Semantic segmentation of terrestrial laser scanning point clouds using locally enhanced image-based geometric representations

Yuanzhi Cai, Lei Fan, Peter Atkinson and Cheng Zhang

***Abstract*—Point cloud data acquired using terrestrial laser scanning (TLS) often need to be semantically segmented to support many applications. To this end, various point-based, voxel-based and image-based methods have been developed. For large scale point cloud data, the former two types of methods often require extensive computational effort. In contrast, image-based methods are favorable from the perspective of computational efficiency. However, existing image-based methods are highly dependent on RGB information and do not provide an effective means of representing and utilizing the local geometric characteristics of point cloud data in images. This not only limits the overall segmentation accuracy, but also prohibits their application to situations where the RGB information is absent. To overcome such issues, this research proposes a novel image enhancement method to reveal the local geometric characteristics in images derived by the projection of the point cloud coordinates. Based on this method, various feature channel combinations were investigated experimentally. It was found that the new combination IZ_eD_e (i.e., intensity, enhanced Z coordinate and enhanced range images) outperformed the conventional IRGB and IRGBD channel combinations. As such, the approach can be used to replace the RGB channels for semantic segmentation. Using this new combination and the pre-trained HR-EHNet considered, a mean Intersection over Union (mIoU) of 74.2% and an Overall Accuracy (OA) of 92.1% were achieved on the Semantic3D benchmark, which sets a new state-of-the-art (SOTA) for the semantic segmentation accuracy of image-based methods.**

***Index Terms*—deep learning, point cloud, semantic segmentation, terrestrial laser scanning, transfer learning.**

I. Introduction

THE rapid development of three-dimensional (3D) data acquisition technologies has led to various types of sensors, such as terrestrial laser scanning (TLS) devices, RGB-D cameras and LiDAR [1]. Among these instruments, TLS stands out for its ability to quickly acquire large-volume (hundreds of millions of points per scan) and high-precision (millimeter level) point cloud data and is, therefore, used widely in applications where high-quality point cloud data are required. These may include, but are not limited to, 3D building reconstruction [2]–[6], vegetation and forest assessments [7]–[10], and cultural heritage management [11], [12].

In addition to the high-precision geometric information provided by TLS point clouds, semantic segmentation is often required as the basis for more complex purposes in the aforementioned applications. The goal of semantic segmentation of point clouds is mainly to annotate each data point with a semantic label, which is often based on the geometry, the reflection intensity and sometimes the color information provided by the data point itself and its neighbors. This can be achieved via traditional supervised classification methods [13]–[15] or deep learning approaches [16]–[20]. Compared to traditional classification methods using handcrafted features (e.g., support vector machines, random forests and conditional random fields), deep learning methods are becoming increasingly popular because they can automatically learn the feature representations needed for segmentation from raw data, avoid complex feature design, and typically result in higher segmentation accuracy [1], [21], [22].

Existing point cloud segmentation methods can be categorized into three major groups based on the form of the input data: point-based, voxel-based and image-based methods. The pioneering work on point-based methods is PointNet [16], which used shared Multi-Layer Perceptrons (MLPs) to learn pre-point features and used symmetrical pooling functions to learn global features. On the basis of PointNet, many other point-based networks have been proposed in recent years, which can be subdivided into pointwise MLP methods [17], [22], [23]–[31], graph-based methods [18], [32]–[37], point convolution methods [38]–[43], and RNN-based methods [44]–[46]. This class of algorithm can typically achieve high accuracy, and the state-of-the-art (SOTA) method is the RFCR [31] in this category, which achieved an Overall Accuracy (OA) of 94.3% and a mean Intersection over Union (mIoU) of 77.8% on the Semantic3D (reduced-8) [31], [47]. However, while point-based methods are focused on increasing the segmentation accuracy of point clouds, their high computational cost makes them too costly for practical application to large-scale TLS point clouds. For example, for a use case where the processing time was revealed [47], it ranges

This research was funded by XJTLU Key Program Special Fund (grant no. KSF-E-40), XJTLU Research Development Fund (grant no. RDF-18-01-40) and XJTLU Research Enhancement Funding (grant no. REF-21-01-003).

Y. Cai, L. Fan (corresponding author), and C. Zhang are with Department of Civil Engineering, Design School, Xi'an Jiaotong-Liverpool University, Suzhou, 215000, China (e-mail: yuanzhi.cai19@student.xjtlu.edu.cn; lei.fan@xjtlu.edu.cn; cheng.zhang@xjtlu.edu.cn).

P. Atkinson is with Faculty of Science and Technology, Lancaster University, Bailrigg, Lancaster LA1 4YW (e-mail: pma@lancaster.ac.uk)

from 10 to 50 minutes to process 4-point clouds containing 80 million points in Semantic3D (reduced-8).

For the second class of voxel-based methods [48]–[52], they first convert the point cloud into a dense/sparse discrete voxel representation and then apply the 3D convolutional neural network (CNN). Since 3D convolutional networks are extremely computationally intensive and consume significant amounts of Graphics Processing Unit (GPU) memory, such methods have to make careful trade-offs in terms of segmentation accuracy and processing time. From the published performance of these methods on various benchmark datasets [47], [53]–[56], such methods are not only less accurate than the first type of method, but also very slow in processing and, therefore, are considered unsuitable for processing large-scale TLS point cloud data.

The image-based methods utilize 2D convolutional neural networks (CNNs) to segment multi-channel images generated from point cloud data. There are two approaches for image generation. The first approach [57]–[59] projects point cloud data from multiple virtual camera views onto a plane, while the second approach [20], [60]–[62] projects the point cloud data as a panoramic image centered at the scanner. The second approach is more efficient than the multi-view ones because processing is limited to only one panoramic image for each point cloud obtained [20], [58]. Coupled with the use of 2D CNNs (much more efficient than those networks used in point-based and voxel-based methods), the panoramic images offer an extremely fast approach to segmenting point cloud data. For example, the SOTA image-based method [20] takes only 5.13s to process the Semantic3D (reduced-8) [47] test dataset. However, it was noticed that its segmentation accuracy [20] was relatively low compared to the SOTA point-based method RFCR [31], achieving an OA of only 89.4% and a mIoU of 63.5% on Semantic3D (reduced-8). Therefore, image-based methods are ideal for processing large-scale TLS point cloud data, but such methods available in the literature suffer from the problems elaborated in the next paragraph, which also form the likely basis for any further improvements in their segmentation accuracy.

Three types of information of TLS point clouds can be considered for semantic segmentation (i.e., geometric information (coordinates and their derivatives), intensity and RGB if images were taken). In the existing image-based methods, it was noticed that combinations of feature channels considered [20], [57]–[59] always included the RGB information, without which the segmentation accuracy degraded significantly. This is not surprising as the true colors include rich information about the objects to be segmented. However, this means that those methods are highly reliant on the RGB information and could not effectively handle the cases where the RGB information is missing (no images taken) or are mismatched to point clouds due to moving objects in the scene or the imperfect matching between images and point clouds taken separately. In addition, the geometric information was either not considered or not used in an effective way. In contrast, point-based and voxel-based methods perform well for point clouds with only coordinate information [18], [22], [41], [50], indicating that geometric features are valuable for point cloud semantic segmentation. Hence, it is reasonable to speculate that the application scope and segmentation accuracy of image-based approaches can be improved further if the geometric information contained in the point cloud is utilized effectively.

Therefore, under the umbrella of image-based methods, this study aims to improve and generalize this class of methods by considering the characterization of the geometric information of scenes/objects in the panoramic images derived from coordinates of point cloud data. The increase in accuracy relates to the semantic segmentation while the generalization refers to cases where the RGB information is missing in the point cloud data. To this end, an image enhancement method is proposed to characterize the local geometric features in the images. Based on the enhanced images, this research proposes a new combination of feature channels without the RGB information. In the CNN used for extracting the semantic information in this study, the Atrous Spatial Pyramid Pooling (ASPP) module [63] is considered to aggregate multi-scale high-level features from HRNet [64]. In the past studies [63], [65], [66], the aggregation was typically executed using coarse-resolution feature maps. However, in our study, the finest-resolution feature maps in HRNet are used for the aggregation, the outputs of which are concatenated with multiple low-level features for segmentation.

The main contributions of this research are the establishment of a new image enhancement method for characterizing effectively the local geometric features in the panoramic images derived from point clouds, and the finding that the utilization of those local geometric features can increase the segmentation accuracy of image-based methods. The approach proposed in this study offers a better alternative channel combination to replace those involving the RGB channels, which is very useful for cases where the RGB information is absent or inaccurate.

II. Methodology

The methodology considered in this research involves the following key steps. Firstly, the information (e.g., intensity and *XYZ* coordinates) contained in the unstructured point cloud data was projected into a multichannel panoramic image using the transformation relationship between the Cartesian coordinate system and the spherical coordinate system. Secondly, the local-based enhancement was applied to the panoramic image channels that contain geometric information such as *XYZ* coordinates and range. Lastly, semantic information was extracted from the panoramic image using a pre-trained customized CNN, and back-projected to the raw point cloud data to obtain semantically segmented point cloud. More detailed descriptions of these steps are provided in Sections II.C-II.F.

A. Study data

The large-scale Sementic3D dataset [47] was used to demonstrate and evaluate the proposed method, which contains a total of 30 labeled TLS point clouds collected at 10 different scenes. Point cloud data were labeled as eight classes, namely: made terrain, natural terrain, high vegetation, low vegetation,

buildings, hard scape, scanning artefacts and cars. The ground reference labels for 15 training point clouds are available from the dataset supplier. The online evaluation frequency of test set results is limited to once every three days. Therefore, except for Section III.D where the test set was used for comparison with the state-of-the-art results, all other experiments were conducted on the training set. More specifically, for Sections III.B-III.C, the performance of our method was evaluated by employing 5-fold cross-validation on the Semantic3D training dataset.

B. Segmentation accuracy metrics

To evaluate the segmentation performance, the same evaluation metrics as used in the Semantic3D online evaluation were used in this study, i.e., OA and mIoU. The OA metric is the ratio of correctly classified points (regardless of class) to the total number of points. The mIoU metric is the mean IoU of all classes. For class i, the IoU metric is the ratio of correctly classified pixels to the total number of ground reference data and predicted pixels in that class. The formulae for the aforementioned metrics are shown in Equations 1-3.

$$\mathrm{OA} = \frac{\mathrm{TP}}{\text{Total number of points}} \tag{1}$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP+FN+FP}} \tag{2}$$

$$\mathrm{mIoU} = \frac{\sum_{i=1}^{N} \mathrm{IoU}_i}{N} \tag{3}$$

where $\mathrm{TP}, \mathrm{FN}, \mathrm{FP}, i, N$ represent the true positive, false negative, false positive points classified, index of class and total number of classes, respectively.

In general, OA provides a quick and computationally inexpensive estimate of the percentage of correctly classified points, while mIoU provides a measurement of accuracy that not only penalizes false positives, but also increases the penalty against segmentation errors in small classes. Since the numbers of points contained in the eight classes of the Semantic3D benchmark dataset are highly imbalanced (shown in Fig.1), mIoU is considered more critical in this research.

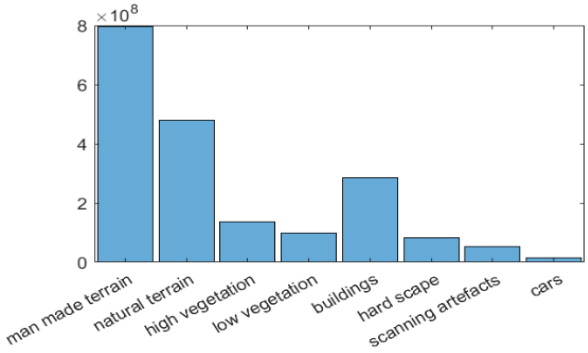


Fig. 1. The distribution of the classes of points in Semantic3D dataset.

C. *Point cloud to image projection*

Many terrestrial laser scanners collect point cloud data through vertically rotating optics that are mounted on a horizontally rotating base. Since their rotational steps are usually fixed throughout a single scan, the point cloud data obtained would theoretically have fixed inclination and azimuthal resolutions. These two resolutions are typically the same. In other words, if the point cloud data are considered as vectors originating from the origin (i.e., the scanner's optical center), these vectors will be uniformly distributed in a spherical space centered at the origin. Therefore, TLS point clouds are inherently suitable to be projected into spherical coordinate systems. Based on this, the following method for point cloud to image projection was used in this study, which is demonstrated using the example shown in Fig.2.a Firstly, the Cartesian coordinates of the point cloud data were transformed into spherical coordinates using Equations 4-6.

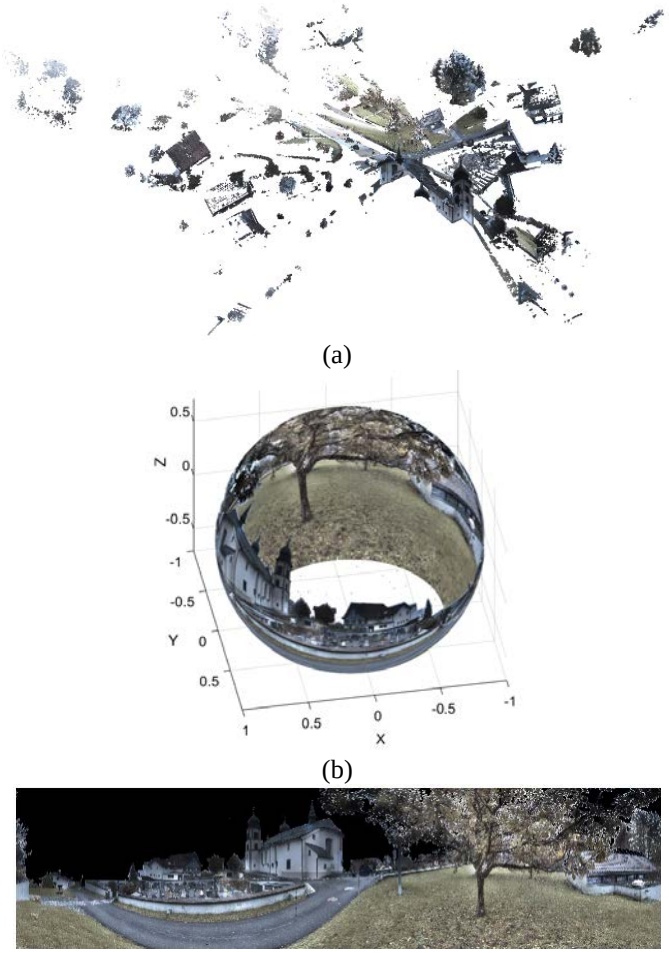


Fig. 2. Key stages in the projection process: (a). The raw input point cloud [47], (b). All points scaled to a spherical surface at a distance of 1 from the origin (i.e., the center of the scanner), (c). The panoramic image rasterized from the spherical surface.

$$range\ (r) = \sqrt{x^2 + y^2 + z^2} \tag{4}$$

$$\text{inclination}\ (\theta) = arccos\frac{z}{r} \tag{5}$$

$$\text{azimuth}\ (\varphi) = arctan\frac{y}{x} \tag{6}$$

Secondly, the position of each data point in the unit spherical surface (i.e., "continuous" spherical image) is determined by its inclination θ and azimuth φ, as shown in Fig.2.b Thirdly, by using a specific angular resolution ω to discretize the "continuous" spherical image, a rasterized spherical image is obtained. To ensure the image continuity, the image angular resolution should be slightly larger than the scanner angular resolution. Finally, by mapping the available information (e.g., RGB, intensity, range) to the rasterized spherical image and splitting it from a certain azimuth (e.g.,180° used in the

subsequent experiments), the multichannel panoramic image is obtained (e.g., the RGB panoramic image in Fig.2.c). More specifically, for a data point of the inclination θ and the azimuth φ in the spherical coordinate system, its pixel location in the panoramic image is determined using Equation 7.

$$\left(\left\lceil\frac{90-\theta}{\omega}\right\rceil, \left\lceil\frac{180-\varphi}{\omega}\right\rceil\right) \quad (7)$$

where the former element represents the row location for the inclination θ, the latter element represents the column location for the azimuth φ, ω is the angular resolution, $\lceil x \rceil$ rounds x to the nearest integer greater than or equal to x.

Because of the fine angular resolutions of laser scanners, the resolution of the projected panoramic image could be ultra-high. For example, the equivalent panoramic image size of the point cloud captured using the RTC360's finest resolution is 8333×20334 pixels.

During the point cloud to image projection, it is often the case that a single image pixel contains multiple data points. In this case, the pixel values in the panoramic feature image (e.g., RGB image) were taken as the average values of multiple data points, while the pixel values (labeled classes) in the labeled panoramic image (labeled image used for training) were taken as the ones corresponding to the rarest class to increase network segmentation accuracy regarding the imbalanced class (typically, the class with fewer data is harder to segment).

D. Enhancement of image-based geometric features

As shown in Fig.3.a, the panoramic RGB image is relatively clear. However, objects in the grayscale images obtained by projecting the *XYZ* coordinates and the range information were not shown clearly, such as the panoramic image of the *Z* coordinate shown in Fig.3.b. Due to this phenomenon, existing image-based methods [20], [57], [58] rely mainly on the RGB information, and this type of grayscale images was usually used as auxiliary information only.

By comparing the pixel value distribution histograms (Fig.3.d and Fig.3.f) of the RGB image (Fig.3.c) and *Z*-coordinate image (Fig.3.e) for the same local area (area within the 256*256 white box in Fig.3.a and Fig.3.b), it was found that the distribution of grayscale values of the *Z* coordinate image was extremely concentrated compared to the RGB image. This is due to the fact that the range of variation in the coordinates of adjacent local data points is relatively small compared to that of the whole dataset. Based on this observation and the fact that CNNs are good at learning local features rather than global ones, the proposed enhancement method is local-based and its detailed description is presented as follows.

Firstly, for a given local area, the grayscale values are redistributed so that their histogram conforms to the Rayleigh Distribution defined in Equation 8.

$$f(z) = \frac{z}{\sigma^2} e^{\left(-\frac{z^2}{2\sigma^2}\right)}, z \geq 0 \quad (8)$$

where the value of σ is taken as 0.4 so that the expected value of mean grayscale values is 0.5. After this local enhancement was applied, the "hidden" geometrical features in Fig.3.e are revealed clearly in Fig.3.g, and the corresponding redistributed histogram is shown in Fig.3.h. Intuitively, the enhanced *Z* coordinate image (Fig.3.g) contains many detailed geometric

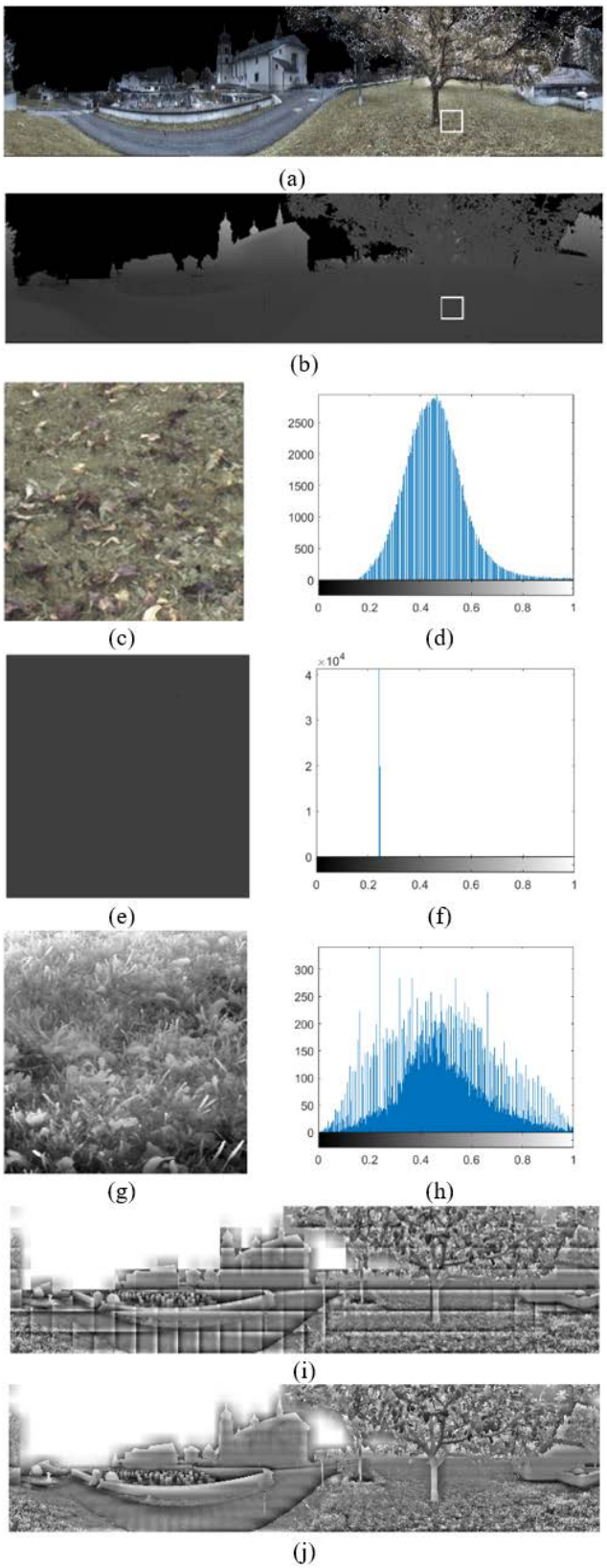


Fig. 3. Illustrations of image enhancement effects: (a). The panoramic image projected from RGB channels, (b). The panoramic image projected from *Z* coordinate, (c). A local RGB image extracted from the box in (a), (d). The distribution histogram of the pixel values in (c), (e). The local *Z* coordinate image extracted from the box in (b), (f). The distribution histogram of the pixel values in (e), (g). The enhanced local *Z* coordinate image. (h) The distribution histogram of the pixel values in (g), (i). The enhanced *Z* coordinate image without overlapping. (j). The enhanced *Z* coordinate image with overlapping.

features that are distinct from the RGB image in Fig.3.c.

In the above example, the local enhancement method essentially magnifies the *Z* coordinate differences within the local area. However, if there is a general trend for the values within adjacent local areas, applying the local enhancement method individually to each area will result in discontinuous pixel values at the edges of the local areas. For example, the *Z*-values of the grass area on the right side of Fig.3.a gradually increases from the bottom to the top. If the local enhancement method is applied without overlap (the sizes of the local areas are taken as 256*256 pixels) between two adjacent local areas, the bottom pixels of the top local area (e.g., Fig.3.g) are set close to black and the top pixels of the bottom local area (i.e., the local area right below the area representing by Fig.3.g) are set close to white. This leads to those horizontal edge discontinuities on the right side of Fig.3.i. This phenomenon is the reason for choosing the Rayleigh distribution instead of a uniform distribution in this research. In general, an image with a uniformly distributed histogram will contain the most information [67]. However, adopting the uniformly distributed histogram means that more points will be distributed close to the two extremes (i.e., zero or one), which will exacerbate the discontinuity at the edges.

To minimize the edge discontinuity, an overlapped local enhancement was used in this study. More specifically, the panoramic image was firstly divided into square areas of the same size that overlap each other by one-eighth of the edge length, and the local enhancement method was applied to each square area. During this process, symmetric padding was used to fill in the blank areas when the actual image area was insufficient. Finally, for the overlapping part, the pixel values were taken as the average of the values of the overlapped pixels. The *Z* coordinate image enhanced using this method is shown in Fig.3.j, where the size of the local square area was taken as 256*256 pixels (same as for Fig.3.i) for this example. It can be observed that the edge discontinuity was effectively mitigated by the overlapping strategy. It should be noticed that the size of the local area has a significant effect on the final enhanced image, and the selection of a proper size is demonstrated in Section III.B.

E. Semantic segmentation network structure

To obtain the semantic information from the fine-resolution panoramic images, a customized CNN was adopted in this research, which consists of two parts: a backbone and a segmentation head. The entire network structure is shown in Fig.4, which is named as HR-EHNet to indicate that it is designed for the segmentation of fine-resolution enhanced panoramic images.

The backbone part is responsible for extracting features from the input images [64]. Although there are various backbone structures available [64], [68]–[73], only HRNet was designed for processing fine-resolution images [64], which has widely been adopted for excellent semantic segmentation results [74]–[77]. As such, it was adopted in this study. More specifically, the HRNet_W48 version (larger version) was adopted, where the number 48 indicates the network width of the finest resolution branch. The basic network structure of HRNet is depicted in Fig.4. Different from mostly used single-branch backbones [78], the HRNet has four parallel branches corresponding to four downsample levels (4, 8, 16, and 32, respectively). As for the width of the network (i.e., the number of feature map channels/ the number of convolutional kernels), HRNet adopts a scheme where the number of channels is doubled accordingly whenever the resolution of a feature map decreases [64]. Compared to single-branch backbones, HRNet increases significantly the network depth (i.e., the number of convolutional layers) with respect to fine-resolution features, and meanwhile retains coarse-resolution features to provide global contextual information. Since a deeper network structure extends the receptive field and enhances the discrimination of each pixel, the fine-resolution segmentation task could benefit from the deep fine-resolution branch in HRNet.

The segmentation head is responsible for interpreting the extracted features from the backbone to assign an appropriate label to each pixel. The ASPP segmentation head was adopted in this study, which was first proposed by [63] and adopted widely by others [20], [65], [66], [79]. The ASPP module

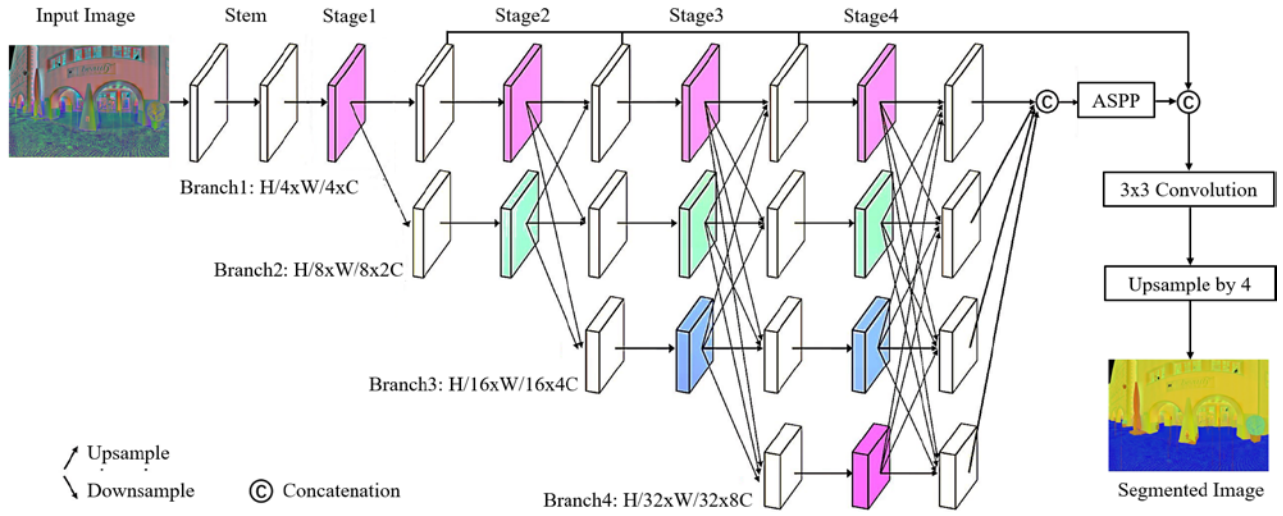


Fig. 4. Illustration of the HR-EHNet network structure: upsampling and downsampling were implemented by bilinear interpolation and strided 3x3 convolution, respectively; The colored blocks that represent multiple residual convolution operations were performed.

employs several parallel atrous (dilated) convolutions with different dilation rates to extract semantic information from different spatial scales [65]. The commonly used output stride for the ASPP module is 16 or 8 (16 most commonly in the literature), which means that its input resolution corresponds to a downsampling level of 16 or 8, respectively. This is because most of the backbones are single-branch structures, which generate only high-level features at a relatively high downsampling level. This is not the case for HRNet. Therefore, the ASPP module is attached to the end of the first branch (corresponding to a downsampling level of 4) to take advantage of the fine-resolution features in HR-Net. It was ascertained in previous research [65], [66] that the proper dilation rate combination for ASPP with an output stride of 16 includes 6, 12 and 18, which should be multiplied by 2 (i.e., 12, 24 and 36) when an output stride of 8 was used. Hence, for an output stride of 4, the dilation rate combination is taken as 24, 48 and 72 in this research. Finally, similar with the DeeplabV3+ [65], the output of ASPP is concatenated with three groups of low-level features (corresponding to the outputs of the first three stages of the first branch) for the final segmentation.

F. Pretraining of network and transfer learning

For image semantic segmentation, it is a consensus that a higher segmentation accuracy can be obtained using pre-trained networks [80]–[82]. This step was also employed in this research where the Cityscapes dataset [83] was used for network pretraining. Similar to Semantic3D, Cityscapes was focused on semantic segmentation in urban scenes and was collected mainly in Europe. Cityscapes contains 5,000 finely labeled fine-resolution RGB images, which were originally divided into 2975, 500, 1525 images for training, validation and testing, respectively [83]. However, since it is beneficial to use a larger dataset for the pretraining, all the training and validation images were used as the training set in this study. Pixels in these images are labeled into 30 classes. Compared to Semantic3D, Cityscapes covers a wider range of urban scenes, has a greater variety of annotations, and suffers from a greater class imbalance.

The training protocol for conducting pre-training followed previous research [63], [64], [84], [85]. The stochastic gradient descent with momentum (SGDM) optimizer was adopted. The base learning rate, the momentum and the weight decay were set to 0.01, 0.9, and 0.0005, respectively. The poly learning rate policy was used for dropping the learning rate, where the power was set to 0.9. The focal loss function [86] was adopted to address the issues of imbalanced classes. The size of the input images was set as 512*1024 pixels. The images were augmented by random cropping, random resize (0.5~2) and random horizontal flipping. Finally, HR-EHNet was trained for 180,000 iterations with a mini-batch size of 8 and synchronized batch normalization.

Since HR-EHNet was pre-trained using the RGB images of Cityscapes, the number of convolutional kernel channels in the first convolutional layer was three, which accepts only three-channel images as its input. However, subsequent experiments in Section III.B-III.C need to use input images with various numbers of channels for comparison. Therefore, in those experiments, the first convolutional layer of the pre-trained HR-EHNet was replaced by a new convolutional layer where its kernel channel number is equal to the number of input features. Meanwhile, the channel number of the convolutional kernels in the last two convolutional layers of the pre-trained HR-EHNet corresponds to the total number of classes (i.e., 19) for Cityscapes. This was replaced by new convolution layers with kernels of 8 channels to accommodate the number of classes in Sementic3D. The weights in these convolution layers were initialized randomly. When HR-EHNet was fine-tuned using the images generated from Semantic3D, almost the same training protocols as those in pre-training were used, expect that the iteration numbers were reduced to 60,000 and 75,000 for the training with five-fold cross-validation and for completing the training with the training set, respectively.

III. EXPERIMENT AND RESULTS

A. Information loss from point clouds to images

One of the most frequently quoted drawbacks of image-based approaches is the inevitable information loss during the process where point cloud data are projected to images [87]. However, based on the literature surveyed in this research, no previous studies have quantitatively evaluated the information loss in that process. Therefore, a quantitative analysis of information loss was carried out for the projection method proposed in the first place. In this study, the degree of information loss was quantified by comparing the labeling information of the Semantic3D training dataset before and after a complete projection process (i.e., point cloud to image, followed by image to point cloud), in which OA and mIoU were used as the

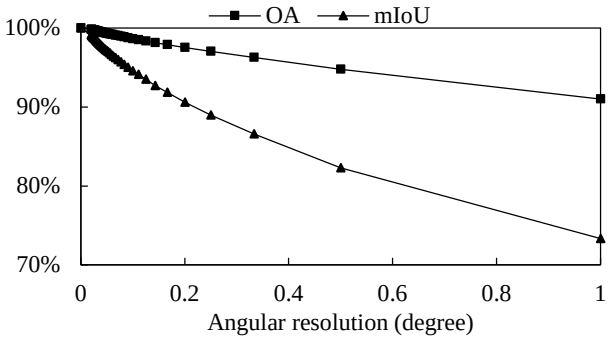
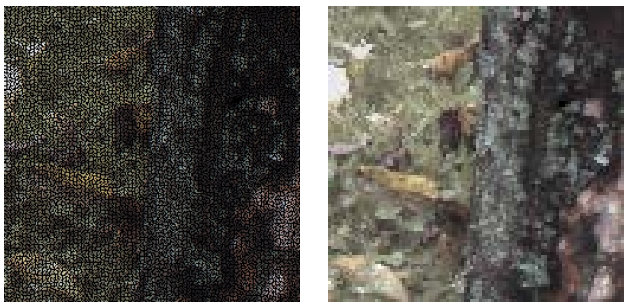


Fig. 5. Plot of accuracy (OA and mIoU) versus angular resolution.



(a) (b)

Fig. 6. Effects of an excessive angular resolution on the projected image: (a). Many black empty pixels for an angular resolution of 1/50 degree, (b). A continuous image without empty pixels for an angular resolution of 1/20 degree.

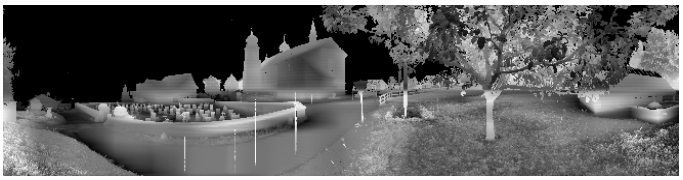
evaluation metrics. Following the projection process described in Section II.C and using a set of angular resolutions equal to 1/*n* degree (where *n* equals 1, 2, 3, ... 50), the corresponding OA and mIoU were recorded and shown in Fig.5.

It is observed in Fig.5 that OA and mIoU decreased gradually with increasing angular resolution, and the decreasing rate of mIoU was much higher. Although there was almost no information loss when extremely small angular resolutions were used (e.g., $OA = 0.998, mIoU = 0.991$ for the angular resolutions of 1/50 degree), this will result in excessive computational demands for subsequent image processing and leave many noisy blank pixels in projected images (e.g., Fig.6.a). The angular resolution of 1/20 degree (i.e., an image size of 3600*7200 pixels) was used in this research to perform the point cloud-image projection, as it can provide visually clean projected images (e.g., Fig.6.b) with a relatively low information loss ($OA = 0.993, mIoU = 0.97$).

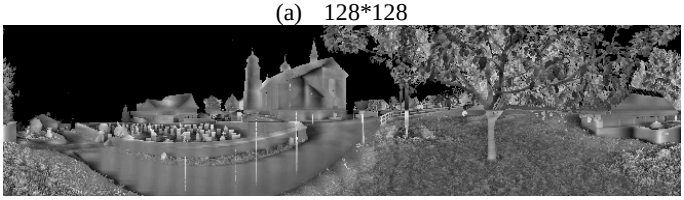
B. Effect of local enhancement area on the segmentation results

As mentioned in Section II.D, the size of the local square area used during enhancement has an impact on the enhanced images produced. For example, the enhanced images of the *Z* coordinate (i.e., Fig.3.b) using a local area of 128*128, 32*32, and 8*8 pixels were shown in Fig.7a, 7b and 7c, respectively, in addition to that using a size of 256*256 pixels in Fig.3.j. It is seen that there are notable differences in the enhanced images when different local patch sizes are used.

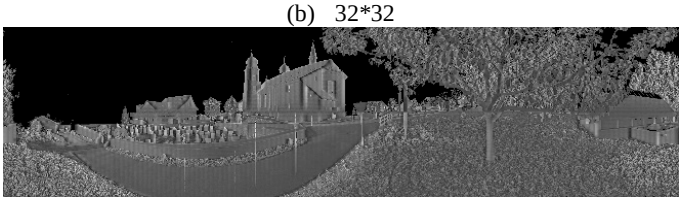
To determine an appropriate local patch size for enhancement and to test its effects on the segmentation results, an experiment was conducted for eight local patch sizes (8*8, 16*16... 1024*1024, i.e., $2^{3\sim10} * 2^{3\sim10}$). Four groups of original grayscale images were used in this experiment, which were projected from *XYZ* coordinates and range (*D*) in Semantic3D, respectively. Each group contain 15 images (corresponding to 15 training point clouds) with a size of 3600*7200 pixels (i.e., an angular resolution of 1/20 degree).



(a) 128*128

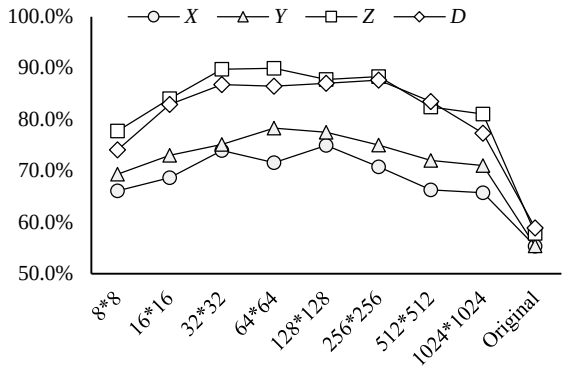


(b) 32*32

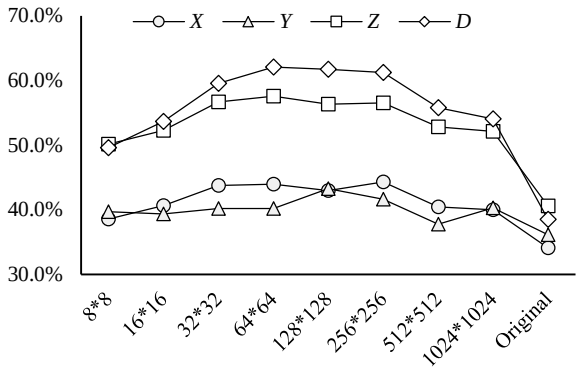


(c) 8*8

Fig. 7. Impacts of the local enhancement area on the enhancement results: (a). 128*128 pixels, (b). 32*32 pixels, (c). 8*8 pixels.



(a)



(b)

Fig. 8. Impacts of the local enhancement area on the enhancement results: (a). OA, (b). mIoU.

These original images were enhanced using each of the eight different sizes, leading to a total of 32 groups of enhanced images. The pre-trained HR-EHNet was fine-tuned on these 36 groups of single-channel images, respectively. The segmentation performances of each group are shown in Fig.8.

From Fig.8, it is seen that the segmentation accuracy of the network was significantly increased by using the image enhancement in this experiment. However, it was also noticed that the accuracies of the networks trained using the enhanced images derived from *X or Y* coordinates were considerably less than those based on *Z* coordinates and range in terms of both OA and mIoU metrics. Therefore, these two types of information (i.e., *X* and *Y* coordinates) were not considered in the subsequent sections. In addition, it is observed that for the images derived from *Z* coordinates and range, the OA and mIoU metrics were relatively similar when the image enhancement was performed using local area sizes from 32*32 pixels to 256*256 pixels. This suggests that the local area size for the image enhancement does not require careful adjustments as long as it is within that range. Nevertheless, since it can be seen from Fig.8.b that the images obtained from *Z* coordinates and range with a local area size of 64*64 pixels produced the highest mIoU index, this size was selected for this research.

C. Selecting combinations of feature channels

In this section, various combinations of the channels were tested, including the enhanced *Z* coordinate images (Z_e), enhanced range images (D_e), and intensity images (*I*) where the raw intensity values of Semantic3D dataset were used without

TABLE I
QUANTITATIVE RESULTS OF DIFFERENT CHANNEL COMBINATIONS ON THE SEMANTIC3D TRAINING SET (FIVE-FOLD CROSS-VALIDATION)

Channels	Index	mIoU	OA	man-made	natural	high veg	low veg	buildings	hard scape	scanning art	cars
Z_eD_e	1	68.4	89.3	85.3	75.0	81.9	41.3	**95.3**	33.7	42.2	**92.5**
IZ_e	2	66.4	90.1	86.5	75.1	68.3	45.3	93.4	26.8	49.2	86.3
ID_e	3	64.5	88.7	85.5	73.9	71.8	24.0	93.6	27.8	**51.5**	88.1
IZ_eD_e	4	**70.8**	**91.9**	86.4	77.7	**88.5**	**60.6**	94.2	37.3	43.5	77.8
*I*RGB	5	63.8	90.0	85.2	76.5	80.5	39.6	92.7	31.4	33.7	71.0
*I*RGB*D*	6	66.0	90.4	85.4	74.4	74.6	31.9	93.0	**45.2**	41.5	82.0
*I*RGBZ_eD_e	7	68.8	90.9	**86.5**	**78.7**	83.7	40.6	95.2	41.3	41.9	82.5
*I*RGB*D*Z_eD_e	8	68.7	90.6	86.4	76.9	81.8	51.0	94.8	36.9	43.5	78.0

any corrections. This is followed by tests on conventional combinations involving RGB channels (*I*RGB*D* and *I*RGB) that were demonstrated to be relatively accurate channel combinations in previous studies [20]. In addition, the combinations of Z_e and D_e with *I*RGB and *I*RGB*D* were tested. A total number of eight combinations of channels were investigated in this research. The test results are shown in Table I. Based on the first four sets of experiments, it is observed that using *I*, Z_e, and D_e together is more accurate than any combination of two of them. In addition, it is clear that the segmentation accuracy achieved by the IZ_eD_e combination was significantly higher than those achieved by the *I*RGB and the *I*RGB*D* combinations. For comparisons of the segmentation accuracy with respect to each class, the segmentation accuracy of IZ_eD_e was found to be higher than the other two combinations (*I*RGB and *I*RGB*D*) in most of the classes, especially in recognizing high vegetation and low vegetation. It was also found that the integration of Z_eD_e to *I*RGB or *I*RGB*D* significantly increased their segmentation accuracy in comparison to *I*RGB or *I*RGB*D* alone, but both cases failed to exceed the segmentation accuracy (mIoU and OA) achieved by the combination IZ_eD_e. However, it was also observed that IZ_eD_e did not perform best for some individual classes. The likely reasons are presented in the following. An individual channel may be favorable to the segmentation of a particular class. However, when multiple channels are combined, their interactions also play an important role in the segmentation accuracy of that particular class. In other words, the network will take into account the trade-off between the contribution of each channel (similar to a weighted average effect) to achieve a higher overall segmentation accuracy for all classes. Consequently, the accuracy of the segmented results of individual classes with or without the use of a particular channel may vary from one to another.

D. Final performance of HR-EHNet

Based on the experimental results in Table I, the channel

TABLE II
IMPACTS OF RETAINING OR REPLACING THE FIRST LAYER OF THE PRE-TRAINED NETWORK ON THE SEGMENTATION RESULTS WHEN IZ_ED_E WERE USED AS THE INPUT CHANNELS (FIVE-FOLD CROSS-VALIDATION)

First layer	mIoU	OA	man-made	natural	high veg	low veg	buildings	hard scape	scanning art	cars
Replaced	70.8	**91.9**	**86.4**	**77.7**	88.5	**60.6**	94.2	37.3	43.5	77.8
Remain	**73.1**	91.6	85.5	76.1	**89.3**	57.3	**95.1**	**46.8**	**46.8**	**88.2**

TABLE III
QUANTITATIVE RESULTS (%) OF DIFFERENT APPROACHES ON SEMANTIC3D (REDUCED-8)

		Time (s)	Params (M)	mIoU	OA	man-made.	natural.	high veg.	low veg.	buildings	hard scape	Scanning art.	cars
Point-based Methods	RF MSSF [23]	1643.75	-	62.7	90.3	87.6	80.3	81.8	36.4	92.2	24.1	42.6	56.6
	ShellNet [30]	3000	0.48	69.3	93.2	96.3	90.4	83.9	41	94.2	34.7	43.9	70.2
	OctreeNet [52]	184.84	-	59.1	89.9	90.7	82.0	82.4	39.3	90.0	10.9	31.2	46.0
	GACNet [34]	1380	-	70.8	91.9	86.4	77.7	88.5	**60.6**	94.2	37.3	43.5	77.8
	SPGraph [18]	3000	0.25	73.2	94	**97.4**	**92.6**	87.9	44	83.2	31.0	63.5	76.2
	KPConv [41]	600	14.9	74.6	92.9	90.9	82.2	84.2	47.9	94.9	40.0	**77.3**	79.7
	RandLA-Net [22]	-	0.95	77.4	**94.8**	95.6	91.4	86.6	51.5	**95.7**	**51.5**	69.8	76.8
	RFCR [31]	-	-	**77.8**	94.3	94.2	89.1	85.7	54.4	95	43.8	76.2	83.7
Projection-based Methods	DeePr3SS [57]	-	134	58.5	88.9	85.6	83.2	74.2	32.4	89.7	18.5	25.1	59.2
	SnapNet [58]	3600	29	59.1	88.6	82.0	77.3	79.7	22.9	91.1	18.4	37.3	64.4
	XJTLU [20]	5.13	70.6	63.5	89.4	85.4	74.4	74.6	31.9	93.0	25.2	41.5	82.0
	HR-EHNet (Our study)	11.72	73.6	74.2	92.1	85.1	75.5	**89.6**	55.9	95.5	50.8	48.3	**92.5**

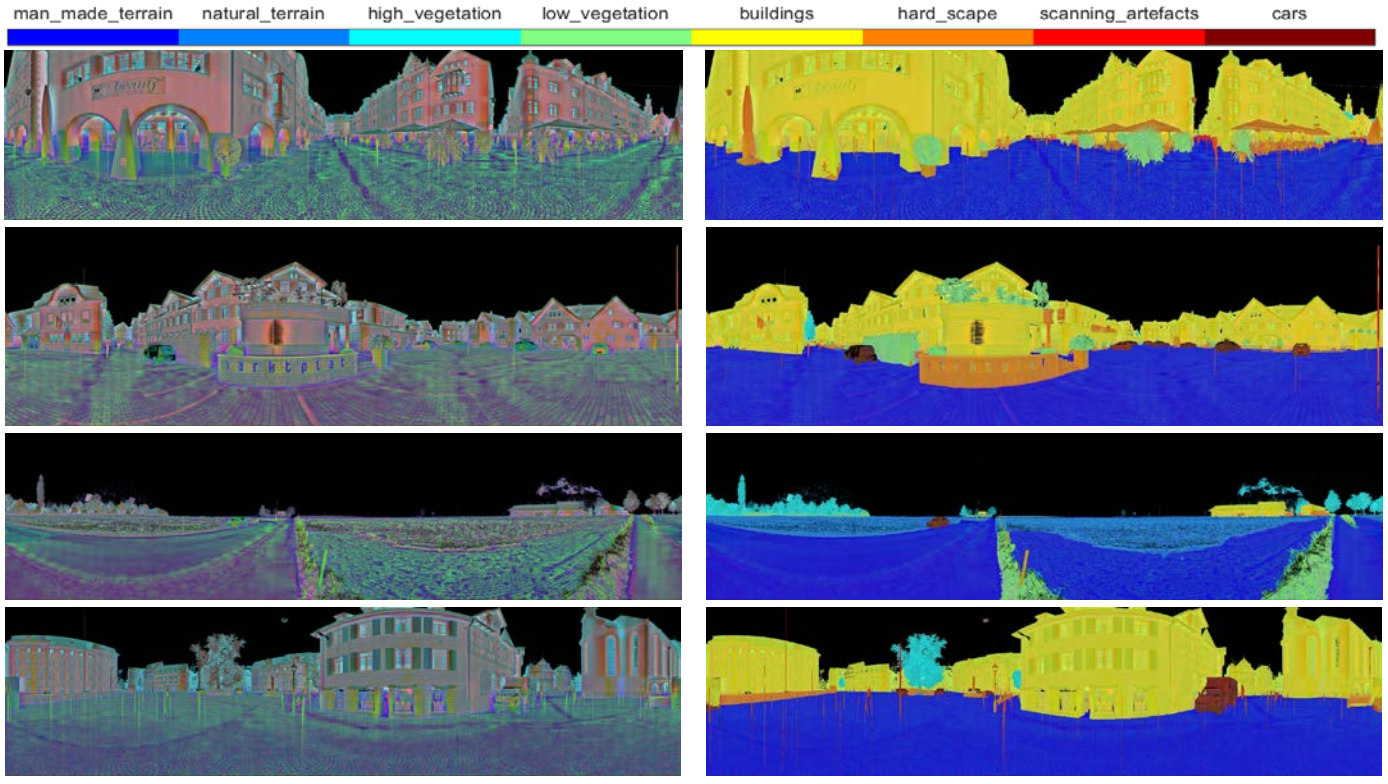
combination IZ_eD_e was selected as the final input to HR-EHNet, which happened to be a three-channel image. This means that for this particular combination, the first convolutional layer of the pre-trained HR-EHNet is unnecessarily replaced with a randomly initialized one. According to previous work [21], the operation of replacing the first convolutional layer could reduce the segmentation accuracy. Therefore, an experiment was conducted to determine whether to retain the pre-trained first convolutional layer in the final version of HR-EHNet. More specifically, the fourth experiment in Table I was repeated on the condition that the first pre-trained convolutional layer of HR-EHNet was retained. The corresponding segmentation results are summarized in Table II. As expected, the strategy of retaining the first pre-trained convolutional layer is beneficial for segmentation accuracy in mIoU and, therefore, adopted in the final version of HR-EHNet. All the prerequisites for performing the final training of HR-EHNet have now been determined. Therefore, the pre-trained HR-EHNet was fine-tuned with the complete training set (i.e., 15 images with IZ_eD_e feature channels and a size of 3600*7200 pixels) for 75,000 iterations according to the training protocols described in Section II.F.

The performance of HR-EHNet was evaluated on the Semantic3D (reduced-8) test dataset, which contains four point clouds. The four pseudo color images of IZ_eD_e and the corresponding segmentation results are illustrated in Fig.9. Through visual inspection, it is observed that the majority of the objects are correctly segmented and that most of the mislabels are concentrated at the edges where different objects intersect. These two-dimensional segmentation results were projected onto each data point in the point clouds to produce the segmented point clouds, which were uploaded to the online evaluation system of Semantic3D. The evaluation results have been made publicly available in the Semantic3D website under the name HR-EHNet (IZ_eD_e). The quantitative results of HR-EHNet and the recently published methods on Semantic3D (reduced-8) are summarized in Table III. Without RGB channels, HR-EHNet significantly outperforms the best outcomes of the previous image-based methods by 2.7% (OA) and 10.7% (mIoU), and meanwhile performed better than most of the point-based methods. It is also noted that HR-EHNet achieved the best segmentation accuracy with respect to high vegetation and cars among all the published methods.

The time spent on each step of HR-EHNet is recorded in Table IV. The data used in this test is the Semantic3D (reduced-8) test dataset, where the four-point clouds contain a total of 78.7 million data points. The inference was conducted with an AMD 3700X @3.6GHz CPU and an NVIDIA RTX2080Ti GPU. The total processing time was 11.72s, which was much faster than the other methods in Table III except XJTLU (Cai et al., 2021a). As shown in Table IV, HR-EHNet is slower than XJTLU because of the additional image enhancement step used.

TABLE IV
THE TIMES TAKEN BY EACH STEP OF HR-EHNET TO PROCESS THE SEMANTIC3D (REDUCED-8) TEST DATASET

	Time (s)	% of total time
Point cloud-image projection	0.17	1.5%
Enhancement	6.89	58.8%
Inference with neural network	4.55	38.8%
Image-point Cloud projection	0.11	1.0%
Total time	11.72	-



(a) (b)

Fig. 9. (a). The pseudo color images of IZ_eD_e feature channels for the point clouds in Semantic3D (reduced-8) test set, (b). The corresponding segmentation results (The legend is only for the visualization of the segmentation results in (b)).

IV. DISCUSSION

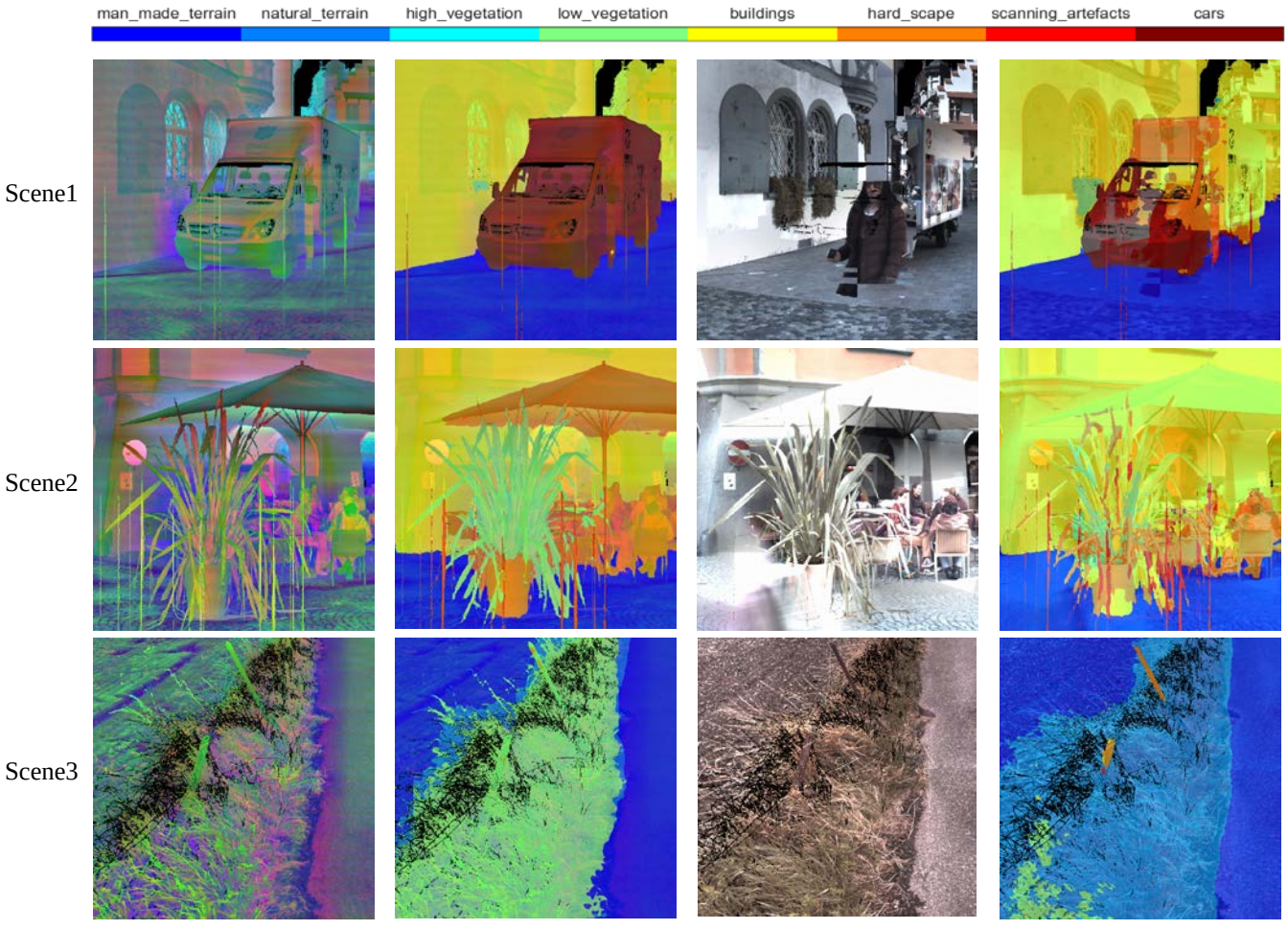
The core idea of HR-EHNet is to provide CNNs with distinguishable local geometric characteristics by enhancements of images derived from point cloud data. In this research, local image enhancement was implemented by a hand-crafted algorithm. Although the image enhancement method proposed was experimentally demonstrated to be effective and insensitive to the local patch size, it consumed more than half of the processing time as shown in Table IV. Considering that image enhancement is a relatively simple task in comparison to image segmentation, it is worth investigating how to reduce its processing time in the future. For example, one potential solution is to use the current image enhancement results as the target images to train a relatively simple neural network.

In this research, not all possible channel combinations were tested and as such there is no guarantee that IZ_eD_e is the best among all possible channel combinations. This is because the computational effort required would be enormous and the focus of this research was not on screening the optimal channel combinations. As such, developing an efficient way to identify optimal channel combinations is highly desirable in future research. Nevertheless, the results in this research showed that the channel combination IZ_eD_e represents a promising choice.

(a) (b)

Fig. 10. Incorrect RGB information in TLS point cloud data: (a). The RGB image contains the cyclist that were not scanned by TLS, (b). The enhanced Z image for the same scene.

The experimental results in Section III.C show that adding additional information (e.g., RGB or RGBD) to IZ_eD_e had a negative impact on the overall results (i.e., mIoU and OA). The primary reason for this phenomenon is the low reliability of the RGB images as mentioned Section I. For example, the RGB and the Z_e images of the same scene were shown in Fig.10.a and Fig.10.b, respectively. The RGB image shows a cyclist that does not exist in the Z_e image because the acquisition was not done simultaneously. The experimental results in Table I indicate that such false RGB information is an obstacle for neural networks to learn correct features. More evidences are shown in Fig.11, where Fig.11.b shows the classes predicted using the pseudo color images of IZ_eD_e in Fig.11.a, and



(a) (b) (c) (d)

Fig. 11. Comparisons between IZ_eD_e and IRGB$Z_eD_e$on segmentation results for three scenes: (a). The pseudo color images of IZ_eD_e, (b). The segmentation results using the corresponding IZ_eD_e images, (c) The RGB images, (d). The segmentation results using the feature combination of IRGBZ_eD_e.

Fig.11.d shows the classes predicted using IRGBZ_eD_e (i.e., IZ_eD_e in Fig.11.a and the RGB images in Fig.11.c). For Scene 1, it is seen that the vehicle in Fig.11.b was correctly segmented using only IZ_eD_e . However, when the erroneous RGB information was added, chaotic segmentation results (Fig.11.d) were obtained. A similar situation occurred for the vase in Scene 2. Nevertheless, Table I shows that for some particular classes (i.e., buildings, hardscape, man-made and natural), combining RGB information with IZ_eD_e improved their segmentation accuracies. This is because the accuracy of the RGB information is uncertain. When the RGB information of a particular class is accurate, it may be beneficial to include the RGB information for the segmentation of that particular class. For example, although the vehicle was segmented incorrectly in Scene 1 in Fig.11.d due to the erroneous RGB information, the vegetation at the windows was segmented correctly due to the correct and high contrast RGB information. However, when the RGB information of two adjacent classes does not show clear contrast, the inclusion of it may be problematic for the segmentation as demonstrated in the next paragraph. Future research may address the quality of RGB information from two perspectives. The most straightforward solution is to design a TLS strategy that simultaneously collects point cloud and imagery data to reduce as much false information as possible. The second possible solution is to design a neural network structure to enhance its ability to discriminate correct information from redundant/false information.

Compared with other segmentation methods, HR-EHNet was found to performance excellently in the recognition of plants and vehicles. This finding suggests its potential application to applications such as forest classification and autonomous driving. The possible reason for lower accuracies in the other methods for these two types of objects is that the use of the RGB information may cause confusion to neural networks if plants or vehicles have similar colors (i.e., spectra) to their surrounding objects. In contrast, HR-EHNet performs semantic segmentation mainly via the geometric features in the enhanced images, which are independent of color and can replace RGB images. For example, although the RGB colors of the vegetation in Scene 2 and Scene 3 seem to be accurate visually, the segmentation results obtained after adding the RGB information became worse due to its similarity to the surrounding objects. Furthermore, this characteristic is presumed to have significant advantages in terms of resistance to adversarial attack, which may offer better security for certain applications such as autonomous driving. There are many studies demonstrating that deep learning relying on RGB images is vulnerable to color perturbation attacks [88]–[90]. For example, shining light from a laser pointer on a stop sign may cause neural networks to fail to recognize the stop sign, which poses a significant safety challenge for autonomous driving [88]. Thus, it may be beneficial to extend the idea in this research to point cloud data that are used typically for a wider range of applications, including autonomous driving.

At present, there is only one TLS point cloud dataset (i.e., Semantic3D) publicly available for evaluating algorithms. Although Semantic3D is a large point cloud dataset in terms of the number of data points, it is small when it is processed as an image dataset (i.e., project each point cloud as a panoramic image), in comparison to image datasets such as the Cityscapes and Mapillary Vistas datasets [47], [83], [91]. Therefore, establishing a larger point cloud dataset would be extremely beneficial to the development of relevant research fields. It is also thought interesting to explore the feasibility of few-shot learning using the relatively small existing point cloud dataset.

Only 15 labeled panoramic images can be derived from the Semantic3D training set, which is not sufficient to support the decent training of HR-EHNet from scratch. Therefore, the Cityscapes dataset - that was taken in similar urban scenes and semantically labeled - was used for the network pre-training in this study. However, such semantically labeled images are often not publicly available. In contrast, unlabeled image datasets can readily be obtained for various application scenarios, through online resources and/or field acquisitions. Therefore, it is interesting to investigate how to effectively use techniques such as self-supervised learning [92]–[96] to pre-train networks using unlabeled images.

V. Conclusions

In this paper, a novel image enhancement method was proposed to characterize effectively the local geometric features in the panoramic images derived from TLS point cloud data. The enhanced images (i.e., enhanced Z-coordinates Z_e, and enhanced range D_e) alone and in various combinations of other popular feature channels (i.e., intensity I, RGB, range D) were used in a pre-trained CNN to assess the potential for semantic segmentation of the Semantic 3D datasets. It was found that compared with the commonly used channel combinations IRGB or IRGBD, our proposed combination IZ_eD_e produced more accurate semantic segmentation predictions. By fine-tuning the customized pre-trained HR-EHNet with the channel combination IZ_eD_e, an OA of 92.1% and a mIoU of 74.2% were obtained on the Semantic3D (reduced-8) test dataset, which substantially outperformed the other image-based methods. This suggests that effective utilization of local geometric features in images can increase the segmentation accuracy of image-based methods. This study also offers a better alternative channel combination to replace those involving the RGB channels, which may be extremely useful for cases where the RGB information is absent or inaccurate.

References

[1] R. Zhang, G. Li, M. Li, and L. Wang, "Fusion of images and point clouds for the semantic segmentation of large-scale 3D scenes based on deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 143, pp. 85–96, Sep. 2018.

[2] Z. Cao, D. Chen, J. Peethambaran, Z. Zhang, S. Xia, and L. Zhang, "Tunnel Reconstruction with Block Level Precision by Combining Data-Driven Segmentation and Model-Driven Assembly," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8853–8872, Oct. 2021, doi: 10.1109/TGRS.2020.3046624.

[3] Y. Cai and L. Fan, "An Efficient Approach to Automatic Construction of 3D Watertight Geometry of Buildings Using Point Clouds," *Remote Sens.*, vol. 13, no. 10, p. 1947, May 2021, doi: 10.3390/rs13101947.

[4] H. Huang, C. Zhang, and A. Hammad, "Effective Scanning Range Estimation for Using TLS in Construction Projects," *J. Constr. Eng.*

Manag., vol. 147, no. 9, p. 04021106, Jul. 2021, doi: 10.1061/(asce)co.1943-7862.0002127.

[5] L. Fan and Y. Cai, "An efficient filtering approach for removing outdoor point cloud data of manhattan-world buildings," *Remote Sens.*, vol. 13, no. 19, p. 3796, Sep. 2021, doi: 10.3390/rs13193796.

[6] Y. Cai, L. Fan, and C. Zhang, "An overview of constructing geometric models of buildings using point clouds," in *International Conference on Image, Video Processing, and Artificial Intelligence*, Nov. 2021, vol. 12076, p. 26, doi: 10.1117/12.2611685.

[7] S. Zheng *et al.*, "Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers," Dec. 2020, doi: 10.1109/cvpr46437.2021.00681.

[8] J. Liu *et al.*, "Comparison of terrestrial LiDAR and digital hemispherical photography for estimating leaf angle distribution in European broadleaf beech forests," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 76–89, Dec. 2019, doi: 10.1016/j.isprsjprs.2019.09.015.

[9] A. H. Safaie, H. Rastiveis, A. Shams, W. A. Sarasua, and J. Li, "Automated street tree inventory using mobile LiDAR point clouds based on Hough transform and active contours," *ISPRS J. Photogramm. Remote Sens.*, vol. 174, pp. 19–34, Apr. 2021, doi: 10.1016/j.isprsjprs.2021.01.026.

[10] L. Fan, W. Powrie, J. Smethurst, P. M. Atkinson, and H. Einstein, "The effect of short ground vegetation on terrestrial laser scans at a local scale," *ISPRS J. Photogramm. Remote Sens.*, vol. 95, pp. 42–52, Sep. 2014, doi: 10.1016/j.isprsjprs.2014.06.003.

[11] A. Montuori *et al.*, "Combined use of ground-based systems for Cultural Heritage conservation monitoring," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, Nov. 2014, pp. 4086–4089, doi: 10.1109/IGARSS.2014.6947384.

[12] J. Moyano, J. León, J. E. Nieto-Julián, and S. Bruno, "Semantic interpretation of architectural and archaeological geometries: Point cloud segmentation for HBIM parameterisation," *Automation in Construction*, vol. 130. Elsevier, p. 103856, Oct. 01, 2021, doi: 10.1016/j.autcon.2021.103856.

[13] B. Yang, Z. Dong, Y. Liu, F. Liang, and Y. Wang, "Computing multiple aggregation levels and contextual features for road facilities recognition using mobile laser scanning data," *ISPRS J. Photogramm. Remote Sens.*, vol. 126, pp. 180–194, Apr. 2017, doi: 10.1016/j.isprsjprs.2017.02.014.

[14] G. Vosselman, M. Coenen, and F. Rottensteiner, "Contextual segment-based classification of airborne laser scanner data," *ISPRS J. Photogramm. Remote Sens.*, vol. 128, pp. 354–371, Jun. 2017.

[15] M. Weinmann, B. Jutzi, S. Hinz, and C. Mallet, "Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers," *ISPRS J. Photogramm. Remote Sens.*, vol. 105, pp. 286–304, Jul. 2015, doi: 10.1016/j.isprsjprs.2015.01.016.

[16] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-Janua, pp. 77–85, doi: 10.1109/CVPR.2017.16.

[17] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems*, 2017, vol. 2017-Decem, pp. 5100–5109.

[18] L. Landrieu and M. Simonovsky, "Large-Scale Point Cloud Semantic Segmentation with Superpoint Graphs," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4558–4567, doi: 10.1109/CVPR.2018.00479.

[19] M. Jaritz, J. Gu, and H. Su, "Multi-view pointnet for 3D scene understanding," in *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019*, Oct. 2019, pp. 3995–4003, doi: 10.1109/ICCVW.2019.00494.

[20] Y. Cai, H. Huang, K. Wang, C. Zhang, L. Fan, and F. Guo, "Selecting Optimal Combination of Data Channels for Semantic Segmentation in City Information Modelling (CIM)," *Remote Sens.*, vol. 13, no. 7, p. 1367, Apr. 2021, doi: 10.3390/rs13071367.

[21] B. Pan, Z. Shi, X. Xu, T. Shi, N. Zhang, and X. Zhu, "CoinNet: Copy Initialization Network for Multispectral Imagery Semantic Segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 816–820, May 2019, doi: 10.1109/LGRS.2018.2880756.

[22] Q. Hu *et al.*, "Randla-Net: Efficient semantic segmentation of large-scale point clouds," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11105–11114, doi: 10.1109/CVPR42600.2020.01112.

[23] H. Thomas, F. Goulette, J. E. Deschaud, B. Marcotegui, and Y. Le Gall, "Semantic classification of 3d point clouds with multiscale spherical neighborhoods," in *Proceedings - 2018 International Conference on 3D Vision, 3DV 2018*, 2018, pp. 390–398, doi: 10.1109/3DV.2018.00052.

[24] M. Jiang, Y. Wu, T. Zhao, Z. Zhao, and C. Lu, "PointSIFT: A SIFT-like Network Module for 3D Point Cloud Semantic Segmentation," *arXiv*, 2018, [Online]. Available: http://arxiv.org/abs/1807.00652.

[25] F. Engelmann, T. Kontogianni, J. Schult, and B. Leibe, "Know what your neighbors do: 3D semantic segmentation of point clouds," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Sep. 2019, vol. 11131 LNCS, pp. 395–409, doi: 10.1007/978-3-030-11015-4_29.

[26] W. Zeng and T. Gevers, "3Dcontextnet: K-d tree guided hierarchical learning of point clouds using local and global contextual cues," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Sep. 2019, vol. 11131 LNCS, pp. 314–330, doi: 10.1007/978-3-030-11015-4_24.

[27] S. Xie, S. Liu, Z. Chen, and Z. Tu, "Attentional ShapeContextNet for Point Cloud Recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Dec. 2018, pp. 4606–4615, doi: 10.1109/CVPR.2018.00484.

[28] H. Zhao, L. Jiang, C. W. Fu, and J. Jia, "Pointweb: Enhancing local neighborhood features for point cloud processing," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2019, vol. 2019-June, pp. 5560–5568, doi: 10.1109/CVPR.2019.00571.

[29] L.-Z. Chen, X.-Y. Li, D.-P. Fan, K. Wang, S.-P. Lu, and M.-M. Cheng, "LSANet: Feature Learning on Point Sets by Local Spatial Aware Layer," *arXiv*, May 2019, Accessed: Apr. 21, 2021. [Online]. Available: http://arxiv.org/abs/1905.05442.

[30] Z. Zhang, B. S. Hua, and S. K. Yeung, "ShellNet: Efficient point cloud convolutional neural networks using concentric shells statistics," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, vol. 2019-Octob, pp. 1607–1616, doi: 10.1109/ICCV.2019.00169.

[31] J. Gong *et al.*, "Omni-supervised Point Cloud Segmentation via Gradual Receptive Field Component Reasoning," May 2021, doi: 10.1109/cvpr46437.2021.01150.

[32] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph Cnn for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, pp. 1–12, Oct. 2019, doi: 10.1145/3326362.

[33] L. Landrieu and M. Boussaha, "Point cloud oversegmentation with graph-structured deep metric learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2019, vol. 2019-June, pp. 7432–7441, doi: 10.1109/CVPR.2019.00762.

[34] L. Wang, Y. Huang, Y. Hou, S. Zhang, and J. Shan, "Graph attention convolution for point cloud semantic segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019, vol. 2019-June, pp. 10288–10297, doi: 10.1109/CVPR.2019.01054.

[35] L. Pan, C. M. Chew, and G. H. Lee, "PointAtrousGraph: Deep Hierarchical Encoder-Decoder with Point Atrous Convolution for Unorganized 3D Points," in *Proceedings - IEEE International Conference on Robotics and Automation*, Jul. 2020, pp. 1113–1120, doi: 10.1109/ICRA40945.2020.9197499.

[36] H. Lei, N. Akhtar, and A. Mian, "Spherical Convolutional Neural Network for 3D Point Clouds," *arXiv*, May 2018, Accessed: Apr. 21, 2021. [Online]. Available: http://arxiv.org/abs/1805.07872.

[37] J. Liu, B. Ni, C. Li, J. Yang, and Q. Tian, "Dynamic points agglomeration for hierarchical point sets learning," in *Proceedings of the IEEE International Conference on Computer Vision*, Oct. 2019, vol. 2019-Octob, pp. 7545–7554, doi: 10.1109/ICCV.2019.00764.

[38] S. Wang, S. Suo, W. C. Ma, A. Pokrovsky, and R. Urtasun, "Deep Parametric Continuous Convolutional Neural Networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Dec. 2018, pp. 2589–2597, doi: 10.1109/CVPR.2018.00274.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2016, vol. 2016-Decem, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[40] A. Boulch, "ConvPoint: Continuous convolutions for point cloud processing," *Comput. Graph.*, vol. 88, pp. 24–34, May 2020, doi: 10.1016/j.cag.2020.02.005.

[41] H. Thomas, C. R. Qi, J. E. Deschaud, B. Marcotegui, F. Goulette, and L. Guibas, "KPConv: Flexible and deformable convolution for point clouds," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, vol. 2019-Octob, pp. 6410–6419, doi: 10.1109/ICCV.2019.00651.

[42] F. Engelmann, T. Kontogianni, and B. Leibe, "Dilated Point Convolutions: On the Receptive Field Size of Point Convolutions on 3D Point Clouds," in *Proceedings - IEEE International Conference on Robotics and Automation*, May 2020, pp. 9463–9469, doi: 10.1109/ICRA40945.2020.9197503.

[43] J. Mao, X. Wang, and H. Li, "Interpolated convolutional networks for 3D point cloud understanding," in *Proceedings of the IEEE International Conference on Computer Vision*, Oct. 2019, vol. 2019-Octob, pp. 1578–1587, doi: 10.1109/ICCV.2019.00166.

[44] Q. Huang, W. Wang, and U. Neumann, "Recurrent Slice Networks for 3D Segmentation of Point Clouds," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Dec. 2018, pp. 2626–2635, doi: 10.1109/CVPR.2018.00278.

[45] F. Engelmann, T. Kontogianni, A. Hermans, and B. Leibe, "Exploring Spatial Context for 3D Semantic Segmentation of Point Clouds," in *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, Jul. 2017, vol. 2018-Janua, pp. 716–724, doi: 10.1109/ICCVW.2017.90.

[46] X. Ye, J. Li, H. Huang, L. Du, and X. Zhang, "3D recurrent neural networks with context fusion for point cloud semantic segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Sep. 2018, vol. 11211 LNCS, pp. 415–430, doi: 10.1007/978-3-030-01234-2_25.

[47] T. Hackel, N. Savinov, L. Ladicky, J. D. Wegner, K. Schindler, and M. Pollefeys, "SEMANTIC3D.NET: A NEW LARGE-SCALE POINT CLOUD CLASSIFICATION BENCHMARK," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, May 2017, vol. 4, no. 1W1, pp. 91–98, doi: 10.5194/isprs-annals-IV-1-W1-91-2017.

[48] L. Tchapmi, C. Choy, I. Armeni, J. Gwak, and S. Savarese, "SEGCloud: Semantic segmentation of 3D point clouds," in *Proceedings - 2017 International Conference on 3D Vision, 3DV 2017*, May 2018, pp. 537–547, doi: 10.1109/3DV.2017.00067.

[49] C. Choy, J. Gwak, and S. Savarese, "4D spatio-temporal convnets: Minkowski convolutional neural networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2019, vol. 2019-June, pp. 3070–3079, doi: 10.1109/CVPR.2019.00319.

[50] H. Y. Meng, L. Gao, Y. K. Lai, and Di. Manocha, "VV-net: Voxel VAE net with group convolutions for point cloud segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, Oct. 2019, vol. 2019-Octob, pp. 8499–8507, doi: 10.1109/ICCV.2019.00859.

[51] B. Graham, M. Engelcke, and L. Van Der Maaten, "3D Semantic Segmentation with Submanifold Sparse Convolutional Networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Dec. 2018, pp. 9224–9232, doi: 10.1109/CVPR.2018.00961.

[52] F. Wang, Y. Zhuang, H. Gu, and H. Hu, "OctreeNet: A Novel Sparse 3-D Convolutional Neural Network for Real-Time 3-D Outdoor Scene Analysis," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 2, pp. 735–747, Apr. 2020, doi: 10.1109/TASE.2019.2942068.

[53] I. Armeni *et al.*, "3D semantic parsing of large-scale indoor spaces," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, pp. 1534–1543, doi: 10.1109/CVPR.2016.170.

[54] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361, doi: 10.1109/CVPR.2012.6248074.

[55] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Rob. Res.*, vol. 32, no. 11, pp. 1231–1237, Aug. 2013, doi: 10.1177/0278364913491297.

[56] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7576 LNCS, no. PART 5, pp. 746–760, doi: 10.1007/978-3-642-33715-4_54.

[57] F. J. Lawin, M. Danelljan, P. Tosteberg, G. Bhat, F. S. Khan, and M. Felsberg, "Deep projective 3D semantic segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol. 10424 LNCS, pp. 95–107, doi: 10.1007/978-3-319-64689-3_8.

[58] A. Boulch, J. Guerry, B. Le Saux, and N. Audebert, "SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks," *Comput. Graph.*, vol. 71, pp. 189–198, 2018, doi: 10.1016/j.cag.2017.11.010.

[59] M. Tatarchenko, J. Park, V. Koltun, and Q. Y. Zhou, "Tangent Convolutions for Dense Prediction in 3D," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Dec. 2018, pp. 3887–3896, doi: 10.1109/CVPR.2018.00409.

[60] B. Wu, A. Wan, X. Yue, and K. Keutzer, "SqueezeSeg: Convolutional Neural Nets with Recurrent CRF for Real-Time Road-Object Segmentation from 3D LiDAR Point Cloud," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2018, pp. 1887–1893, doi: 10.1109/ICRA.2018.8462926.

[61] B. Wu, X. Zhou, S. Zhao, X. Yue, and K. Keutzer, "SqueezeSegV2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a LiDAR point cloud," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2019, vol. 2019-May, pp. 4376–4382, doi: 10.1109/ICRA.2019.8793495.

[62] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss, "RangeNet ++: Fast and Accurate LiDAR Semantic Segmentation," in *IEEE International Conference on Intelligent Robots and Systems*, 2019, pp. 4213–4220, doi: 10.1109/IROS40897.2019.8967762.

[63] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.

[64] J. Wang *et al.*, "Deep High-Resolution Representation Learning for Visual Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Apr. 2021, doi: 10.1109/TPAMI.2020.2983686.

[65] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Sep. 2018, vol. 11211 LNCS, pp. 833–851, doi: 10.1007/978-3-030-01234-2_49.

[66] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," Jun. 2017, Accessed: Aug. 23, 2021. [Online]. Available: https://arxiv.org/abs/1706.05587v3.

[67] R. C. Gonzalez, R. E. Woods, and B. R. Masters, "Digital Image Processing, Third Edition," *J. Biomed. Opt.*, vol. 14, no. 2, p. 029901, 2009, doi: 10.1117/1.3115362.

[68] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.

[69] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Nov. 2017, vol. 2017-Janua, pp. 1800–1807, doi: 10.1109/CVPR.2017.195.

[70] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 2017, pp. 4278–4284.

[71] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.

42, no. 8, pp. 2011–2023, Aug. 2020, doi: 10.1109/TPAMI.2019.2913372.

[72] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10425–10433, doi: 10.1109/CVPR42600.2020.01044.

[73] S. H. Gao, M. M. Cheng, K. Zhao, X. Y. Zhang, M. H. Yang, and P. Torr, "Res2Net: A New Multi-Scale Backbone Architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021, doi: 10.1109/TPAMI.2019.2938758.

[74] S. Borse, Y. Wang, Y. Zhang, and F. Porikli, *InverseForm: A Loss Function for Structured Boundary-Aware Segmentation*. 2021, pp. 5901–5911.

[75] C. Yu *et al.*, "Lite-HRNet: A Lightweight High-Resolution Network," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Nov. 2021, pp. 10435–10445, doi: 10.1109/cvpr46437.2021.01030.

[76] Y. Yuan, X. Chen, and J. Wang, "Object-Contextual Representations for Semantic Segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Aug. 2020, vol. 12351 LNCS, pp. 173–190, doi: 10.1007/978-3-030-58539-6_11.

[77] Z. Xu, W. Zhang, T. Zhang, and J. Li, "Hrcnet: High-resolution context extraction network for semantic segmentation of remote sensing images," *Remote Sens.*, vol. 13, no. 1, pp. 1–23, Dec. 2021, doi: 10.3390/rs13010071.

[78] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, doi: 10.1109/CVPR.2016.90.

[79] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: Gated shape CNNs for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, Oct. 2019, vol. 2019-Octob, pp. 5228–5237, doi: 10.1109/ICCV.2019.00533.

[80] S. Niu, Y. Liu, J. Wang, and H. Song, "A Decade Survey of Transfer Learning (2010–2020)," *IEEE Trans. Artif. Intell.*, vol. 1, no. 2, pp. 151–166, Jun. 2021, doi: 10.1109/tai.2021.3054609.

[81] H. Liang, W. Fu, and F. Yi, "A Survey of Recent Advances in Transfer Learning," in *International Conference on Communication Technology Proceedings, ICCT*, Oct. 2019, pp. 1516–1523, doi: 10.1109/ICCT46805.2019.8947072.

[82] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 26, no. 5, pp. 1019–1034, May 2015, doi: 10.1109/TNNLS.2014.2330900.

[83] M. Cordts *et al.*, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, pp. 3213–3223, doi: 10.1109/CVPR.2016.350.

[84] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Nov. 2017, vol. 2017-Janua, pp. 6230–6239, doi: 10.1109/CVPR.2017.660.

[85] H. Zhao *et al.*, "PSANet: Point-wise spatial attention network for scene parsing," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Sep. 2018, vol. 11213 LNCS, pp. 270–286, doi: 10.1007/978-3-030-01240-3_17.

[86] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.

[87] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep Learning for 3D Point Clouds: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4338–4364, Jun. 2020, doi: 10.1109/tpami.2020.3005434.

[88] R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. Yang, "Adversarial camouflage: Hiding physical-world attacks with natural styles," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020, pp. 997–1005, doi: 10.1109/CVPR42600.2020.00108.

[89] K. Eykholt *et al.*, "Robust Physical-World Attacks on Deep Learning Visual Classification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Dec. 2018, pp. 1625–1634, doi: 10.1109/CVPR.2018.00175.

[90] A. Shahin Shamsabadi, R. Sanchez-Matilla, and A. Cavallaro, "ColorFool: Semantic Adversarial Colorization," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1148–1157, doi: 10.1109/CVPR42600.2020.00123.

[91] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, "The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, vol. 2017-Octob, pp. 5000–5009, doi: 10.1109/ICCV.2017.534.

[92] X. Zhan, X. Pan, Z. Liu, D. Lin, and C. C. Loy, "Self-supervised learning via conditional motion propagation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2019, vol. 2019-June, pp. 1881–1889, doi: 10.1109/CVPR.2019.00198.

[93] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context Encoders: Feature Learning by Inpainting," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Dec. 2016, vol. 2016-Decem, pp. 2536–2544, doi: 10.1109/CVPR.2016.278.

[94] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised Feature Learning via Non-parametric Instance Discrimination," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Dec. 2018, pp. 3733–3742, doi: 10.1109/CVPR.2018.00393.

[95] R. Zhang, P. Isola, and A. A. Efros, "Split-brain autoencoders: Unsupervised learning by cross-channel prediction," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Nov. 2017, vol. 2017-Janua, pp. 645–654, doi: 10.1109/CVPR.2017.76.

[96] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9726–9735, doi: 10.1109/CVPR42600.2020.00975.