information Embedding

Point initial information (PII) includes the position (X, Y, Z) and the color information (R, G, B) of each point. The PII of the input N_i points form a $N_i \times 6$ matrix. To efficiently extract and assemble features at different levels, PIIE processes the point-cloud data by two branches: In branch one, Pointnet++ [5,26] based SOTA backbones can be employed to extract deep features from the N_i input points, and the PAConv encoder [38] is used in our framework. As down-sampling happens during PAConv encoder, we create a point ID index to memorize the IDs of all N kept points and $(N_i - N)$ dropped points.

Deep features of the N kept points extracted by PAConv are formed as Equation (1), where C_D is the length of deep feature of each point, and C_D is 64 from the PAConv encoder output.

$$DF = \{DF_1, DF_2, \ldots, DF_N\}_{N \times C_D} \tag{1}$$

Then, the point initial information (PII) of the N kept points are re-picked and input into the other branch. An expansion-dimension net (EDN) is used to expand the PII (six dimensions) to a higher dimension and generate PII features as Equation (2), where C_P is the length of each PII feature of each point. EDN is introduced mainly to balance the dimension difference between PII feature and deep feature; thus, C_P = 64 is used in this paper.

$$PF = \{PF_1, PF_2, \ldots, PF_N\}_{N \times C_P} \tag{2}$$

The structure of the used EDN is shown in Figure 2. PII is encoded by four different cascaded 'Conv(1 × 1) + Batch Normalization (BN)'.

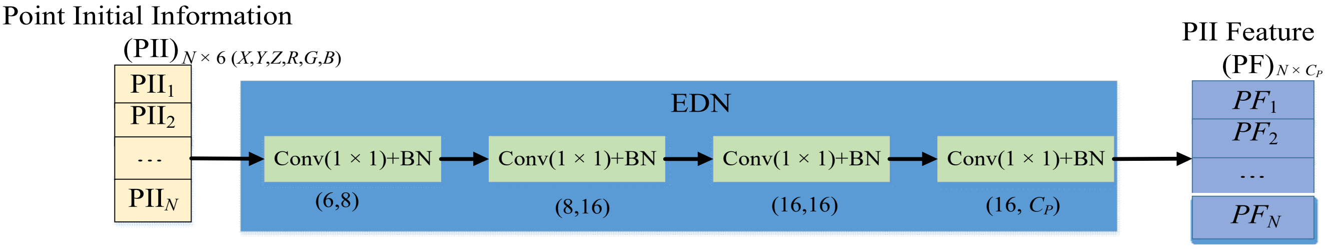

Figure 2. The structure of the expansion-dimension net (EDN) used in the PIIE module.

DF and *PF* are combined by a concatenated operation in the channel splicing module and generate an $N \times (C_D + C_P)$ feature *PIIE_F*, which contains the initial position, color information and deep features of the initial points.

2.3. Dynamic Self-Attention

The self-attention mechanism [58] is derived from text feature extraction. It is used to adaptively learn the correlation between different features and the importance of different features. By back-propagation learning, the weights to different features were assigned, which achieves optimized feature extraction. Similar to shown in Figure 1, the structure of conventional self-attention mechanism directly uses the original data or extracted features as input, and the three matrices Queries (Q), Keys (K) and Values (V) were the products of the input and each learnable weights W_Q, W_K and W_V.

The product of matrix Q and transpose of matrix K characterizes the cross-correlation between features. After that, an attention matrix is obtained through the soft-max layer. The attention matrix describes the weight distribution of different degrees of importance to the input. The attention matrix is multiplied with the matrix V and finally obtains the feature matrix. In this way, the weights of the features that input into Q, K and V are the same and fixed.

The dynamic self-attention module is designed for optimized organizing the extracted *PIIE_F*. Different from the structure of a conventional self-attention mechanism, a transformation tensor T_{LWT} is proposed to make the weights learnable, which multiplies with the input *PIIE_F* before inputting into Q, K and V. T_{LWT} is a 3 × (C_D + C_P) × C_D tensor formed as $\{T_{LWT\_1}, T_{LWT\_2}, T_{LWT\_3}\}$, and T_{LWT_1}, T_{LWT_2} and T_{LWT_3} are (C_D + C_P) × C_D matrices. T_{LWT} is initialized by the method of 'He initialization' [59] before forward-propagation and is updated during back-propagation. The input self-attention Q, K and V matrices (N × C_D) are generated by the products T_{LWT_1}, T_{LWT_2} and T_{LWT_3} with *PIIE_F*, respectively, as shown in Equation (3):

$$\begin{aligned} Q &= W_Q(PIIE\_F \times T_{LWT\_1}) \\ K &= W_K(PIIE\_F \times T_{LWT\_2}) \\ V &= W_V(PIIE\_F \times T_{LWT\_3}) \end{aligned} \tag{3}$$

The Q, K matrices are multiplied with W_Q and W_K and then reshaped to Q_R (C_D × N) and K_R (C_D × N), respectively and then generate A_R (C_D × C_D) as in Equation (4). The attention matrix A (C_D × C_D) is the normalized A_R by a softmax layer.

$$A_R = \frac{Q_R \times K_R^T}{\sqrt{C_D}} \tag{4}$$

The final fused feature F (N × C_D) used for semantic segmentation is calculated by V (N × C_D) and A (C_D × C_D) by Equation (5):

$$F = V \times A \tag{5}$$

In addition to the conventional single-head self-attention structure [58] used in our PIIE-DSA-net, there is also a multi-head self-attention mechanism [60]. We tested different self-attention mechanisms in the experiments.

2.4. Loss Function Used in PIIE-DSA-Net

The loss functions used in PIIE-DSA-net include cross-entropy loss L_{ce} and matrix similarity loss L_{ms}. As in Equation (6), λ_{ce} and λ_{ms} are weight coefficients.

$$L = \lambda_{ce} L_{ce} + \lambda_{ms} L_{ms} \tag{6}$$

The cross-entropy loss L_{ce} constrains the correctness of probability prediction in multi-classification. As in Equation (7), N_t is the number of samples, i represents the i-th sample, c is the c-th class, y_{ic} is 1 when the correct predict i-th sample is the c-th class, otherwise y_{ic} = 0. p_{ic} is the probability of predicting i-th sample as the c-th class. When the more correct predicted samples are, the higher the probability of correct predicted samples is, the smaller the cross-entropy is and vice versa.

$$L_{ce} = -\frac{1}{N_t} \sum_i \sum_{c=1}^{classes} y_{ic} \log(p_{ic}) \tag{7}$$

The matrix similarity loss L_{ms} is defined according to the weight regularization used in PAConv [38]. However, we use it in the learnable weights transformation tensor T_{LWT}. It maintains the independence of the feature extraction methods learned by multiple weight matrices after initialization and constrains the correlation between different weights to minimize the redundancy and duplication of extracted features. As shown in Equation (8), B represents the sets of the defined weight matrices, B_i and B_j respectively represent two different weight matrices. When the similarity between the weight matrices is smaller, the loss is smaller and vice versa.

$$L_{ms} = \sum_{B_i, B_j \in B, i \neq j} \frac{\left|\sum B_i B_j\right|}{\|B_i\|_2 \|B_j\|_2} \tag{8}$$

3. Experiments

In this section, experiments on 3D semantic segmentation are performed on three datasets respectively to verify the effectiveness of PIIE-DSA-net. In Section 3.1, the detailed descriptions of datasets are introduced. In Section 3.2, evaluation metrics used in the experiments are given. In Section 3.3, 3D semantic segmentation performances of different methods are compared in two datasets, and ablation experiments are further analyzed.

3.1. Description of the Datasets

Experiments of 3D semantic segmentation are performed on the indoor dataset S3DIS [61], outdoor datasets SensatUrban [62] and Hessigheim 3D [63].

Statement of the Datasets

1. Stanford Large-Scale 3D Indoor Spaces Dataset (S3DIS)

Stanford Large-Scale 3D Indoor Spaces Dataset (S3DIS) was proposed by Stanford University. It is an indoor scene benchmark dataset in the task of 3D semantic segmentation. The point cloud was collected by a Matterport camera. As shown in Figure 3, the high-resolution aerial imagery sequences were captured by fixed-wing drone Ebee X, and the dense image matching method was used [61]. It consists of 271 rooms scanned from 11 kinds of buildings including office, conference room, hallway, auditorium, open space, lobby, lounge, pantry, copy room, storage and WC.

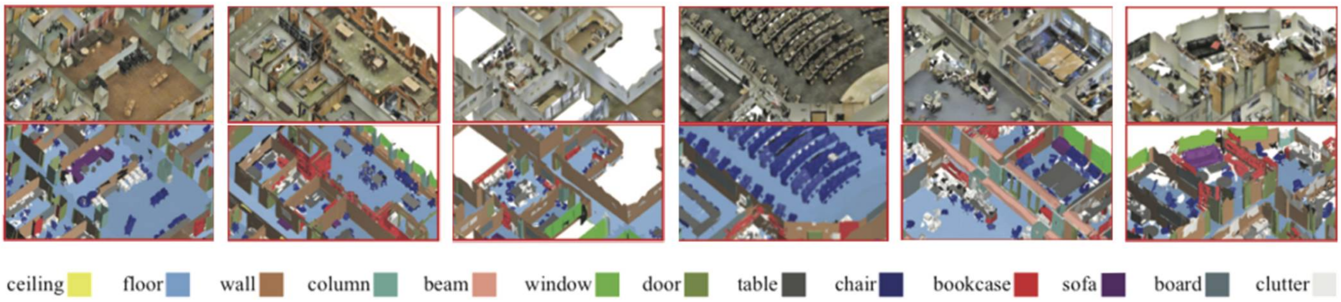

Figure 3. Stanford Large-Scale 3D Indoor Spaces Dataset (S3DIS) [61].

The official dataset is divided into six areas with a total of 273 million points. The dataset is labeled into 13 categories as ‘ceiling’, ‘floor’, ‘wall’, ‘beam’, ‘column’, ‘window’, ‘door’, ‘table’, ‘chair’, ‘sofa’, ‘bookcase’, ‘board’ and ‘clutter’. The data is pre-processed according to the processing steps of PAConv [38], and the points of each room are divided into several 1 m × 1 m blocks according to the horizontal direction. A total of 4096 points are randomly picked from each block. Each point contains six-dimensional initial information, including normalized XYZ coordinates and RGB colors.

In the experiment, all six areas were used by its official grouped training, verifying and testing sets. Specially, Area5 is an officially designated area that could be used for separate testing. Random scaling, selection and random jittering are used for data augmentation during training.

2. SensatUrban dataset

The airborne SensatUrban dataset was released in CVPR2021 with nearly 3 billion labeled points. The point cloud was obtained by UAV photogrammetry technology, which covers large areas of two British cities: Birmingham and Cambridge, covering about 6 square kilometers of the urban area, including the 1.2 square kilometers of Birmingham and the 3.2 square kilometers of Cambridge. As is shown in Figure 4, the point clouds are labeled into 13 categories, including ‘ground’, ‘vegetation’, ‘building’, ‘wall’, ‘bridge’, ‘parking’, ‘rail’, ‘car’, ‘footpath’, ‘bike’, ‘water’, ‘traffic road’ and ‘street furniture’. The dataset uses the registered optical image mapping RGB information for 3D point clouds. In the experiments, the point-cloud data is divided into multiple 30 m × 30 m blocks according to the horizontal direction. A total of 4096 points are also randomly picked from each part. Each point also contains six-dimensional initial information, which are normalized XYZ coordinates and RGB colors. The data augmentation steps during training are the same as S3DIS.

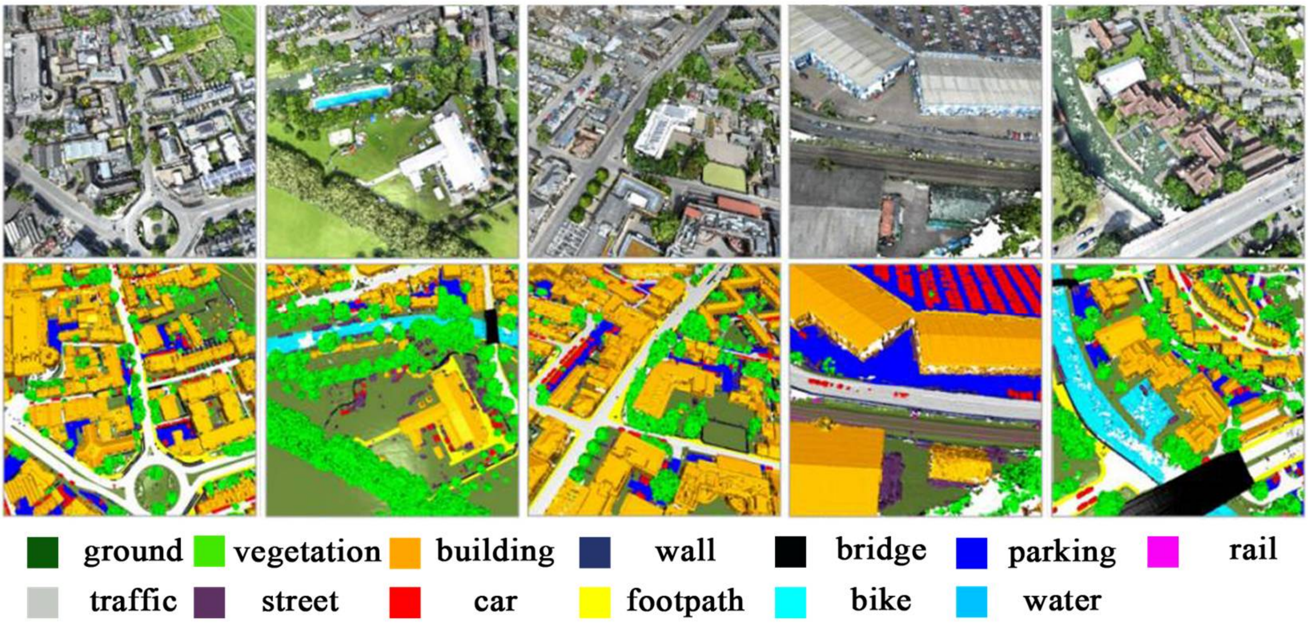

Figure 4. SensatUrban Dataset.

Since the competition of SensatUrban dataset is over, the labels of the competition testing set are not published, and thus the testing set cannot be used for evaluation anymore currently. Thus, in our experiments, we divided the competition training set of SensatUrban and redefined the training and testing sets. The competition training set of SensatUrban contains 37 blocks from the two cities, which are all published with labels. To be fair, only part of them are selected to make the ratio of training set and testing set similar to the competition dataset. The blocks (number 3, 6, 7, 9 and 10) in Birmingham and the blocks (number 3, 4, 6, 8, 14, 18, 19, 20, 21, 23, 25, 28 and 33) in Cambridge are selected as our new training set. The blocks (number 1 and 5) in Birmingham and the blocks (number 7, 10, 12 and 17) in Cambridge are selected as our new testing set.

3. Hessigheim 3D dataset

The Hessigheim 3D dataset (H3D) was proposed by University of Stuttgart and is a benchmark in the task of 3D semantic segmentation [63]. The H3D dataset, shown in Figure 5, consists of High density LiDAR data of 800 points/m^2 enriched by RGB colors of on board cameras incorporating a ground sample distance (GSD) of 2–3 cm. Multi-temporal datasets are available for four different epochs. The dataset is labeled into 11 categories: ‘Low Vegetation’, ‘Impervious Surface’, ‘Vehicle’, ‘Urban Furniture’, ‘Roof’, ‘Facade’, ‘Shrub’, ‘Tree’, ‘Soil/Gravel’, ‘Vertical Surface’ and ‘Chimney’. A total of 4096 points are randomly picked from each part. Each point contains six-dimensional initial information, including normalized XYZ coordinates and RGB colors.

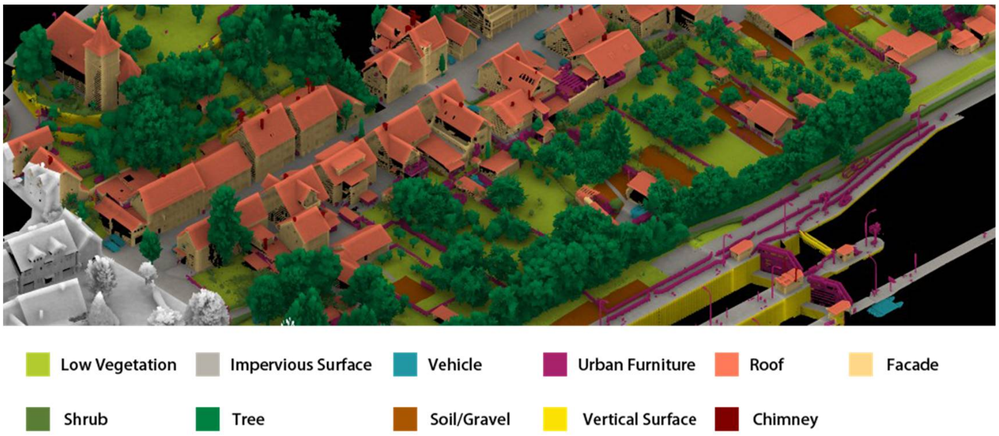

Figure 5. Hessigheim 3D Dataset [63].

3.2. Evaluation Metrics

The mean intersection over union (*mIoU*), mean accuracy (*mAcc*) and overall accuracy (*OA*) are the most widely used for evaluation 3D semantic segmentation of point cloud.

IoU is defined in Equation (9), where P_{pre} and P_{GT} represent the predicted category and ground truth category, respectively. *mIoU* is defined in Equation (10), where IoU_i represents the *IoU* of the *i*-th category, and *C* represents the number of categories.

$$IoU = \frac{P_{pre} \cap P_{GT}}{P_{pre} \cup P_{GT}} \tag{9}$$

$$mIoU = \frac{\sum_{i}^{C} IoU_i}{C} \tag{10}$$

As in Equation (11), *mAcc* is calculated by the proportion of correct predictions on each category and averages it according to the number of categories. *C* represents the number of categories, and N_i is the number of points in the *i*-th category. P_{pre_j} and P_{GT_j} are the predicted category and ground truth category of the *j*-th point in the *i*-th category.

$$mAcc = \frac{\sum_{i}^{C} \frac{1}{N_i} \sum_{j}^{N_i} P_{pre\_j} = P_{GT\_j}}{C} \tag{11}$$

$$OA = \frac{TP + TN}{TP + FP + TN + FN} \tag{12}$$

The True Positive (*TP*), False Positive (*FP*), True Negative (*TN*) and False Negative (*FN*) are used to define overall accuracy (*OA*) as in Equation (12).

3.3. 3D Semantic Segmentation Experiments

The 3D semantic segmentation experiments were performed on a Silver with 4210R CPU, NVIDIA RTX 3090 GPU and 40GB RAM. To train our PIIE-DSA-net, the batch size of training samples is set to 16, and the initial learning rate is set to 0.05, the number of iterations is set to 80, the weight decay is set to 0.00005, 16 weight matrices are set in PAConv encoder, and the SGD optimizer is used.

3.3.1. Performance of PIIE-DSA-Net on Indoor Dataset S3DIS

First, we tested the performance of PIIE-DSA-net on all six areas of S3DIS. As shown in Table 1, the *IoUs* of each category obtained by PIIE-DSA-net are listed. Overall, PIIE-DSA-net can give high evaluation values in Area1, Area3 and Area6, and the results were relative lower in the other three areas. In particular, in different areas, the higher OA shows most of the points were correctly segmented, and the relative lower *mIoU* can further indicate segmentation results of the categories with fewer points are not ideal. It can also be reflected by the *mAcc*. In Table 2, the most recent published works and their six-fold experiments performance on S3DIS are listed with form of overall ranking. Our PIIE-DSA-net is seventh of the current top ten methods. The ranking is from the website (https://paperswithcode.com/sota/semantic-segmentation-on-s3dis, accessed on 10 July 2022).

Table 1. The six-fold experiments on the S3DIS dataset (%).

Categories/Test Area	Area1	Area2	Area3	Area4	Area5	Area6
ceiling	97.98	90.67	95.94	93.26	93.27	96.31
floor	97.24	77.66	98.30	97.38	98.51	97.33
wall	93.61	79.62	83.00	78.11	82.75	85.31
column	86.87	31.80	23.26	34.58	28.42	62.85
beam	93.05	15.25	62.10	0.87	0.00	81.75
window	94.30	54.55	82.96	33.23	62.26	85.63
door	94.51	65.72	91.93	64.12	67.63	89.64
table	86.91	60.48	77.70	63.04	79.01	78.01
chair	93.30	28.54	83.90	72.64	88.86	80.95
bookcase	94.59	28.00	75.13	63.02	60.65	53.52
sofa	90.26	47.55	75.53	54.61	74.51	75.46
board	93.18	19.75	90.21	45.20	74.98	81.26
clutter	88.27	37.13	75.53	58.72	58.86	71.66
mIoU	92.62	48.98	78.11	58.37	66.90	79.98
mAcc	96.23	62.41	86.46	68.15	73.90	88.35
OA	96.77	79.46	91.59	85.86	89.44	92.24

Table 2. Current overall TOP-10 ranking of the six-fold experiments on the S3DIS dataset (%).

Rank	Methods	*mIoU*	*mAcc*	*OA*
1	RepSurf-U [31]	74.3	82.6	90.8
2	PointNeXt [32]	74.9	83.0	90.3
3	PointTransformer [45]	73.5	81.9	90.2
4	DeepViewAgg [54]	74.7	83.8	90.1
5	CBL [55]	73.1	79.4	89.6
6	BAAF-Net [53]	72.2	83.1	88.9
7	PIIE-DSA-net (OURS)	71.66	81.24	88.89
8	PointASNL [33]	68.7	79.0	88.8
9	ConvPoint [37]	68.2	N/A	88.8
10	JSNet [56]	61.7	71.7	88.7

The detailed performance of PIIE-DSA-net can be further found in the individually testing on Area5. Since our PIIE-DSA-net is an improved method based on PAConv, we mainly show the visualized results of PAConv for comparison. Figure 6(a1–a3) are the original input PII data of three different scenes; Figure 6(b1–b3) are the ground truths of the input data; Figure 6(c1–c3) are the prediction results by PAConv; and Figure 6(d1–d3) are the predicted result by PIIE-DSA-net.

PIIE-DSA-net has improved performance on S3DIS in the detailed prediction of categories, such as ‘ceiling’, ‘door’, ‘sofa’, ‘table’ and ‘chair’. Specifically, as the circled areas in Figure 6(c1), a part of the ‘ceiling’ area is incorrectly predicted to ‘beam’ by PAConv, while this area can be correctly predicted by PIIE-DSA-net as shown in Figure 6(d1). A group of points are incorrectly predicted in ‘sofa’ areas by PAConv; however, these points can be correctly predicted by PIIE-DSA-net. In Figure 6(c2,d2), PIIE-DSA-net improves the prediction in the circled ‘wall’ part, which is incorrectly predicted by PAConv. In Figure 6(c3,d3), PAConv performs poorly in the circled ‘tables’ and ‘chairs’ areas; however, PIIE-DSA-net improves the results. PIIE-DSA-net obtained better segmentation completeness on the areas of the same category with more points.

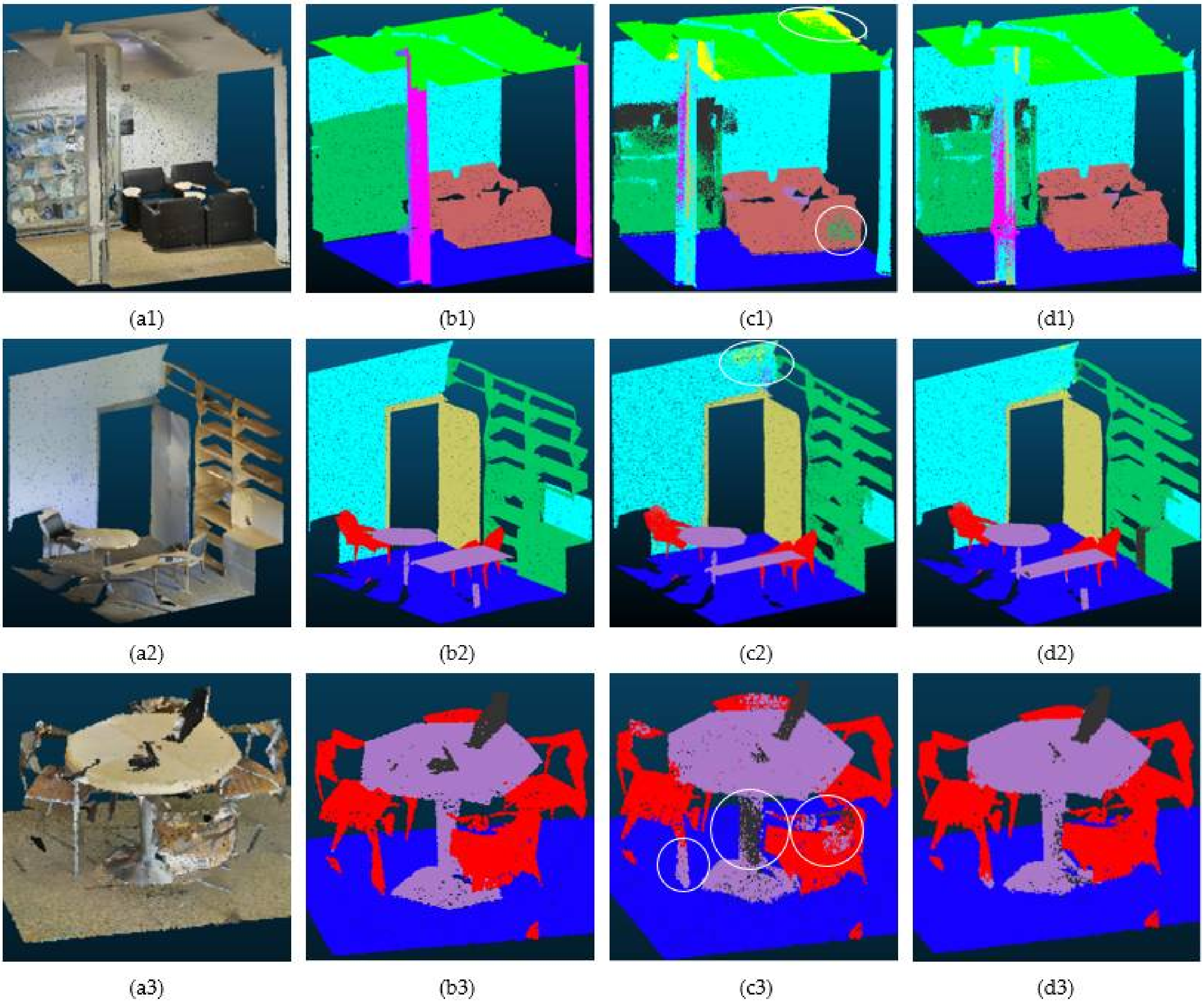

Figure 6. 3D semantic segmentation of S3DIS. (**a1**) Data of Scene 1; (**b1**) Ground Truth; (**c1**) PAConv; (**d1**) PIIE-DSA-net; (**a2**) Data of Scene 2; (**b2**) Ground Truth; (**c2**) PAConv; (**d2**) PIIE-DSA-net; (**a3**) Data of Scene 3; (**b3**) Ground Truth; (**c3**) PAConv; (**d3**) PIIE-DSA-net.

We re-performed the Area5 experiments of the following typical methods: PointNet, PointNet++, PointCNN, KPconv rigid, RandLA-Net and PAConv and listed the *IoU* of each class in Table 3. PIIE-DSA-net obtained the best *IoU* on six categories of all the 13, which were 'wall', 'beam', 'chair', 'sofa', 'board' and 'clutter'. The best *mIoU* and *mAcc* were obtained by PIIE-DSA-net compared with other six methods. Furthermore, we collected the most recent published works and their Area5 experiments performance on S3DIS from the website and listed them in Table 4. Our PIIE-DSA-net is sixth of the current top ten methods. The ranking is from the website: (https://paperswithcode.com/sota/semantic-segmentation-on-s3dis-area5, accessed on 10 July 2022).

Table 3. The Area5 experiments on the S3DIS dataset (%).

Categories/Methods	PointNet	PointNet++	PointCNN	KPconv Rigid	RandLA-Net	PAConv	PIIE-DSA-Net
ceiling	88.80	91.31	92.31	92.6	91.69	94.55	93.72
floor	97.33	96.92	98.24	97.3	96.90	98.59	98.51
wall	69.80	78.73	79.41	81.4	78.45	82.37	82.75
column	3.92	15.99	17.60	16.5	27.07	26.43	28.42
beam	0.05	0.00	0.00	0.00	0.00	0.00	0.00
window	46.26	54.93	22.77	54.5	64.19	57.96	62.26
door	10.76	31.88	62.09	69.5	37.53	59.96	67.63
table	52.61	83.52	80.59	90.1	73.97	89.73	79.01
chair	58.93	74.62	74.39	80.2	83.94	80.44	88.86
bookcase	40.28	67.24	66.67	74.6	66.39	74.32	60.65
sofa	5.85	49.31	31.67	66.4	67.94	69.80	74.51
board	26.38	54.15	62.05	63.7	61.96	73.50	74.98
clutter	33.22	45.89	56.74	58.1	50.37	57.72	58.86
mIoU	41.09	57.27	57.26	65.4	61.57	66.58	66.90
mAcc	49.98	63.54	63.86	70.9	71.50	73.00	73.90

Table 4. Current overall TOP-10 ranking of Area5 experiments on the S3DIS dataset (%).

Rank	Methods	*mIoU*	*mAcc*	*OA*
1	StratifiedFormer [61]	72.0	78.1	91.5
2	PointNeXt [32]	71.1	77.2	91.0
3	PointTransformer [45]	70.4	76.5	90.8
4	CBL [55]	69.4	75.2	90.6
5	RepSurf-U [31]	68.9	76.0	90.2
6	PIIE-DSA-net (OURS)	66.9	73.9	89.44
7	BAAF-Net [53]	65.4	73.1	88.9
8	MuG-net [51]	63.5	N/A	88.1
9	SSP + SPG [50]	61.7	68.2	87.9
10	HPEIN [57]	61.85	68.3	87.18

- Ablation experiments on the indoor dataset

Area5 of S3DIS was taken to perform the ablation experiments, and the results are given in Table 5. First, we compared original PAConv, PAConv with only PIIE module (PAConv + PIIE) and PIIE-DSA-net. Except for the case that PAConv + PIIE won *mAcc*, PIIE-DSA-net obtained the highest *mIoU*. Secondly, before the self-attention part, there are also different methods to transform PIIE for effective feature extraction. Convolution transformation, full-connection transformation and matrix transformation are compared.

Table 5. Ablation experiments on the indoor dataset (%).

S3DIS	Module Ablation		PIIE Transformation Methods		Selection of Multi-Head Attention Operation		PIIE-DSA-Net
	PAConv	PAConv + PIIE	Convolution Transform	Full-Connection Transform	Two-Head Attention	Four-Head Attention	
ceiling	94.55	93.67	93.63	94.01	92.58	94.90	93.72
floor	98.59	98.50	98.21	98.05	98.51	98.44	98.51
wall	82.37	82.63	82.27	82.36	82.30	82.14	82.75
column	26.43	32.62	20.04	21.49	26.41	17.34	28.42
beam	0.00	0.00	0.00	0.00	0.00	0.00	0.00
window	57.96	59.55	59.50	60.93	59.50	57.52	62.26
door	59.96	65.92	67.95	68.72	62.80	54.77	67.63
table	80.44	79.64	79.13	77.29	79.31	80.10	79.01
chair	89.73	88.34	86.13	85.68	88.21	88.54	88.86
bookcase	74.32	60.92	64.12	59.67	58.23	61.02	60.65
sofa	69.80	74.40	72.28	71.20	73.89	75.49	74.51
board	73.50	71.67	72.36	69.55	75.71	73.67	74.98
clutter	57.72	59.12	58.36	58.48	56.96	58.82	58.86
mIoU	66.58	66.69	65.69	65.19	65.72	64.83	66.90
mAcc	73.00	73.98	71.96	71.65	72.13	70.87	73.90

The method of matrix transformation, used in the DSA module, obtained higher *mIoU* and *mAcc* compared with the others. Thirdly, in the DSA module, we also compared the performance when using different multi-head attention operations. Single-head attention, two-head attention and four-head attention were tried in the DSA module. Single-head attention used in our PIIE-DSA-net was more effective than the other ways. The PIIE module made a greater impact in PIIE-DSA-net on the indoor dataset.

3.3.2. Performance of PIIE-DSA-Net on the Outdoor Dataset SensatUrban and H3D

1. Result analysis of SensatUrban dataset

Some examples of the segmentation results on SensatUrban are shown in Figure 7, where Figure 7(a1–a4) are the original input PII; Figure 7(b1–b4) are the ground truth; Figure 7(c1–c4) are the prediction results by PAConv; Figure 7(d1–d4) are the prediction results by PIIE-DSA-net. Since the point clouds of SensatUrban are collected from airborne sensors, 'ground', 'parking', 'footpath' and 'traffic road', which have similar heights, are always confusing categories. As in Figure 7(c1), a large area of 'traffic roads' and 'ground' are incorrectly divided into 'parking' by PAConv. In Figure 7(c2), some 'ground' and 'footpath' are misclassified into each other, and some 'ground' and 'traffic road' are wrongly divided into 'parking'. Similar bad predictions also appeared in Figure 7(c3,c4). In contrast, when using PIIE-DSA-net, the above problems are improved, as shown in Figure 7(d1–d4). Similarly, PIIE-DSA-net obtained better segmentation completeness in the areas of the same category with more points.

2. Ablation experiments on the outdoor dataset

The SensatUrban dataset was taken to perform ablation experiments of the outdoor dataset, and the results are given in Table 6. First, in the comparisons between original PAConv, PAConv with only PIIE module (PAConv + PIIE) and PIIE-DSA-net, PIIE-DSA-net obtained the highest *mIoU* and *mAcc*. Secondly, the method of matrix transformation also won higher *mIoU* and *mAcc* than the convolution transformation and full-connection transformation. Thirdly, single-head attention used in PIIE-DSA-net was more effective than the two-head attention and four-head attention.

Table 6. The experiments on the SensatUrban dataset (%).

SensatUrban	Module Ablation		PIIE Transformation Methods		Selection of Multi-Head Attention Operation		PIIE-DSA-Net
	PAConv	PAConv + PIIE	Convolution Transform	Full-Connection Transform	Two-Head Attention	Four-Head Attention	
ground	72.11	73.53	73.10	74.48	74.43	73.37	73.92
vegetation	97.54	97.30	97.11	97.72	97.86	97.56	97.69
building	93.01	92.90	91.98	93.65	92.92	93.22	93.05
wall	44.43	49.98	49.68	47.22	50.71	49.69	49.81
bridge	5.78	3.77	2.42	0.01	2.01	6.73	17.43
parking	39.94	40.21	37.51	43.27	43.11	41.52	40.60
rail	0.00	0.00	0.00	0.00	0.00	0.00	0.00
car	73.04	77.87	77.09	75.90	77.66	77.56	77.75
footpath	21.78	23.75	21.68	24.31	24.66	22.58	24.57
bike	0.00	0.00	0.00	0.00	0.00	0.00	0.00
water	57.97	63.87	60.22	58.62	62.87	62.47	57.10
traffic road	58.78	63.00	59.97	62.96	62.99	62.07	61.45
Street furniture	29.42	33.78	33.64	32.54	32.52	32.68	31.38
mIoU	45.68	47.69	46.49	46.98	47.82	47.65	48.09
mAcc	53.76	55.03	53.73	54.79	54.83	54.98	55.16

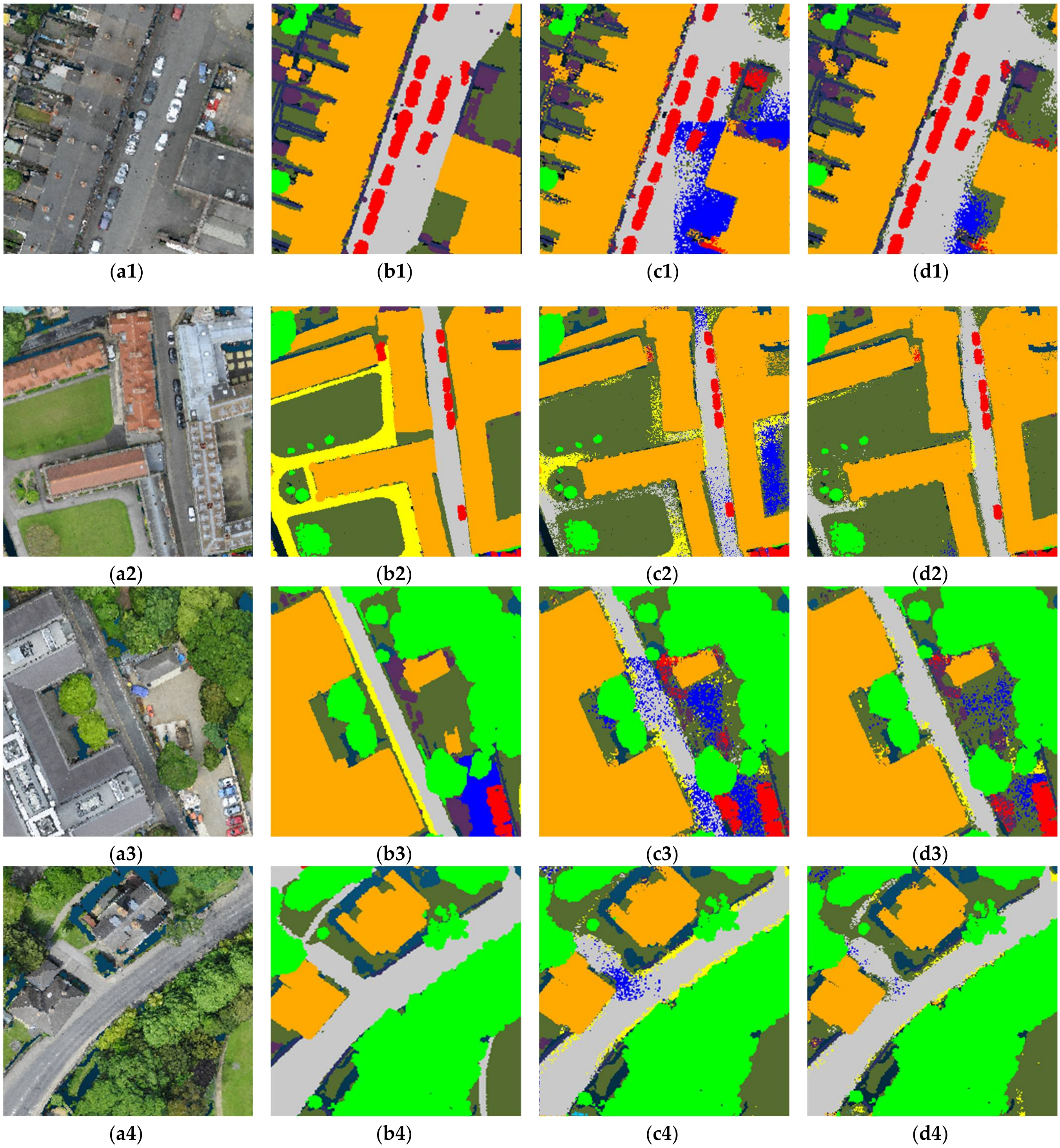
(**a1**) (**b1**) (**c1**) (**d1**)
(**a2**) (**b2**) (**c2**) (**d2**)
(**a3**) (**b3**) (**c3**) (**d3**)
(**a4**) (**b4**) (**c4**) (**d4**)

Figure 7. 3D semantic segmentation on the SensatUrban dataset. (**a1**) Data of Scene 1. (**b1**) Ground Truth. (**c1**) PAConv. (**d1**) PIIE-DSA-net. (**a2**) Data of Scene 2. (**b2**) Ground Truth. (**c2**) PAConv. (**d2**) PIIE-DSA-net. (**a3**) Data of Scene 3. (**b3**) Ground Truth. (**c3**) PAConv. (**d3**) PIIE-DSA-net. (**a4**) Data of Scene 4. (**b4**) Ground Truth. (**c4**) PAConv. (**d4**) PIIE-DSA-net.

3. Result analysis for the H3D dataset

To further verify the performance of PIIE-DSA-net, the H3D dataset was tested. Since the ground truth of the testing set of H3D is not published, the verifying set was used for testing. As is shown in Figure 8(a1–a4), some typical scenes are visualized, and their ground truths are shown in Figure 8(b1–b4). Similar to the experiment results on S3DIS and SensatUrban, PAConv obtained poor performance in the circled areas in the four different scenes as shown in Figure 8(c1–c4), especially on the edge areas between different categories and on categories with few numbers of points. In contrast, PIIE-DSA-net greatly improved the results in these areas as shown in Figure 8(d1–d4).

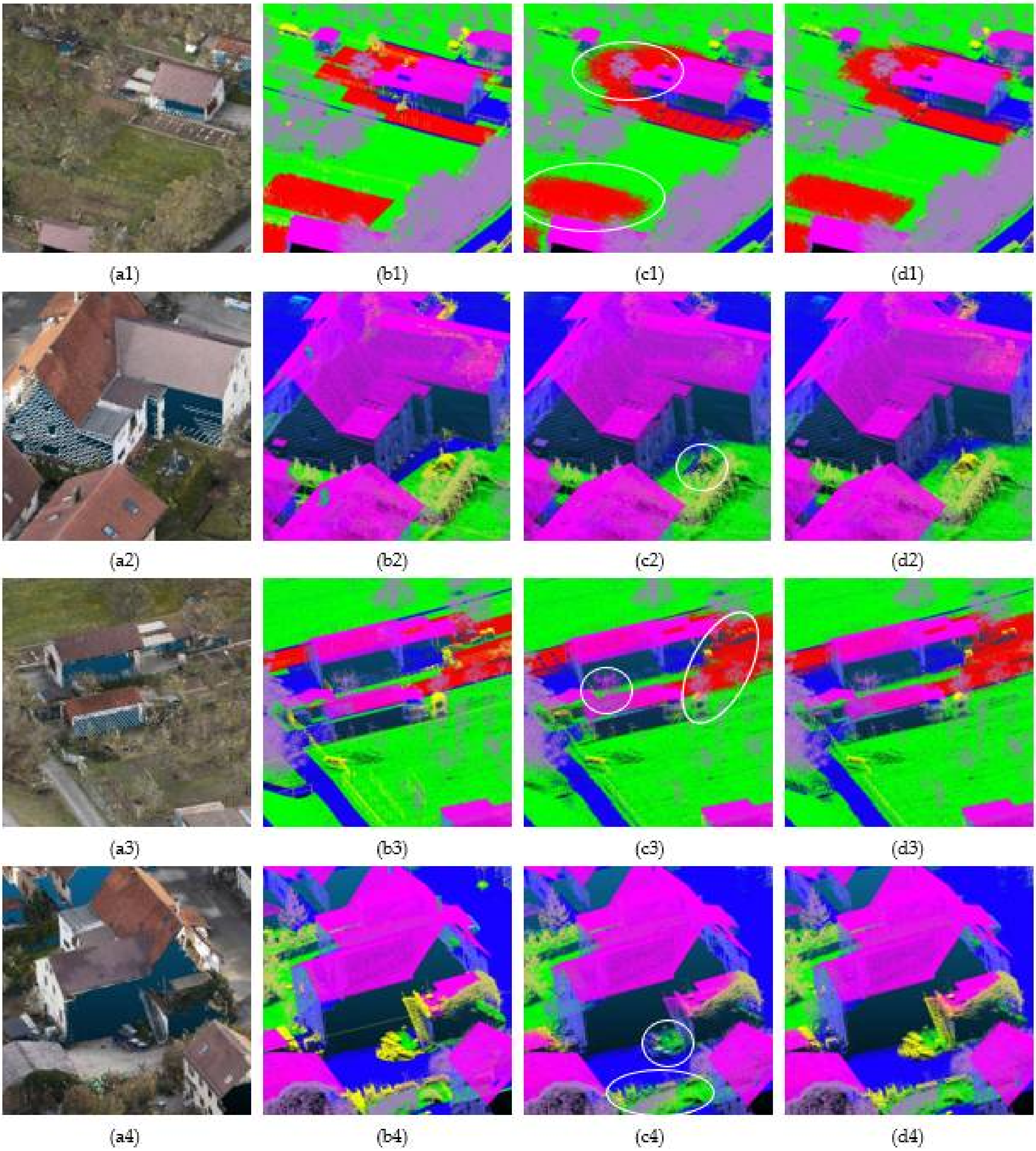

Figure 8. 3D semantic segmentation on the H3D dataset. (**a1**) Data of Scene 1; (**b1**) Ground Truth; (**c1**) PAConv; (**d1**) PIIE-DSA-net; (**a2**) Data of Scene 2; (**b2**) Ground Truth; (**c2**) PAConv; (**d2**) PIIE-DSA-net; (**a3**) Data of Scene 3; (**b3**) Ground Truth; (**c3**) PAConv; (**d3**) PIIE-DSA-net; (**a4**) Data of Scene 4; (**b4**) Ground Truth; (**c4**) PAConv; (**d4**) PIIE-DSA-net.

For the published testing dataset of H3D (without labels), we used PAConv and PIIE-DSA-net obtaining the testing data and submitted them to the official website of H3D and obtained the evaluation results. Since the ground truth of the testing set of H3D is not available to us, we do not show the visualized results. The results are shown in Table 7, where only data accurate to 1% are given by their website. Apart from the given OA, we also calculated *mIoU* by ourselves. A similar conclusion can be found in that, although PAConv won in some of the specific categories, PIIE-DSA-net greatly improved the overall performance compared with PAConv. Better segmentation completeness on the areas of the same category with more points was obtained by PIIE-DSA-net.

Table 7. The experiments on the H3D dataset (%).

Categories	PAConv	PIIE-DSA-Net
Low vegetation	74	81
Impervious Surface	90	84
Vehicle	66	75
Urban Furniture	43	48
Roof	94	94
Facade	77	71
Shrub	55	63
Tree	96	95
Soll/Gravel	41	58
Vertical Surface	68	74
Chimney	100	85
mIoU	64.09	75.27
OA	74	81

4. Conclusions and Discussion

In this paper, we proposed PIIE-DSA-net for 3D semantic segmentation on urban indoor and outdoor datasets. Most of the recently published SOTA works of 3D semantic segmentation benefited from different novel feature augmentation strategies. However, they did not pay sufficient attention to low-level features, and the asymmetry between the length of the low-level features and deep features led to poor results of feature fusion and further segmentation. Our PIIE-DSA-net was based on PAconv. The PIIE module was employed to enhance the low-level features, and the DSA module was proposed to optimize the fusion of the extracted low-level features and deep features.

Overall, the results of the experiments on one indoor dataset and two outdoor datasets proved the reliability and advancement of PIIE-DSA-net. Compared with the original PAConv, PIIE-DSA-net had more reliable results on edge areas between different categories. Moreover, it was also more effective in the categories with few points. Furthermore, the segmentation completeness of PIIE-DSA-net was good on the areas of the same category with more points.

In the ablation experiments on both indoor and outdoor datasets, we found that the PIIE module had more contributions on the segmentation results, and the DSA module also improved the results. Moreover, the method of matrix transformation and single-head attention were more effective than other tricks.

Our work verified the importance of low-level features for 3D semantic segmentation. The idea of PIIE-DSA-net can be modified and used in other backbones for 3D segmentation. The feature augmentation methods of low-level features and the fusion methods of low-level and deep features can be researched in greater depth in the future. Finally, by optimizing the parameter settings, fully tuning and training PIIE-DSA-net may result in further potential improvement.

Author Contributions: Conceptualization, F.G. and Y.Y.; methodology, Y.Y.; software, H.L. and R.S.; validation, F.G. and Y.Y.; data curation, H.L. and R.S.; writing—original draft preparation, F.G.; writing—review and editing, F.G. and Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation, grant number 62071136.

Data Availability Statement: As illustrated in the statement of the datasets, the datasets used in our work are public datasets. The three datasets can be downloaded from following links. S3DIS dataset: http://buildingparser.stanford.edu/dataset.html#Download, accessed on 19 July 2022. SensatUrban dataset used in our experiments: https://drive.google.com/file/d/1ckFhM_Qe_j9YvxCUTurIchHg5f9vlbI7/view, accessed on 19 July 2022. H3D dataset: https://ifpwww.ifp.uni-stuttgart.de/benchmark/hessigheim/default.aspx, accessed on 19 July 2022. Our work can been download from Github: https://github.com/WrenchShi/PIIE-DSA-net, accessed on 4 July 2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hu, Q.; Wang, S.; Fu, C.; Ai, M.; Yu, D.; Wang, W. Fine Surveying and 3D Modeling Approach for Wooden Ancient Architecture via Multiple Laser Scanner Integration. *Remote Sens.* **2016**, *8*, 270. [CrossRef]
2. Siranec, M.; Höger, M.; Otcenásová, A. Advanced Power Line Diagnostics Using Point Cloud Data-Possible Applications and Limits. *Remote Sens.* **2021**, *13*, 1880. [CrossRef]
3. Çakir, A.; Akpancar, S. 3D Simultaneous Positioning and Mapping in Dark, Closed Spaces with an Autonomous Flying Robot. *Acta Polytech. Hung.* **2020**, *17*, 7–23. [CrossRef]
4. Li, Y.; Ma, L.; Zhong, Z.; Liu, F.; Cao, D.; Li, J.; Chapman, M.A. Deep Learning for LiDAR Point Clouds in Autonomous Driving: A Review. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 3412–3432. [CrossRef] [PubMed]
5. Chen, Y.; Liu, G.; Xu, Y.; Pan, P.; Xing, Y. PointNet++ Network Architecture with Individual Point Level and Global Features on Centroid for ALS Point Cloud Classification. *Remote Sens.* **2021**, *13*, 472. [CrossRef]
6. Elsner, P.; Dornbusch, U.; Thomas, I.; Amos, D.F.; Bovington, J.T.; Horn, D. Coincident beach surveys using UAS, vehicle mounted and airborne laser scanner: Point cloud inter-comparison and effects of surface type heterogeneity on elevation accuracies. *Remote Sens. Environ.* **2018**, *208*, 15–26. [CrossRef]
7. Mathias, L. Mobile Laser Scanning Point Clouds. Gim International. Available online: https://www.gim-international.com/content/article/mobile-laser-scanning-point-clouds (accessed on 3 August 2017).
8. Zhu, J.; Xu, Y.; Ye, Z.; Hoegner, L.; Stilla, U. Fusion of urban 3D point clouds with thermal attributes using MLS data and TIR image sequences. *Infrared Phys. Technol.* **2021**, *113*, 103622. [CrossRef]
9. Babahajiani, P.; Fan, L.; Kämäräinen, J.; Gabbouj, M. Comprehensive Automated 3D Urban Environment Modelling Using Terrestrial Laser Scanning Point Cloud. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 652–660.
10. Poli, D.; Caravaggi, I. 3D modeling of large urban areas with stereo VHR satellite imagery: Lessons learned. *Nat. Hazards* **2013**, *68*, 53–78. [CrossRef]
11. Xie, Y.; Tian, J.; Zhu, X.X. Linking Points With Labels in 3D: A Review of Point Cloud Semantic Segmentation. *IEEE Geosci. Remote Sens. Magzine* **2020**, *8*, 38–59. [CrossRef]
12. Bello, S.A.; Yu, S.; Wang, C.; Adam, J.M.; Li, J. Review: Deep learning on 3D point clouds. *Remote Sens.* **2020**, *12*, 1729. [CrossRef]
13. Han, X.; Jin, J.S.; Wang, M.; Jiang, W.; Gao, L.; Xiao, L. A review of algorithms for filtering the 3D point cloud. Signal Process. *Image Commun.* **2017**, *57*, 103–112.
14. Cheng, S.; Chen, X.; He, X.; Liu, Z.; Bai, X. PRA-Net: Point Relation-Aware Network for 3D Point Cloud Analysis. *IEEE Trans. Image Process.* **2021**, *30*, 4436–4448. [CrossRef]
15. Chen, Y.; Liu, X.; Xiao, Y.; Zhao, Q.; Wan, S. Three-Dimensional Urban Land Cover Classification by Prior-Level Fusion of LiDAR Point Cloud and Optical Imagery. *Remote Sens.* **2021**, *13*, 4928. [CrossRef]
16. Wang, Y.; Shi, T.; Yun, P.; Tai, L.; Liu, M. PointSeg: Real-Time Semantic Segmentation Based on 3D LiDAR Point Cloud. *arXiv* **2018**, arXiv:1807.06288.
17. Milioto, A.; Vizzo, I.; Behley, J.; Stachniss, C. RangeNet ++: Fast and Accurate LiDAR Semantic Segmentation. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 4213–4220.
18. Lyu, Y.; Huang, X.; Zhang, Z. Learning to Segment 3D Point Clouds in 2D Image Space. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 12252–12261.
19. Poux, F.; Billen, R. Voxel-based 3D Point Cloud Semantic Segmentation: Unsupervised Geometric and Relationship Featuring vs Deep Learning Methods. *ISPRS Int. J. Geo Inf.* **2019**, *8*, 213. [CrossRef]
20. Liu, Z.; Tang, H.; Lin, Y.; Han, S. Point-Voxel CNN for Efficient 3D Deep Learning. *arXiv* **2019**, arXiv:1907.03739.

21. Graham, B.; Engelcke, M.; Maaten, L.V. 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9224–9232.
22. Le, T.; Duan, Y. PointGrid: A Deep Network for 3D Shape Understanding. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9204–9214.
23. Meng, H.; Gao, L.; Lai, Y.; Manocha, D. VV-Net: Voxel VAE Net With Group Convolutions for Point Cloud Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 10–17 October 2019; pp. 8499–8507.
24. Triess, L.T.; Peter, D.; Rist, C.B.; Zöllner, J.M. Scan-based Semantic Segmentation of LiDAR Point Clouds: An Experimental Study. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 19 October–13 November 2020; pp. 1116–1121.
25. Qi, C.; Su, H.; Mo, K.; Guibas, L.J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 77–85.
26. Qi, C.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2017; Volume 30.
27. Huang, Q.; Wang, W.; Neumann, U. Recurrent Slice Networks for 3D Segmentation of Point Clouds. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2626–2635.
28. Zhao, H.; Jiang, L.; Fu, C.; Jia, J. PointWeb: Enhancing Local Neighborhood Features for Point Cloud Processing. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5560–5568.
29. Zhang, Z.; Hua, B.; Yeung, S. ShellNet: Efficient Point Cloud Convolutional Neural Networks Using Concentric Shells Statistics. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 10–17 October 2019; pp. 1607–1616.
30. Qian, G.; Hammoud, H.A.; Li, G.; Thabet, A.K.; Ghanem, B. ASSANet: An Anisotropic Separable Set Abstraction for Efficient Point Cloud Representation Learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2021; Volume 34, pp. 28119–28130.
31. Ran, H.; Liu, J.; Wang, C. Surface Representation for Point Clouds. *arXiv* **2022**, arXiv:2205.05740.
32. Qian, G.; Li, Y.; Peng, H.; Mai, J.; Hammoud, H.A.; Elhoseiny, M.; Ghanem, B. PointNeXt: Revisiting PointNet++ with Improved Training and Scaling Strategies. *arXiv* **2022**, arXiv:2206.04670.
33. Yan, X.; Zheng, C.; Li, Z.; Wang, S.; Cui, S. PointASNL: Robust Point Clouds Processing Using Nonlocal Neural Networks With Adaptive Sampling. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5588–5597.
34. Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B. PointCNN: Convolution On X-Transformed Points. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2018; Volume 31.
35. Thomas, H.; Qi, C.; Deschaud, J.; Marcotegui, B.; Goulette, F.; Guibas, L.J. KPConv: Flexible and Deformable Convolution for Point Clouds. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 10–17 October 2019; pp. 6410–6419.
36. Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, A.; Markham, A. RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11105–11114.
37. Boulch, A. ConvPoint: Continuous convolutions for point cloud processing. *Comput. Graph.* **2020**, *88*, 24–34. [CrossRef]
38. Xu, M.; Ding, R.; Zhao, H.; Qi, X. PAConv: Position Adaptive Convolution with Dynamic Kernel Assembling on Point Clouds. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 3172–3181.
39. Deng, S.; Dong, Q. GA-NET: Global Attention Network for Point Cloud Semantic Segmentation. *IEEE Signal Process. Lett.* **2021**, *28*, 1300–1304. [CrossRef]
40. Chen, X.; Li, Y.; Fan, J.; Wang, R. RGAM: A novel network architecture for 3D point cloud semantic segmentation in indoor scenes. *Inf. Sci.* **2021**, *571*, 87–103. [CrossRef]
41. Geng, X.; Ji, S.; Lu, M.; Zhao, L. Multi-Scale Attentive Aggregation for LiDAR Point Cloud Segmentation. *Remote Sens.* **2021**, *13*, 691. [CrossRef]
42. Marsocci, V.; Scardapane, S.; Komodakis, N. MARE: Self-Supervised Multi-Attention REsu-Net for Semantic Segmentation in Remote Sensing. *Remote Sens.* **2021**, *13*, 3275. [CrossRef]
43. Chen, Z.; Li, D.; Fan, W.; Guan, H.; Wang, C.; Li, J. Self-Attention in Reconstruction Bias U-Net for Semantic Segmentation of Building Rooftops in Optical Remote Sensing Images. *Remote Sens.* **2021**, *13*, 2524. [CrossRef]
44. Li, J.; Chen, B.M.; Lee, G.H. SO-Net: Self-Organizing Network for Point Cloud Analysis. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9397–9406.
45. Zhao, H.; Jiang, L.; Jia, J.; Torr, P.H.; Koltun, V. Point Transformer. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 16239–16248.

46. Cheng, Z.; Wan, H.; Shen, X.; Wu, Z. PatchFormer: An Efficient Point Transformer with Patch Attention. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
47. Lai, X.; Liu, J.; Jiang, L.; Wang, L.; Zhao, H.; Liu, S.; Qi, X.; Jia, J. Stratified Transformer for 3D Point Cloud Segmentation. *arXiv* **2022**, arXiv:2203.14508.
48. Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S.E.; Bronstein, M.M.; Solomon, J.M. Dynamic Graph CNN for Learning on Point Clouds. *ACM Trans. Graph. (TOG)* **2019**, *38*, 1–12. [CrossRef]
49. Wang, C.; Samari, B.; Siddiqi, K. Local Spectral Graph Convolution for Point Set Feature Learning. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
50. Landrieu, L.; Boussaha, M. Point Cloud Oversegmentation With Graph-Structured Deep Metric Learning. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7432–7441.
51. Xie, L.; Furuhata, T.; Shimada, K. Multi-Resolution Graph Neural Network for Large-Scale Pointcloud Segmentation. *arXiv* **2020**, arXiv:2009.08924.
52. Lu, T.; Wang, L.; Wu, G. CGA-Net: Category Guided Aggregation for Point Cloud Semantic Segmentation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 11688–11697.
53. Qiu, S.; Anwar, S.; Barnes, N. Semantic Segmentation for Real Point Cloud Scenes via Bilateral Augmentation and Adaptive Fusion. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 1757–1767.
54. Robert, D.L.; Vallet, B.; Landrieu, L. Learning Multi-View Aggregation In the Wild for Large-Scale 3D Semantic Segmentation. *arXiv* **2022**, arXiv:2204.07548.
55. Tang, L.; Zhan, Y.; Chen, Z.; Yu, B.; Tao, D. Contrastive Boundary Learning for Point Cloud Segmentation. *arXiv* **2022**, arXiv:2203.05272.
56. Zhao, L.; Tao, W. JSNet: Joint Instance and Semantic Segmentation of 3D Point Clouds. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
57. Jiang, L.; Zhao, H.; Liu, S.; Shen, X.; Fu, C.; Jia, J. Hierarchical Point-Edge Interaction Network for Point Cloud Semantic Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 10–17 October 2019; pp. 10432–10440.
58. Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-Attention with Relative Position Representations. *arXiv* **2018**, arXiv:1803.02155.
59. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
60. Voita, E.; Talbot, D.; Moiseev, F.; Sennrich, R.; Titov, I. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. *arXiv* **2019**, arXiv:1905.09418.
61. Armeni, I.; Sener, O.; Zamir, A.R.; Jiang, H.; Brilakis, I.K.; Fischer, M.; Savarese, S. 3D Semantic Parsing of Large-Scale Indoor Spaces. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1534–1543.
62. Hu, Q.; Yang, B.; Khalid, S.; Xiao, W.; Trigoni, A.; Markham, A. Towards Semantic Segmentation of Urban-Scale 3D Point Clouds: A Dataset, Benchmarks and Challenges. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 4975–4985.
63. Kölle, M.; Laupheimer, D.; Schmohl, S.; Haala, N.; Rottensteiner, F.; Wegner, J.D.; Ledoux, H. The Hessigheim 3D (H3D) Benchmark on Semantic Segmentation of High-Resolution 3D Point Clouds and Textured Meshes from UAV LiDAR and Multi-View-Stereo. *arXiv* **2021**, arXiv:2102.05346. [CrossRef]