2023 International Symposium on Nonlinear Theory and Its Applications
NOLTA2023, September 26-29, 2023, Catania and Online

Scale-Equivariant Convolution for Projection-based Point Cloud Segmentation

Hidetaka Marumo†, Takashi Matsubara†

†Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka, 560-8531 Japan
Email: marumo@hopf.sys.es.osaka-u.ac.jp, matsubara@sys.es.osaka-u.ac.jp

Abstract— With the progress in Artificial intelligence (AI) computer vision, semantic segmentation of Light Detection and Ranging (LiDAR) point clouds using deep neural networks (DNNs) has attracted attention in autonomous driving. Considering recognition accuracy and computational complexity, a promising approach is to project a point cloud into a 2-dimensional (2D) range image and process it with 2D convolutional neural networks (CNNs). Since distant objects appear smaller than nearby objects in an image, it is crucial to incorporate scale-equivariance into the CNN to improve parameter efficiency and recognition accuracy, but no method has focused on it. We proposed a new scale-equivariant convolution method, focusing on the relationship between object distance and scale ratio in images as well as the theoretical properties of partial differential operators. Evaluation experiments on the LiDAR point cloud dataset demonstrate the effectiveness of our method.

1. Introduction

An accurate and robust understanding of the environment is essential for autonomous driving. Therefore, various sensors have been utilized, in addition to cameras. Light Detection and Ranging (LiDAR) scanner is one of the promising sensors to be utilized further in the future because it can acquire 3-dimensional (3D) information accurately and be easily integrated into later decision-making and operations. On the other hand, research on recognition techniques such as semantic segmentation of LiDAR point clouds using deep neural networks (DNNs) has attracted much attention with the progress in Artificial intelligence (AI) computer vision. Semantic segmentation of point clouds is assigning an object class label to each point in a point clouds, which is used to recognize objects and locate drivable areas.

Recent semantic segmentation of point clouds can be roughly divided into two categories. The first is for small-scale, high-density point clouds used in object analysis and indoor scene understanding. In these cases, high shape extraction performance is required, while processing speed is often not as critical. Therefore, point-wise processing methods based on PointNet [2] are the mainstream in this case. The other is for large-scale and sparse point clouds, such as outdoor scene understanding in autonomous driving. In this case, the effectiveness of PointNet-based methods is limited due to the need for real-time processing and the difficulty of acquiring contextual information. Therefore, projection-based methods such as RangeNet++ [5] and point-voxel-based methods such as SPVConv [3] are the mainstream. Projection-based methods do not directly process 3D point clouds but rather transform them into a dense 2-dimensional (2D) representation, such as range images, projecting it onto a sphere centered on the sensor. This process not only increases processing efficiency, but also allows the use of standard CNNs for RGB images. The results of segmentation in 2D space are then re-projected onto 3D point cloud space to achieve point cloud segmentation. Therefore, the projection-based method is effective regarding both computational efficiency and accuracy. However, 2D projection causes differences in scale within the image; for example, distant objects are represented as smaller than nearby objects. As a result, even if the objects are the same, they are learned separately, reducing the network parameter efficiency. Therefore, it is crucial to incorporate scale-equivariance into CNN such that the same results are obtained regardless of local-scale differences. However, to the best of our knowledge, no projection-based method has focused on scale-equivariance.

This study proposes a new convolution method, range-equivariant convolution (REconv), which effectively utilizes the range information. The experiments on the Semantic KITTI dataset [1] showed that by replacing the first three convolutional layers of RangeNet21 [5] with REconv resulted in a 1.1% improvement of mIoU. In addition, we defined a measure called equivariance error, and evaluated scale-equivariance by comparing the similarity of the feature maps for inputs of different scales. The equivariance error in the feature maps for the layers using REconv was smaller than the one using standard convolution. This result indicates that REconv has scale-equivariance. The above two results demonstrate that REconv is an effective method for semantic segmentation of LiDAR point clouds.

2. Related Works

2.1. LiDAR Point Clouds Segmentation

2.1.1. Point-Voxel-based Methods

Point-voxel-based methods such as SPVConv [3, 4] combine point-wise processing with 3D convolution by dividing space into voxels. These methods can achieve high

ORCID iDs First Author: 0009-0002-2776-4193, Second Author: 0000-0003-0642-4800 recognition accuracy but tend to be computationally expensive. In addition, due to the imbalance in the density of the point cloud, many voxels do not contain any points, which reduces the computational efficiency. Cylinder3D [4] alleviated some of this problem by rethinking the voxel shape but did not solve it completely.

2.1.2. Projection-based Methods

Among the methods for projecting LiDAR point clouds into 2D range images, RangeNet++ [5] was proposed in the early stages and has since become the baseline for many studies, including this study. In RangeNet++, the LiDAR point cloud is first projected onto a sphere, as follows:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \frac{1}{2}[1 - \arctan(y/x)\pi^{-1}] \;\; w \\ [1 - (\arcsin(zr^{-1} + \mathrm{f_{up}})\mathrm{f}^{-1}] \;\; h \end{pmatrix}, \tag{1}$$

where $(h,\ w)$ are the height and width of the desired range image representation, $\mathrm{f} = \mathrm{f_{up}} + \mathrm{f_{down}}$ is the vertical field of view of the sensor, and $r = ||\mathbf{p}_i||_2$ is the range of each point. This equation provides the correspondence between points and pixels, and the five channels of the range value r, (x, y, z) coordinate values, and reflection intensity i were used as inputs to the CNN. Semantic labels in the image segmented by the 2D CNN were projected onto the point cloud space using the correspondence between points and pixels. In addition, the k-nearest neighbor method was applied in a 3D space to refine "shadow-like artifacts" caused by the "blurring" of object boundaries. The k-nearest neighbor method used here applies several approximations, so the increase in computational complexity is kept to a minimum. Subsequent research has proposed various improvement based on this method. SqueezesegV3 [6] introduced a new convolution method that focuses on the differences in modalities, such as range r and (x, y, z) coordinate values. MiNet [7] used standard and depth-wise convolution for each resolution to reduce the computational complexity while maintaining high accuracy. SalsaNext [8] achieves state-of-the-art performance among projection-based methods by expanding the receptive field to facilitate contextual information extraction. However, to the best of our knowledge, no method focuses on the differences in scale between the objects in an image.

2.2. Scale-Equivariant Convolution

In many image-based tasks, differences in object position, orientation, and distance cause transformations, such as shifting, rotating, or scaling the input image. Because these transformations significantly affect the discriminative power of the model, there is a discussion on equivariance, which is the property of being able to extract the same features from the input with these transformations. Standard convolution is shift-equivariant but not rotation or scale-equivariant. In autonomous driving, rotation-equivariance is not so important since the situation where a rotated object appears is unlikely. However, the scale-equivariance is vital for 2D projected LiDAR images because objects at a distance are represented as smaller than the nearby objects.

In the research on scale-equivariance in the framework of CNN for RGB images, multiple methods [9, 10] have been proposed. SESN [9] introduced a manipulatable filter, allowing it to handle arbitrary scales. DISCO [10] uses a kernel derived by solving a constraint equation that must be satisfied to be scale-equivariant and minimizes the equivariance error. The above scale-equivariant convolution methods assumes that the scale ratio is unknown. However, for LiDAR images, the scale ratio can be calculated using the range values. Partial differential operators (PDOs) have straightforward theoretical scaling properties. Therefore, we develop a new scale-equivariant convolution method by defining a convolution filter as a linear combination of multiple PDOs, as in PDO-eConvs [11], which allow the filter for manipulation by applying weights according to the range and differential order of each PDO.

3. Method

3.1. Preliminaries

3.1.1. Scale Transformation

The scale transformation of the scale ratio s to the feature function $f : \mathbb{R} \to \mathbb{R}$ can be expressed as follows:

$$L_s[f](\boldsymbol{\xi}) = f(s^{-1}\boldsymbol{\xi}), \qquad \forall s > 0. \tag{2}$$

3.1.2. Approximation of PDOs

When the partial derivative in the x direction is denoted by D_x, the result of applying D_x to the feature $f(\boldsymbol{\xi})$ can be discretized as follows:

$$D_x[f](\boldsymbol{\xi}) \approx \frac{f(\boldsymbol{\xi} + \Delta x) - f(\boldsymbol{\xi} - \Delta x)}{2\Delta x}. \tag{3}$$

Consequently, D_x can be represented by a following filter:

$$D_x^d = \frac{1}{\Delta x}\begin{bmatrix} 0 & 0 & 0 \\ -1/2 & 0 & 1/2 \\ 0 & 0 & 0 \end{bmatrix}. \tag{4}$$

Similarly, $D_0, D_y, D_{xx}, D_{xy}, D_{yy}, D_{xxy}, D_{xyy}$, and D_{xxyy} can be approximated by 3×3 filters. In the following, we denote each discretized PDOs by D_*^d for $* \in \{0, x, y, xx, xy, yy, xxy, xyy, xxyy\} = \mathcal{D}$.

3.1.3. Partial Differentiation of Scaled Features

Applying the discretized D_x^d to the scaled features yields the following equation:

$$\begin{aligned} D_x^d[L_s[f]](s\boldsymbol{\xi}) &= \frac{L_s[f](s\boldsymbol{\xi} + s\Delta x) - L_s[f](s\boldsymbol{\xi} - s\Delta x)}{2s\Delta x} \\ &= \frac{1}{s} \cdot \frac{f(\boldsymbol{\xi} + \Delta x) - f(\boldsymbol{\xi} - \Delta x)}{2\Delta x} \\ &= \frac{1}{s} \cdot D_x^d[f](\boldsymbol{\xi}). \quad (\because \text{ Eq. (3)}) \end{aligned} \tag{5}$$

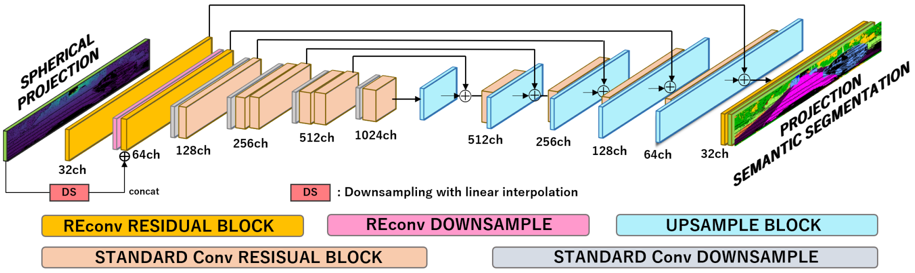

Figure 1: Network architecture of REconvNet21

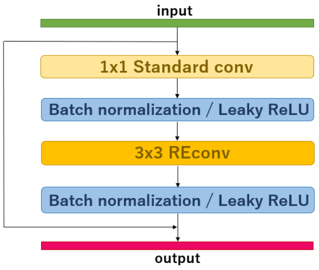
input
1x1 Standard conv
Batch normalization / Leaky ReLU
3x3 REconv
Batch normalization / Leaky ReLU
output

Figure 2: REconv residual block

Similarly, the results of applying each of the discretized PDOs to the scaled features can be expressed as follows:

$$D^d_*[L_s[f]](s\boldsymbol{\xi}) = \mathrm{s}_* D^d_*[f](\boldsymbol{\xi}), \tag{6}$$

$$\mathrm{s}_* = \frac{1}{s^{i_*+j_*}}, \tag{7}$$

where i_* and j_* representing the differential orders of D_* in the x and y directions respectively.

3.2. Range-Equivariant Convolution (REconv)

From the Eq. (4), each filter is multiplied by a coefficient corresponding to the actual length per pixel and derivative factor. The scale ratio s of a feature in a LiDAR image is inversely proportional to the range r of the object corresponding to that pixel. Based on these properties, the variable r_* is defined as follows such that $\mathrm{r}_*\mathrm{s}_* = 1$:

$$\mathrm{r}_* = \frac{{k_x}^{i_*}{k_y}^{j_*}}{r^{i_*+j_*}}. \tag{8}$$

Note that k_x and k_y are the coefficients that depend on the resolution in the x and y directions, respectively. i_* and j_* are the differential orders in the x and y directions of D_*, respectively. We define a new weighted partial differential operator using r_* as follows:

$$D^r_* = \mathrm{r}_* D^d_*. \tag{9}$$

Furthermore, a convolution filter is defined as a linear combination of nine D^r_* filters, parameterized by the learnable coefficient parameter $\boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots, \beta_9\}$ as follows:

$$D^d_r = \sum_{*\in\mathcal{D}} \beta_i D^r_*. \tag{10}$$

The convolution with the filter above applied to the scaled features is represented as follows:

$$\begin{aligned} D^d_r[L_s[f]](s\boldsymbol{\xi}) &= \sum_{*\in\mathcal{D}} \beta_i D^r_*[L_s[f]](s\boldsymbol{\xi}) \\ &= \sum_{*\in\mathcal{D}} \beta_i \mathrm{r}_* \mathrm{s}_* D^d_*[f](\boldsymbol{\xi}) \\ &= \sum_{*\in\mathcal{D}} \beta_i D^d_*[f](\boldsymbol{\xi}) \\ &= D^d_r[f](\boldsymbol{\xi}). \end{aligned} \tag{11}$$

From the above, the convolution using D^r_* is a scale-equivariant for LiDAR range images.

3.3. Usage of REconv

When incorporating REconv into a CNN, you should consider the treatment of absolute values of range. It should be added range information by post-processively concatenating range images to the feature map processed by REconv. When extracting features using PDOs, the absolute range values are likely to be lost because of the focus on the relative values between pixels. However, the absolute range value is significant because it works as a prior distribution for each class; for example, “vegetation” and “buildings,” which correspond to the background tend to appear at the distance. Therefore, we can compensate for the weaknesses of REconv while benefiting from its advantages by adding range information.

4. Experiments

4.1. Dataset

We conducted evaluation experiments using the Semantic KITTI dataset [1], which is a LiDAR point cloud dataset, as in RangeNet++ [5]. The sequences {0–7} and {9, 10} are the training set and {8} is the validation set in this dataset. The model was trained on the training set and evaluated on the validation set in this study.

4.2. Segmentation Performance

4.2.1. Metrics

We used mIoU (mean intersection-over-union), which is a commonly used metric for evaluating segmentation performance.

$$\mathrm{mIoU} = \frac{1}{C}\sum_{c=1}^{C} \frac{\mathbf{TP}_c}{\mathbf{TP}_c + \mathbf{FP}_c + \mathbf{FN}_c}. \tag{12}$$

where $\mathbf{TP}_c$, $\mathbf{FP}_c$ and $\mathbf{FN}_c$ denote the numbers of true positives, false positives, and false negatives for class c; where C denotes the number of classes;

4.2.2. Results and Discussions

The network structure is shown in Fig. 1 and Fig. 2 to rigorously evaluate the effectiveness of REconv. RangeNet21 [5] was used as baseline, and the first three convolution layers of the encoder were replaced with REconv. Hereafter, this network is referred to as REconvNet21. Tab. 1 compares the segmentation performances

Table 1: Segmentation performance on Semantic KITTI dataset

method	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign	mean IoU
RangeNet21++ [5]	87.3	**19.0**	30.1	24.2	23.4	24.3	47.6	0.0	93.0	39.8	78.8	**0.3**	80.3	40.8	**81.3**	**45.9**	69.5	40.2	**30.8**	45.1
REconvNet21(ours)	**89.7**	13.2	**38.5**	**29.6**	**25.2**	**26.4**	**48.0**	0.0	**93.3**	**42.9**	**79.9**	0.2	**81.5**	**43.0**	80.7	45.1	**69.7**	**40.5**	30.1	**46.2**

Table 2: Equivariance Error on Semantic KITTI dataset

methods	metrics		1st layer	3rd layer
RangeNet21[5]	MSE		0.0329	0.0151
	variance	64 × 2048	0.0486	0.0317
		32 × 1024	0.0662	0.0506
	Error		0.5664	0.3778
REconvNet21(ours)	MSE		0.0032	0.0039
	variance	64 × 2048	0.0428	0.0329
		32 × 1024	0.0413	0.0352
	Error		**0.0823**	**0.1211**

on the Semantic KITTI dataset. This table shows that REconvNet21 outperformed RangeNet21++ by 1.1% in mIoU. Furthermore, more considerable IoU gains were observed mainly for large object classes, such as ”car” and ”building”. This is because large objects tend to retain their shapes in images even when the scale is reduced, and the effect of incorporating scale-equivariance is significant and. This result indicates that it is practical to incorporate scale-equivariance into the CNN for processing LiDAR images.

4.3. Equivariance

4.3.1. Metrics

We evaluated scale-equivariance using the similarity of feature maps for inputs of different scales. Specificaly, considering the mean squared error (MSE) between the feature maps and the variance of the feature maps, the equivariance error is defined as

$$\text{Error} = \frac{\text{MSE}}{\text{average of the two variances}}. \tag{13}$$

The smaller this value is, the stronger equivariance.

4.3.2. Results and Discussion

Tab. 2 compares the equivariance errors on the Semantic KITTI dataset. This result shows that REconvNet21 has a much smaller equivariance error, indicating that REconv certainly has scale-equivariance.

5. Conclusion

This study proposed a new range-equivariant convolution for the semantic segmentation of LiDAR images. The experimental results show that replacing a part of the standard convolution layer of RangeNet21 with REconv in an appropriate way improved mIoU on outdoor point cloud datasets. Furthermore, the evaluation of the equivariance error showed that REconv exhibited scale-equivariance. In conclusion, incorporating the scale-equivariance into the CNN using the proposed method is effective for the segmentation of LiDAR point clouds.

Acknowledgments

This work was supported by JST PRESTO (JPMJPR21C7), JST CREST (JPMJCR1914) and JSPS KAKENHI (19K20344, 19H04172), and JST-Mirai Program (JPMJMI20B8).

References

[1] Behley, Jens, et al. ”Semantickitti: A dataset for semantic scene understanding of lidar sequences.” in Proc. of the IEEE/CVF ICCV (2019).

[2] Qi, Charles R., et al. ”Pointnet: Deep learning on point sets for 3d classification and segmentation.” in Proc. of the IEEE/CVF CVPR (2017).

[3] Tang, Haotian, et al. ”Searching efficient 3d architectures with sparse point-voxel convolution.” in Proc. of IEEE/CVF ECCV (2020).

[4] Zhou, Hui et al. “Cylinder3D: An Effective 3D Framework for Driving-scene LiDAR Semantic Segmentation. ” ArXiv abs/2008.01550 (2020): n. pag.

[5] Milioto, Andres et al. “ RangeNet ++: Fast and Accurate LiDAR Semantic Segmentation. ” Proc. of the IEEE/RSJ IROS (2019).

[6] Xu, Chenfeng et al. “ SqueezeSegV3: Spatially-Adaptive Convolution for Efficient Point-Cloud Segmentation. ” in Proc. of the IEEE/CVF ECCV (2020).

[7] Li, Shijie et al. “ Multi-Scale Interaction for Real-Time LiDAR Data Segmentation on an Embedded Platform. ” IEEE Robotics and Automation Letters 7 (2020): 738-745.

[8] Cortinhal, Tiago et al. “ SalsaNext: Fast, Uncertainty-Aware Semantic Segmentation of LiDAR Point Clouds. ” in Proc. of ISVC (2020).

[9] Sosnovik, Ivan et al. “Scale-Equivariant Steerable Networks. ” ArXiv abs/1910.11093 (2019): n. pag.

[10] Sosnovik, Ivan et al. “ How to Transform Kernels for Scale-Convolutions. ” in Proc. of IEEE/CVF ICCV Workshops (2021): 1092-1097.

[11] Shen, Zhengyang et al. “ PDO-eConvs: Partial Differential Operator Based Equivariant Convolutions. ”in Proc. of ICML (2020).