

Regularization Strategy for Point Cloud via Rigidly Mixed Sample

Dogyoon Lee¹ Jaeha Lee¹ Junhyeop Lee¹ Hyeongmin Lee¹ Minhyeok Lee¹
Sungmin Woo¹ Sangyoun Lee^{1*}

¹Yonsei University

{nemotio, jaeha0725, jun.lee, minimonia, hydragon516, smw3250, syleee}@yonsei.ac.kr

Abstract

Data augmentation is an effective regularization strategy to alleviate the overfitting, which is an inherent drawback of the deep neural networks. However, data augmentation is rarely considered for point cloud processing despite many studies proposing various augmentation methods for image data. Actually, regularization is essential for point clouds since lack of generality is more likely to occur in point cloud due to small datasets. This paper proposes a Rigid Subset Mix (RSMix)¹, a novel data augmentation method for point clouds that generates a virtual mixed sample by replacing part of the sample with shape-preserved subsets from another sample. RSMix preserves structural information of the point cloud sample by extracting subsets from each sample without deformation using a neighboring function. The neighboring function was carefully designed considering unique properties of point cloud, unordered structure and non-grid. Experiments verified that RSMix successfully regularized the deep neural networks with remarkable improvement for shape classification. We also analyzed various combinations of data augmentations including RSMix with single and multi-view evaluations, based on abundant ablation studies.

1. Introduction

Deep neural networks have achieved outstanding performances in various fields regardless of the data domains, such as image, video, speech, and point cloud. In particular, three-dimensional (3D) point cloud processing is attracting considerable interest following the pioneering network PointNet [22] development, since point clouds can be applied directly to deep learning without preprocessing. Although various tasks have been successfully addressed using point clouds, such as 3D object shape classification and part segmentation, inherent drawback of deep learning is still less considered in the point cloud domain. Due to

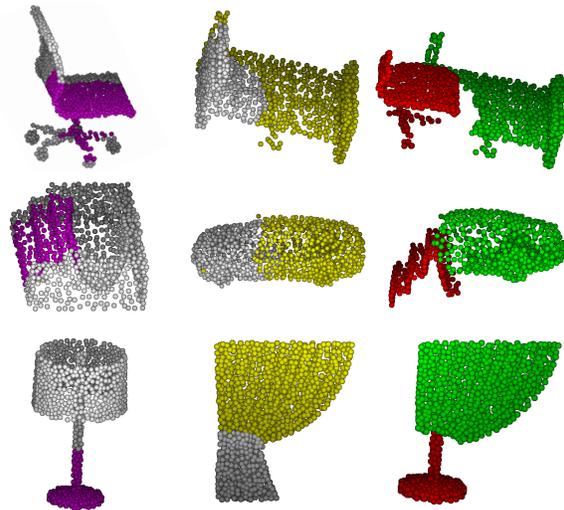


Figure 1: Qualitative results with RSMix. Purple (left) and yellow (middle) colored points indicate Rigid Subsets to be extracted from each sample to synthesize red and green colored mixed samples (right).

the typical nature of deep neural networks (DNNs) that approximates the model from the given data distribution, the trained model tends to be overfitted regardless of the data domain. This lack of generality is a fundamental deep learning problem. One way to alleviate overfitting and generalize the model is data augmentation, which improves diversity of the training data.

Various data augmentation methods have been recently proposed in the image domain as network regularization strategies, but data augmentation for point clouds has only rarely been considered. Actually, regularization is essential for point clouds since it is easier to be biased to the distribution of training samples than that of image. That is largely because point cloud datasets [31, 2, 6] are typically considerably smaller and less diverse than image datasets, such as ImageNet [7] and MSCOCO [20], which have millions of training data. For example, ModelNet40 [31], one of the

¹Project page: <https://github.com/dogyoonlee/RSMix>

most widely used point cloud dataset, includes only 12,311 models with 40 categories. Therefore, it is essential to improve the generality of models for point cloud.

In the image domain, regional dropout [38, 25, 10, 8] and mixup-based methods [34, 28, 33, 14, 11] have been proposed as data augmentation strategies, which are different to conventional methods, to generate virtual training samples. These methods are designed to improve generality of the neural network, preventing the model from being significantly affected by only small part of the sample that has discriminative characteristics by eliminating or mixing the part of the data. However, it is difficult to apply this intuition directly to the point clouds due to two inherent properties of point cloud: non-grid and order-invariance. Although Chen *et al.* [4] applied the concept of Mixup [34] to point clouds handling the properties of point cloud with linear interpolation based on optimal assignment, generated samples lost the structural information of the original sample due to distortion.

This paper proposes Rigid Subset Mix (RSMix), the shape-preserving data augmentation method for point clouds that can partially mix two samples, preserving the partial shapes of the original samples. We redefine the concept of mask region from image analysis and adapt it to 3D space to extract parts from each sample while preserving structural information of the point cloud. We also define a Rigid Subset (RS) derived from the redefined mask region, a group of adjacent points within a certain distance from a specific query point using a neighboring function to address unique characteristics: unordered structure and non-grid. In contrast to PointMixup [4], we can utilize structural information of the original point cloud sample intactly by using RS. In addition, we designed RS scale to vary, to improve diversity of the training sample, and hence increase regularization effects. Furthermore, RSMix can be used in conjunction with the existing data augmentation since it utilizes the part of the given data intactly. In the end, by introducing RS, we can improve generality of DNNs and give attention them to recognize parts of the object. In Section 3.2, we describe in detail how to generate the virtual sample preserving shape of the source sample by extracting RS. In advance, we provide visualized RS samples to be extracted and resultant mixed samples in Figure 1.

We provide the experimental results for shape classification on ModelNet40 [31] and ModelNet10 [31] with the most representative DNN approaches [23, 29] for point clouds. RSMix successfully improved the network performance, outperforming the existing data augmentation methods. Moreover, abundant ablation studies for various combinations of existing data augmentation and RSMix verified that RSMix improved the model regardless of which conventional data augmentation method was employed.

Meanwhile, we analyzed the experimental results with

respect to two evaluation mechanisms to ensure fair comparisons. In fact, although the evaluation methods of shape classification on point cloud are divided into two ways: single and multi-view, many studies present their experimental results without clearly specifying their mechanism. This makes hard to quantitatively compare results among studies. Our experiments show that the results evaluated by single and multi-view approaches have significant differences. Therefore, it is essential to analyze experimental results along the evaluation methods. Section 4.2 presents analysis with single and multi-view approaches based on ablation studies.

To summarize, this paper provides the following major contributions.

- *Shape preserving augmentation.* We propose new data augmentation method for point clouds that mixes training samples with preserved structures by using Rigid Subset (RS).
- *Significant improvement.* The proposed method remarkably improves DNN performances and robustness for shape classification and outperforms existing data augmentation strategies.
- *Complementary method.* RSMix can be used in conjunction with other data augmentation approaches. Abundant ablation studies verify that RSMix can be combined well with other augmentations and further improves the target model.

2. Related Work

Data Augmentation for Images. Data augmentation is a regularization methods that expands the knowledge range that can be learned from training data by transforming data while retaining the essential sample meaning. Thus, the model becomes less dependent on the specific given data. Various methods have been proposed in the image domain in addition to conventional methods, such as random rotation, flip and crop. Some works have enabled the model to learn spatially distributed representation by removing the part of the data on pixel [38, 25, 8] or feature map [10] basis. Furthermore, several mixup-based methods [34, 28, 33, 14, 11] have been proposed that generate virtual samples by combining the two samples.

Mixup [34] generates virtual training samples by linearly interpolating two images and defining the mixed area ratio as a corresponding label. By introducing the combination between data, Mixup brings out the regularization effect and shows improved performance for several tasks. After Mixup, Verma *et al.* [28] extended Mixup by applying the concept to the feature map, and Yun *et al.* [33] fused the concept of [34] and [8] to improve localization and classification ability of the model. In addition, Kim *et al.* [14]

Method	Mix function $f_{M_d}(x_\alpha, x_\beta)$
Mixup [34]	$(1 - \lambda)x_\alpha + \lambda x_\beta$
Manifold Mixup [28]	$(1 - \lambda)f(x_\alpha) + \lambda f(x_\beta)$
CutMix [33]	$(1 - \mathcal{M}) \odot x_\alpha + \mathcal{M} \odot x_\beta$
Puzzle Mix [14]	$(1 - z) \odot \Pi_\alpha^T x_\alpha + z \odot \Pi_\beta^T x_\beta$
F-Mix [11]	$(1 - \mathcal{H}(\mathcal{G})) \odot x_\alpha + \mathcal{H}(\mathcal{G}) \odot x_\beta$
PointMixup [4]	$(1 - \lambda)x_\alpha + \lambda \mathcal{J}_{\phi^*}(x_\alpha, x_\beta)$

Table 1: Various mixup functions for image and point cloud domains.

and Harris *et al.* [11] utilized saliency maps [14] and Fourier transform [11], respectively, to use semantically representative parts of the data when generating virtual samples. However, these approaches can only be applied to image based models rather than point clouds, because they have different data structures. Therefore, we propose the RSMix, novel mixup-based augmentation strategy for the point cloud that generates virtual samples considering the unique properties of point clouds.

Point Cloud Structural Properties. In contrast with images, point clouds have 3D coordinate information, including implicit geometric feature, which is essential to understand them. Due to the unique properties of point cloud: non-grid and unordered structures, it is difficult to extract the local and geometric feature from point clouds. Various networks have been proposed with different structures, such as point-wise multi layer perceptron (MLP) [23, 13, 37, 32], convolution [21, 26, 30, 1, 12, 17, 19, 16], graph [29, 35, 24, 36], and spatial partitioning structured [9, 15] based networks to extract local and geometric features. For example, [23, 26, 29, 15] extract local and geometric features by applying point-wise grouping, radius-based kernel, graph-structure, and space partitioning tree, respectively. All these networks demonstrated that considering structural information is significantly important for the DNN model to understand point clouds. Therefore, regardless of how data augmentation occurs, structural information of point cloud should be regarded as core characteristics since it is a critical component for the model.

Data Augmentation on Point Cloud. Data augmentation has not been extensively explored in the point cloud domain, aside from general conventional methods, such as randomly scaling, rotation, and jittering. Few studies [18, 5, 4] dealt with data augmentation in the point cloud domain. Liu *et al.* [18] proposed auto-augmentation network for point clouds to find an optimal combination of conventional data augmentation methods corresponding to each sample. Choi *et al.* [5] divided the sample into specific partitions and transformed or mixed each part independently. However, there is a limit to diversity of virtual sample because mixing is performed on inter-classes and specified grids are used for partitioning. PointMixup [4],

which is the closest method to our proposed approach, extends the concept of Mixup [34] to point clouds through linear interpolation with optimal assignments between two samples. However, the generated samples have distorted structures which lead to the loss of structural information. Structural information is a core point cloud feature since they have no textural information. Therefore, we propose a more general data augmentation method for point clouds that can preserve shape of the original data.

3. Method

3.1. Preliminary

Neural networks aim to model function f that describes the true distribution P for given data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, where samples $x \in \mathcal{X}$ have corresponding labels $y \in \mathcal{Y}$. It has been proved through *Empirical Risk Minimization* [27] that f can be approximated by minimizing empirical risk $R_\xi(f)$ of the model by computational optimization using loss \mathcal{L} and empirical distribution P_ξ for given data distribution as

$$R_\xi(f) = \int \mathcal{L}(f(x), y) dP_\xi(x, y) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(f(x_i), y_i) \quad (1)$$

In data augmentation, P_ξ can be expanded to P_ψ with additional augmented data through *vicinal risk minimization* [3],

$$P_\psi(\tilde{x}, \tilde{y}) = \frac{1}{n} \sum_{i=1}^n \psi(\tilde{x}, \tilde{y} | x_i, y_i), \quad (2)$$

where ψ is a vicinity distribution, *i.e.*, the probability that virtual sample and label pair (\tilde{x}, \tilde{y}) , are sampled from the vicinity of given sample and label pair (x_i, y_i) . For image data, Zhang *et al.* [34] designed a vicinal distribution ψ that generated a virtual mixed sample-label pair (\tilde{x}, \tilde{y}) from two paired data (x_α, y_α) and (x_β, y_β) using mix function f_{M_d} , for sample and f_{M_l} , for label, as

$$\begin{aligned} \tilde{x} &= f_{M_d}(x_\alpha, x_\beta) = (1 - \lambda)x_\alpha + \lambda x_\beta, \\ \tilde{y} &= f_{M_l}(y_\alpha, y_\beta) = (1 - \lambda)y_\alpha + \lambda y_\beta, \end{aligned} \quad (3)$$

where $\lambda \sim$ beta distribution $\text{Beta}(\theta, \theta)$, for $\theta \in (0, \infty)$. Table 1 shows the deformations of f_{M_d} in various ways using features from model f [28] or masking approaches such as binary mask \mathcal{M} [33], salient data included mask z [14], or thresholding mask \mathcal{H} [11] with filtered data \mathcal{G} in the frequency domain. However, these mask-based approaches cannot be applied directly to point cloud, since point clouds have no grid and points can exist anywhere in 3D real space. Though Chen *et al.* [4] solved this problem by linear interpolation between two point clouds, introducing optimal assignment \mathcal{J}_{ϕ^*} , they could not generate virtual samples preserving shape of the original sample. Therefore, our goal is to generate a shape-preserved virtual sample that has combined information from both samples as well as proposing an adapted spatial mask for 3D data. We are inspired by

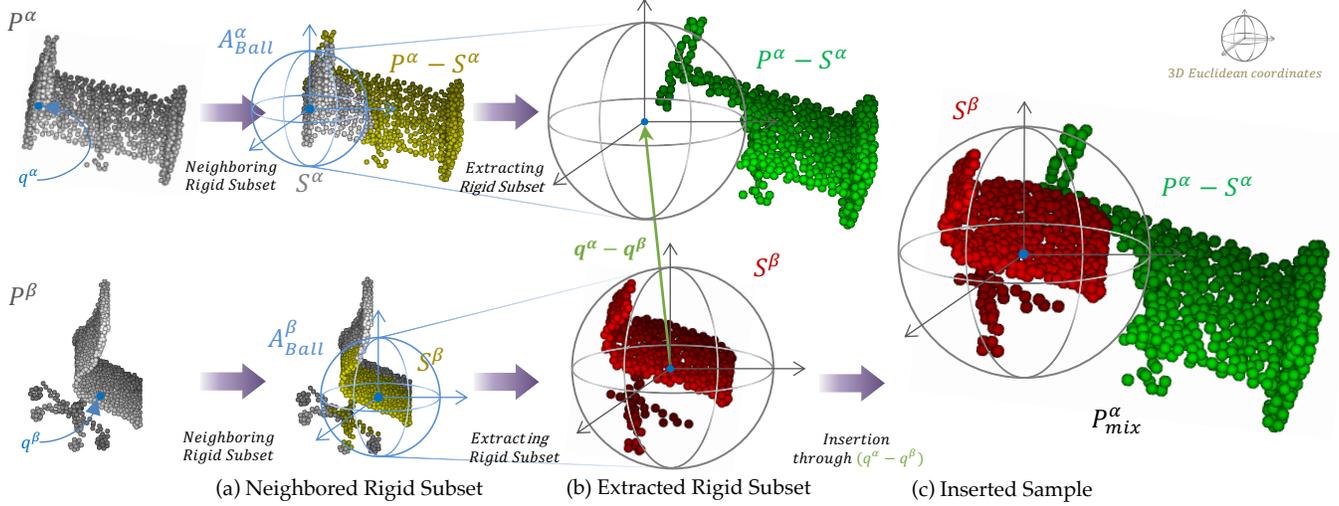


Figure 2: Overall pipeline of RSMix. Three steps to synthesize the mixed samples (\mathcal{P}_{mix}^α) using Rigid Subset (RS).

concept of the mask region from image analysis, which preserves the part of the original data intactly.

Mask as Region of Neighboring Data. CutMix [33] defined the mix function as

$$f_{M_d}(x_\alpha, x_\beta) = (1 - \mathcal{M}) \odot x_\alpha + \mathcal{M} \odot x_\beta, \quad (4)$$

where \odot represents the element-wise multiplication; mask \mathcal{M} denotes a $d_u \times d_w$ binary rectangular region represented as $[u_M, u_M + d_u] \times [w_M, w_M + d_w]$ with mixture ratio, $\lambda = \frac{d_u d_w}{WH}$, where (u_M, w_M) is a randomly chosen pixel of the image. To utilize inherent definition of the mask region, we redefine mask \mathcal{M} as a group of successive adjacent pixels within distances $d_u/2$ and $d_w/2$ in image. Hence, mask \mathcal{M} can be denoted as

$$\mathcal{M} = \{(u_i, w_j) \mid |u_i - u_c| \leq \frac{d_u}{2}, |w_j - w_c| \leq \frac{d_w}{2}\}, \quad (5)$$

where (u_i, w_j) is the (i, j) th pixel for the given image; $i = \{1, 2, \dots, W\}$; $j = \{1, 2, \dots, H\}$; and (u_c, w_c) is the center of the mask $(u_M + \frac{d_u}{2}, w_M + \frac{d_w}{2})$. Thus, the mask can be regarded as an adjacent group of data from a particular point (u_c, w_c) . We adapt this definition of mask to point clouds.

3.2. Rigid Subset Mix

Rigid Subset Mix (RSMix) mixes parts of two point cloud samples by extracting the Rigid Subset (RS), which preserves each samples's shape. RSMix is divided in three steps: neighboring, extraction, and insertion. First, we utilize the redefined mask region concept from Section 3.1 with neighboring function \mathcal{A} to prevent deformation of original data. Then we extract RSs from each sample to mix the samples. Finally, we mix two RSs in the insertion step. Figure 2 shows the RSMix algorithm pipeline.

Neighboring Rigid Subset. We define two n sampled point sets normalized in the unit sphere as $\mathcal{P}^t = \{p_i^t \mid i =$

$1, 2, \dots, n\}$, where $t \in \{\alpha, \beta\}$. $p_i^t \in \mathbb{R}^3$ is its Euclidean coordinates, which represents location of the point. We only consider coordinate information since RSMix operates on point-wise coordinates.

We adapt the regional mask for image to spatial subset of each point sets from given point sets \mathcal{P}^α and \mathcal{P}^β , by grouping adjacent points from a certain query point, q^t , randomly chosen from \mathcal{P}^t . These subsets are denoted as \mathcal{S}^α and \mathcal{S}^β according to below Equation (6).

$$\mathcal{S}^t = \mathcal{A}(\mathcal{P}^t; q^t), \quad (6)$$

which are grouped using the specific neighboring function \mathcal{A} . We define these subsets as Rigid Subset (RS) since they preserve the sample shape rigidly.

We introduce two instantiations for \mathcal{A} to retain the original point set shape: K-Nearest Neighbor(KNN) to given q^t and Ball-query method that neighboring points in certain distance r_{rigid} from q^t as

$$\begin{aligned} \mathcal{A}_{knn}(\mathcal{P}^t; q^t) &= \{p^t \mid p^t \text{ is KNN of } q^t, p^t \in \mathcal{P}^t\}, \\ \mathcal{A}_{ball}(\mathcal{P}^t; q^t) &= \{p^t \mid \|p^t - q^t\| \leq r_{rigid}, p^t \in \mathcal{P}^t\}, \end{aligned} \quad (7)$$

respectively, where r_{rigid} is sampled from beta distribution $\text{Beta}(\theta, \theta)$, with parameter $\theta = 1.0$ as default, *i.e.*, the uniform distribution, since P^t is normalized in the unit sphere. Both \mathcal{A} are based on Euclidean distance in 3D space considering the point cloud's unordered structure and free space around them. Each method has different characteristics on neighboring subsets with respect to the density or directional bias of given point sets.

Meanwhile, we limit $|\mathcal{S}^\beta| \leq n^{max}$, where n^{max} and $|\cdot|$ denote the upper bound number of points in RS and cardinality for the point set, respectively. We usually set $n^{max} = |\mathcal{P}^t|/2$ to preserve at least half of the original point sets. In addition, when using \mathcal{A}_{ball} , we randomly sample

points in \mathcal{S}^β along the difference between $|\mathcal{S}^\alpha|$ and $|\mathcal{S}^\beta|$ to maintain the $|\mathcal{P}_{mix}^\alpha|$, where \mathcal{P}_{mix}^α denotes a mixed sample described in below Insertion part. We compare and analyze the two methods for quantitative and visualized results in Section 4.2. Further, we also provide experiments with various θ values in Section 4.2. Figure 2(a) shows neighboring the RS from the each sample.

Extraction. Neighbored RSs, \mathcal{S}^α and \mathcal{S}^β , are used to a generate mixture sample \mathcal{P}_{mix}^α . To mix two samples, we remove the \mathcal{S}^α from \mathcal{P}^α and replace the empty space with \mathcal{S}^β . Hence, extracted RSs from each point cloud sample to generate mixture samples are denoted as $\mathcal{P}^\alpha - \mathcal{S}^\alpha$ and \mathcal{S}^β as shown in Figure 2(b).

Insertion. However, q^α and q^β are usually different because they are randomly chosen from \mathcal{P}^α and \mathcal{P}^β , respectively. Hence, before insertion, \mathcal{S}^β should be translated by the difference between the two query points. We introduce translation function $\mathcal{T}^{\beta \rightarrow \alpha}$ to translate \mathcal{S}^β by $q^\alpha - q^\beta$ as

$$\mathcal{T}^{\beta \rightarrow \alpha}(\mathcal{S}^\beta; q^\alpha, q^\beta) = \{p^{\beta \rightarrow \alpha} \mid p^{\beta \rightarrow \alpha} = p^\beta + (q^\alpha - q^\beta)\}, \quad (8)$$

where p^β is a point in \mathcal{S}^β . Applying $\mathcal{T}^{\beta \rightarrow \alpha}$ to \mathcal{S}^β , the translated subset $\mathcal{S}^{\beta \rightarrow \alpha}$ is denoted as

$$\mathcal{S}^{\beta \rightarrow \alpha} = \mathcal{T}^{\beta \rightarrow \alpha}(\mathcal{S}^\beta; q^\alpha, q^\beta). \quad (9)$$

Therefore, mixed sample \mathcal{P}_{mix}^α is defined as

$$\mathcal{P}_{mix}^\alpha = (\mathcal{P}^\alpha - \mathcal{S}^\alpha) \cup \mathcal{S}^{\beta \rightarrow \alpha}, \quad (10)$$

and Figure 2(c) describes the inserted mixture sample. Thus, mix function $f_{M_d}(x_\alpha, x_\beta)$ for RSMix can be expressed using as follow Equation (11) using \mathcal{P}^α and \mathcal{P}^β instead of x_α and x_β .

$$f_{M_d}(\mathcal{P}^\alpha, \mathcal{P}^\beta) = (\mathcal{P}^\alpha - \mathcal{A}(\mathcal{P}^\alpha)) \cup \mathcal{T}^{\beta \rightarrow \alpha}(\mathcal{A}(\mathcal{P}^\beta)), \quad (11)$$

where input arguments related to query points q^α and q^β are omitted for clarity.

3.3. Mixture Ratio λ for Training

In this Section, we define the mixture ratio λ , the ratio of $|\mathcal{S}^{\beta \rightarrow \alpha}|$ w.r.t. $|\mathcal{P}_{mix}^\alpha|$, to train the network for shape classification. In contrast to \mathcal{A}_{knn} or some previous image masks, $|\mathcal{P}^\alpha - \mathcal{S}^\alpha|$ and $|\mathcal{S}^\beta|$ are often different when using the \mathcal{A}_{ball} , since we apply same r_{rigid} to \mathcal{P}^α and \mathcal{P}^β , despite of their different densities. Hence, we define

$$\lambda = \begin{cases} 0, & \text{if } \mathcal{P}^\alpha = \mathcal{S}^\alpha, \\ 0, & \text{if } \mathcal{S}^\beta = \emptyset, \\ |\mathcal{S}^\beta| / (|\mathcal{P}^\alpha - \mathcal{S}^\alpha| + |\mathcal{S}^\beta|), & \text{Otherwise,} \end{cases} \quad (12)$$

To explicitly consider the relation between $|\mathcal{P}^\alpha - \mathcal{S}^\alpha|$ and $|\mathcal{S}^\beta|$. Finally, we define the label mix function as

$$f_{M_l}(y_\alpha, y_\beta) = (1 - \lambda)y_\alpha + \lambda y_\beta, \quad (13)$$

which is same as in CutMix [33], to generate virtual label \tilde{y} for classification training. Detailed implementations are available in the supplementary material with pseudo-code.

4. Experiments

Datasets. We evaluate RSMix on ModelNet40 [31] and ModelNet10 [31], which are widely used point cloud classification benchmark datasets. ModelNet40 is small dataset which comprises 12,311 CAD models from 40 man-made object categories, and ModelNet10 is subset of ModelNet40 that includes only 4899 CAD models from 10 categories. We utilized the preprocessed data provided by PointNet [22] for ModelNet40 with same train-test split, which were 1024 uniformly sampled points on mesh faces according to face area and normalized onto the unit sphere, and preprocessed ModelNet10 similarly. In particular, we ignored normals of samples since they are not available for real-data.

Backbone Networks. We considered three representative point-wise DNNs for point cloud: PointNet++ [23], DGCNN [29], and PointNet [22] as our backbone network architecture. We applied RSMix to several neural networks to emphasize RSMix is model agnostic.

Single and Multi-view Evaluations. Single and multi-view evaluations are separated depending on whether objects were evaluated from different angles or not. These approaches can be separated into two cases: with or without voting strategy to predict an object multiple times by rotating about an axis. The experiments adopted voting strategy of evaluating an object 12 times, rotating it 30° on its vertical (y) axis between evaluations. Meanwhile, ModelNet40 [31] has 10 classes with aligned poses/headings. Thus, it is trivial to separate the 10 classes with the remaining 30 classes if we don't do random rotation on test samples when evaluating the model. Hence, there are obvious differences between the results from single and multi-view evaluations. Appropriate combinations of augmentation strategies also vary depending on evaluation type. We investigated results from both evaluation strategies and explored optimal combinations of augmentations for different models by abundant ablation studies (Sections 4.2).

Implementation details. We implemented RSMix using PointNet++ [23] and DGCNN [29] with conventional data augmentation, ConvDA (comprising jittering($\sigma^2=0.01$); scaling(0.8~1.25); rotation along the y-axis *i.e.*, gravity axis; and shifting (range=0.1) for the training dataset. Further details are included in the supplementary material.

4.1. Shape Classification

Evaluations. We evaluate RSMix for shape classification using three backbone networks on ModelNet40 and ModelNet10. All experiments were implemented using official codes and results are shown in Table 2. "Multi" indicates the evaluation with multi-view. To ensure fair comparison given the rotational bias in ModelNet40, we exclude experimental results of point-wise MLP networks [22, 23] trained without rotational augmentation in Table 2. Section 4.2 presents a rotation-related ablation study.

Method	#Points	Evaluation Accuracy(%)	
		ModelNet40	ModelNet10
PointNet[22]	1k	88.5	93.1
PointNet[22] (Multi)	1k	88.4	92.5
PointNet++[23]	1k	91.0	93.3
PointNet++[23] (Multi)	1k	91.0	93.5
DGCNN[29]	1k	92.8	94.8
PointNet +Ours	1k	88.7(0.2) ↑	93.1(-)
PointNet +Ours (Multi)	1k	88.5(0.1) ↑	92.6(0.1) ↑
PointNet++ +Ours	1k	91.6(0.6) ↑	94.3(1.0) ↑
PointNet++ +Ours (Multi)	1k	92.1(1.1) ↑	94.4(0.9) ↑
DGCNN +Ours	1k	93.5(0.7) ↑	95.9(1.1) ↑

Table 2: Quantitative results for single and Multi-view evaluations of RSMix on ModelNet40 [31]. We only present the results of PointNet [22] and PointNet++ [23] with rotational augmentation included model for fair comparison.

Method	Augmentation	ACC.(%)	Dataset Align	Eval
PointNet++ [23]	PointMixup [4]	92.7	Pre-aligned	-
DGCNN [29]	PointMixup [4]	93.1	Pre-aligned	-
PointNet [22]	PointMixup [4]	89.9	Unaligned	-
PointNet++ [23]	PointMixup [4]	91.7	Unaligned	-
PointNet++ [23]	Ours	92.7	Raw	Single-View
DGCNN [29]	Ours	93.5	Raw	Single-View
PointNet [22]	Ours	88.5	Raw	Multi-View
PointNet++ [23]	Ours	92.1	Raw	Multi-View

Table 3: Comparing RSMix and PointMixup [4] on ModelNet40 [31].

All the results reveal that RSMix improved the network accuracies regardless of network type or evaluation methods, verifying the effectiveness of our shape-preserved mixture approach with significant improvements for PointNet++ [23] and DGCNN [29], which encode local or geometric features of object through hierarchical grouping or graph structure, respectively.

Comparison against PointMixup [4]. We demonstrate results of RSMix for two evaluation methods against PointMixup [4], the closest work to us, in Table 3. We compared pre-aligned and unaligned settings to for single and multi-view accuracies, respectively, since PointMixup [4] do not specify their evaluation method but each are similar. They follow the PointCNN [19] setting discriminating pre-aligned and unaligned with horizontal rotation on point cloud samples. They randomly rotate the training point cloud along the y-axis for unaligned settings. For natural evaluation, we do not preprocess the dataset as pre-aligned or unaligned (denoted as Raw in Table 3). RSMix achieves more competitive performance than PointMixup [4] for networks that use local information [23, 29], and further enhances the networks ability to recognize local information.

Visualization. Supplementary material provides additional examples synthesized with RSMix.

ConvDA	RandDrop	RSMix	$ACC_S^{knn}(\%)$	$ACC_S^{ball}(\%)$
		✓	93.0(0.5)↑	93.3(0.8)↑
✓		✓	93.3(0.7)↑	93.4(0.8)↑
✓	✓	✓	93.4(0.6)↑	93.5(0.7)↑

Table 4: Quantitative Comparison of neighboring functions for DGCNN [29] on ModelNet40. ACC_S^{knn} and ACC_S^{ball} indicate single-view accuracy with \mathcal{A}_{knn} and \mathcal{A}_{ball} .

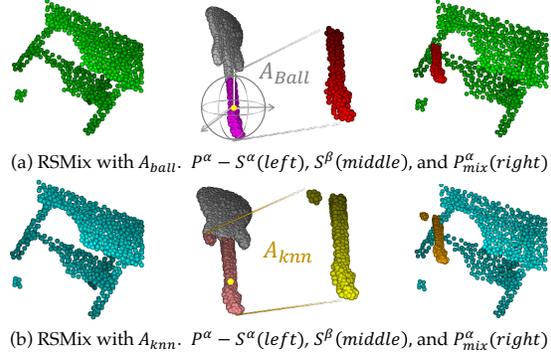


Figure 3: Differences depending on directional bias of point cloud.

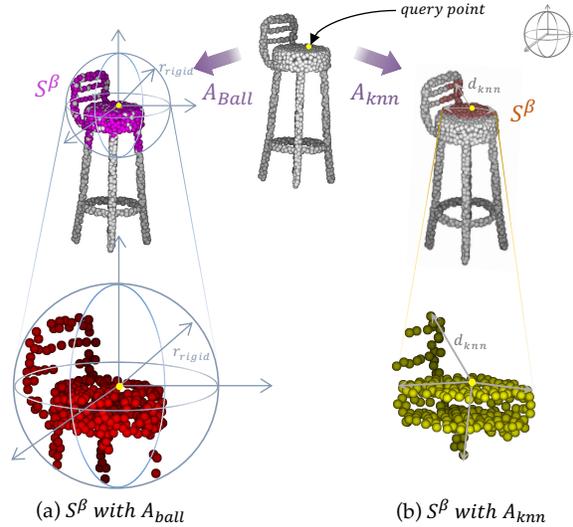


Figure 4: Differences depending on density of point cloud.

4.2. Ablation Study

Neighboring Function. This Section describes the two types of neighboring function, \mathcal{A}_{ball} and \mathcal{A}_{knn} , employed to extract RSs from point cloud samples. Figure 3 and 4 show the differences between these functions under directional bias or different densities between point cloud samples qualitatively. Although \mathcal{A}_{knn} extracts the subset pre-

Jitter+Shift	Rotation	Scaling	RandDrop	RSMix	ACC _S	ACC _M
					91.5	78.5
				✓	92.7(1.2↑)	71.5(7.0↓)
✓					91.4	73.5
✓				✓	92.0(0.6↑)	74.4(1.1↑)
✓			✓		91.4	67.1
✓			✓	✓	91.8(0.4↑)	72.8(5.7↑)
	✓	✓			91.0	90.8
	✓	✓		✓	91.6(0.6↑)	92.1(1.3↑)
	✓	✓	✓		91.0	91.0
	✓	✓	✓	✓	91.3(0.3↑)	91.2(0.2↑)
✓	✓	✓			90.3	90.7
✓	✓	✓		✓	90.8(0.5↑)	91.4(0.7↑)
✓	✓	✓	✓		90.6	90.7
✓	✓	✓	✓	✓	91.0(0.4↑)	91.1(0.4↑)

Table 5: Ablation studies on evaluation accuracy with single(ACC_S(%)) and multi-view(ACC_M(%)) for PointNet++ [23] on ModelNet40.

serving shape of the sample based on Euclidean distance d_{knn} , it is prone to be overlapped with the other extracted part, e.g. Figure 3(b), if there is directional bias on the sample. On the other hand, \mathcal{A}_{ball} alleviates overlapping by selecting points within the distance, r_{rigid} e.g. Figure 3(a). However, in contrast with \mathcal{A}_{knn} , if the density of sample around the query point is high or r_{rigid} is too large, the number of points in RS from \mathcal{A}_{ball} must be controlled to maintain the number of points in a mixed sample. Therefore, we randomly sample the extracted points in \mathcal{S}^β to preserve the overall shape of the extracted part. Figure 4 illustrates the difference between processed RS with \mathcal{A}_{ball} and \mathcal{A}_{knn} depending on density of point cloud. We also compare two methods with quantitative results of single-view evaluation on DGCNN [29] in Table 4, using scaling and shift augmentations as ConvDA. Both approaches achieved improved results over existing networks and \mathcal{A}_{ball} achieves superior results since it is more robust to directional bias as well as considers the density of point cloud. Therefore, we used \mathcal{A}_{ball} as our main neighboring function.

Single and multi-view evaluations with various combinations of augmentations. RSMix can be applied in conjunction with existing ConvDA methods to further increase the diversity of mixed data since they are independent approaches. However, some combinations of augmentations can cause excessive deformation on the data sample, reducing the network’s ability to recognize objects itself. Therefore, it is essential to analyze various combinations of augmentation strategies. We provide quantitative results in Table 5 with single and multi-view evaluations for PointNet++ [23] on ModelNet40. RandDrop is the data augmentation method proposed in [23] that randomly drops the points from sample so that network can extract the global features better. The results show that models with RSMix alone achieved the highest accuracy for single-view evaluation. In addition, overall experiments show better results without rotational augmentation for single-view evaluation.

ConvDA	RandDrop	RSMix	ACC _S (%)	Dataset
			92.5	MN40
		✓	93.3(0.8↑)	MN40
✓			92.6	MN40
✓		✓	93.4(0.8↑)	MN40
✓	✓		92.8	MN40
✓	✓	✓	93.5(0.7↑)	MN40
			94.6	MN10
		✓	95.9(1.3↑)	MN10
✓			94.8	MN10
✓		✓	95.4(0.6↑)	MN10
✓	✓		94.8	MN10
✓	✓	✓	95.5(0.7↑)	MN10

Table 6: Ablation studies for DGCNN[29] on ModelNet40(MN40) and ModelNet10(MN10). Random scaling augmentation was applied as ConvDA.

However, results with multi-view evaluation reveal that if the model is trained without rotational augmentation, network can be overfitted to directional bias of the dataset, ModelNet40. Hence, rotational augmentation is essential for multi-view evaluation. However, RSMix improves discriminative ability of the model with appropriate combinations with other augmentations regardless of evaluation type, because diversity of datasets increases significantly when RSMix is used in conjunction with rotation and scaling augmentations.

In addition, we also provide the results for DGCNN [29] in Table 6 with single-view evaluation on ModelNet40 and ModelNet10 with scaling augmentation as ConvDA, since single-view evaluation shows better results without rotational augmentation. We also obtained remarkable improvements with RSMix for all presented combinations.

Therefore, we can notice three things as follows.

- Single and multi-view evaluation performances differ significantly depending on the presence of rotational biases in the dataset.
- Rotational augmentation reduces single-view evaluation performance, but must be included when training if evaluation is performed with multi-view evaluation.
- RSMix successfully improved model generality by appropriate combination with other augmentation strategies regardless of evaluation type.

Robustness Test. We tested the robustness of RSMix with PointNet++ [23] to four noisy environments: jitter, rotation, scaling, and DropPoint, in order to verify that our method makes the model robust to noise. Table 7 verifies the impact of RSMix with single-view and multi-view evaluation for 2 cases against the use of the ConvDA. Especially, multi-view evaluation for rotational noisy environment along the y-axis was performed by rotating the sample

Transform	×	RSMix	ConvDA w/o RSMix	ConvDA w/ RSMix	Eval
Jitter ($\sigma^2=0.01$)		90.0	90.8	91.0	Single
		82.3	78.5	90.9	Multi
Jitter ($\sigma^2=0.05$)		15.3	23.4	21.2	Single
		10.3	9.8	20.3	Multi
X-axis 90°		17.3	16.5	19.9	Single
		18.0	16.7	19.8	Multi
Y-axis 90°		56.7	90.0	91.0	Single
		57.0	90.1	91.1	Multi
Z-axis 90°		14.6	15.2	18.3	Single
		14.9	14.8	18.4	Multi
X-axis 180°		44.6	43.3	45.7	Single
		44.9	43.5	45.6	Multi
Y-axis 180°		75.5	90.2	91.0	Single
		75.0	90.1	91.1	Multi
Z-axis 180°		42.3	43.8	44.4	Single
		41.7	43.3	44.4	Multi
Scale (0.6)		90.7	90.4	91.0	Single
		83.7	82.5	90.5	Multi
Scale (1.4)		90.7	90.5	91.2	Single
		83.4	82.7	90.6	Multi
Scale (2.0)		90.7	90.4	91.0	Single
		83.6	82.6	90.5	Multi
DropPoint (0.2)		88.2	87.4	91.0	Single
		77.1	88.2	90.9	Multi

Table 7: Robustness test of RSMix with or without ConvDA for PointNet++ [23] on ModelNet40 using Random shift, scaling, rotation, and jitter augmentations as ConvDA.

along the x-axis for fair comparison. ConvDA includes jittering, shifting, scaling, and rotational augmentations with default settings as same as PointNet++ [23]. The results in Table 7 reveal that RSMix improves the robustness of model whether or not ConvDA was applied for rotation and Drop-Point noise, since shape and scale of original point cloud were preserved. However, we achieved lower results with multi-view evaluation when only RSMix was applied for scaling noise. The reason is that if scaling noise is large, it is difficult for the model to interpret the data when viewed from different angles since shape is preserved but scale compared with the original data. Meanwhile, results differed greatly depending on the level of noise for jittering noise, where the shape of an object was not preserved. Although RSMix alone cannot improve robustness for multi-view evaluation, RSMix provided improvements when jitter noise was small for single-view evaluation regardless of ConvDA usage. However, robustness was reduced for both evaluation methods for large jitter noise when RSMix was applied because it was difficult for subsets extracted from each sample by RSMix to have sufficient shape information since the original sample shape was greatly distorted prior to applying RSMix.

Various θ values. We introduced beta function $\text{Beta}(\theta, \theta)$, to sample r_{rigid} from the beta distribution when using \mathcal{A}_{ball} in Section 3.2. We demonstrate the experimental results for

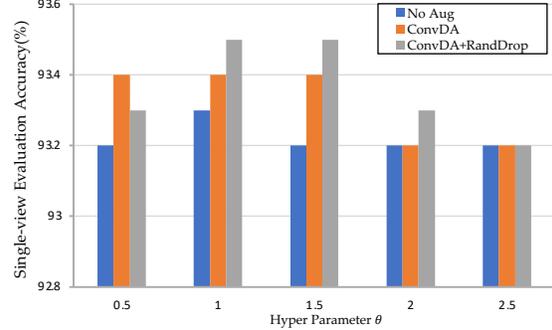


Figure 5: Ablation studies for various θ values on DGCNN [29] with single-view evaluation on ModelNet40.

various θ for DGCNN [29] on ModelNet40 [31] in Figure 5 to figure out the effect of θ to our model using same experimental settings. Due to a property of beta function, larger θ implies higher probability that r_{rigid} was sampled close to 0.5. However, since we set the $n^{max} =$ half of the number of points in sample, more frequent sampling of large r_{rigid} does not have much effect. Best accuracy was achieved for $\theta = 1.0$ for all cases. Therefore, we set $\theta = 1.0$ as default for all experiments.

5. Conclusion

This paper proposes RSMix, a novel data augmentation method for point clouds, that generates virtual mixed samples from extracted subsets from each point cloud without additional learnable parameters. We extracted the subsets from samples without shape distortion by redefining the rectangular mask for images as a subset of adjacent points from a query point in 3D space. Various experiments verified that RSMix improved deep neural networks to extract discriminative feature effectively by increasing diversity of data. In addition, extensive tests demonstrated that RSMix also improved robustness of the model to various types of noise. We further analyze the two types of evaluation method for shape classification: single and multi-view, which are utilized as evaluation metrics for the overall experiments. Experiments verified explicit differences between two methods and necessity of selecting appropriate combination with various data augmentation strategies. Extensive ablation studies also verified generic effectiveness of RSMix with various combinations with existing data augmentations.

Acknowledgements. This work was supported by the Institute for Information and Communications Technology Promotion (IITP) funded by the Korean Government (MSIP) under Grant 2016-0-00197. This research was also supported by R&D program for Advanced Integrated-intelligence for Identification (AIID) through the National Research Foundation of KOREA(NRF) funded by Ministry of Science and ICT (NRF-2018M3E3A1057289).

References

- [1] Alexandre Boulch. Convpoint: Continuous convolutions for point cloud processing. *Computers & Graphics*, 2020.
- [2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.
- [3] Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. In *Advances in neural information processing systems*, pages 416–422, 2001.
- [4] Yunlu Chen, Vincent Tao Hu, Efstratios Gavves, Thomas Mensink, Pascal Mettes, Pengwan Yang, and Cees GM Snoek. Pointmixup: Augmentation for point clouds. *arXiv preprint arXiv:2008.06374*, 2020.
- [5] Jaeseok Choi, Yeji Song, and Nojun Kwak. Part-aware data augmentation for 3d object detection in point cloud. *arXiv preprint arXiv:2007.13373*, 2020.
- [6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017.
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [8] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.
- [9] Matheus Gadelha, Rui Wang, and Subhransu Maji. Multiresolution tree networks for 3d point cloud processing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018.
- [10] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. In *Advances in Neural Information Processing Systems*, pages 10727–10737, 2018.
- [11] Ethan Harris, Antonia Marcu, Matthew Painter, Mahesan Niranjan, and Adam Prügel-Bennett Jonathon Hare. Fmix: Enhancing mixed sample data augmentation. *arXiv preprint arXiv:2002.12047*, 2020.
- [12] Binh-Son Hua, Minh-Khoi Tran, and Sai-Kit Yeung. Pointwise convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 984–993, 2018.
- [13] Mor Joseph-Rivlin, Alon Zvirin, and Ron Kimmel. Momen (e) t: Flavor the moments in learning to classify shapes. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.
- [14] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. *arXiv preprint arXiv:2009.06962*, 2020.
- [15] Roman Klokov and Victor Lempitsky. Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 863–872, 2017.
- [16] Shiyi Lan, Ruichi Yu, Gang Yu, and Larry S Davis. Modeling local geometric structure of 3d point clouds using geo-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 998–1008, 2019.
- [17] Huan Lei, Naveed Akhtar, and Ajmal Mian. Octree guided cnn with spherical kernels for 3d point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9631–9640, 2019.
- [18] Ruihui Li, Xianzhi Li, Pheng-Ann Heng, and Chi-Wing Fu. Pointaugument: an auto-augmentation framework for point cloud classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6378–6387, 2020.
- [19] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *Advances in neural information processing systems*, pages 820–830, 2018.
- [20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [21] Yongcheng Liu, Bin Fan, Shiming Xiang, and Chunhong Pan. Relation-shape convolutional neural network for point cloud analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8895–8904, 2019.
- [22] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.
- [23] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017.
- [24] Yiru Shen, Chen Feng, Yaoqing Yang, and Dong Tian. Mining point cloud local structures by kernel correlation and graph pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4548–4557, 2018.
- [25] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *2017 IEEE international conference on computer vision (ICCV)*, pages 3544–3553. IEEE, 2017.
- [26] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6411–6420, 2019.
- [27] VN Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.
- [28] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Ben-

- gio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019.
- [29] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019.
- [30] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9621–9630, 2019.
- [31] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.
- [32] Xu Yan, Chaoda Zheng, Zhen Li, Sheng Wang, and Shuguang Cui. Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5598, 2020.
- [33] Sangdoon Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6023–6032, 2019.
- [34] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.
- [35] Kuangen Zhang, Ming Hao, Jing Wang, Clarence W de Silva, and Chenglong Fu. Linked dynamic graph cnn: Learning on point cloud via linking hierarchical features. *arXiv preprint arXiv:1904.10014*, 2019.
- [36] Yingxue Zhang and Michael Rabbat. A graph-cnn for 3d point cloud classification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6279–6283. IEEE, 2018.
- [37] Hengshuang Zhao, Li Jiang, Chi-Wing Fu, and Jiaya Jia. Pointweb: Enhancing local neighborhood features for point cloud processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5565–5573, 2019.
- [38] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, pages 13001–13008, 2020.