

Learning Rotation-Invariant Representations of Point Clouds Using Aligned Edge Convolutional Neural Networks

Junming Zhang
University of Michigan
junming@umich.edu

Ming-Yuan Yu
University of Michigan
myyu@umich.edu

Ram Vasudevan
University of Michigan
ramv@umich.edu

Matthew Johnson-Roberson
University of Michigan
mattjr@umich.edu

Abstract

Point cloud analysis is an area of increasing interest due to the development of 3D sensors that are able to rapidly measure the depth of scenes accurately. Unfortunately, applying deep learning techniques to perform point cloud analysis is non-trivial due to the inability of these methods to generalize to unseen rotations. To address this limitation, one usually has to augment the training data, which can lead to extra computation and require larger model complexity. This paper proposes a new neural network called the Aligned Edge Convolutional Neural Network (AECNN) that learns a feature representation of point clouds relative to Local Reference Frames (LRFs) to ensure invariance to rotation. In particular, features are learned locally and aligned with respect to the LRF of an automatically computed reference point. The proposed approach is evaluated on point cloud classification and part segmentation tasks. This paper illustrates that the proposed technique outperforms a variety of state of the art approaches (even those trained on augmented datasets) in terms of robustness to rotation without requiring any additional data augmentation.

1. Introduction

The development of low-cost 3D sensors has the potential to revolutionize the way robots perceive the world. For this revolution to be realized, algorithms to interpret and classify the large volumes of point clouds generated by these sensors must be developed. To construct such algorithms, one could be inspired by the successes of deep learning approaches that robustly interpret 2D images in the presence of noise or lighting, rotation, and scaling variability. These deep learning approaches achieve impressive perfor-

mance by relying on representations that enforce lighting, rotation, and scaling invariance. Unfortunately, the lack of a representation that is able to enforce rotation invariance has hindered the application of deep learning techniques to analyze point clouds.

To address this challenge, researchers have typically converted point clouds into regular 2D [3, 20, 6, 21] or 3D [12, 16, 28] grids before applying convolutional neural networks (CNNs) to learn a meaningful representation. Unfortunately, this conversion process degrades the resolution of measured objects which can adversely affect point cloud analysis. More recently, PointNet [15] was proposed to preserve some of this geometric information by using a symmetric kernel that could enforce permutation invariance. This approach allowed a user to directly treat a point cloud as an input into a Deep Neural Network (DNN) and get a global vector representing the input point cloud as an output. This work was subsequently extended in a variety of ways to preserve local structure within a point cloud that proved to be important while performing classification [17, 26, 10]. However, the representations developed by these methods rely on an individual point's absolute position, which hinders their ability to develop algorithms that are invariant to rigid body transformations. Fig. 1, for instance, illustrates the deficiency of these methods when they are applied to perform part segmentation on views that are unseen during training.

Typically, one can address this limitation and improve the robustness of DNNs to rigid body transformations by augmenting the training set with additional examples. However, this requires additional computation and increased model capacity. For instance, during classification, a model must learn a function that maps the same object under different rigid body transformations into a similar feature in a feature space. Rather than augment the training set, other approaches have focused on developing representations that

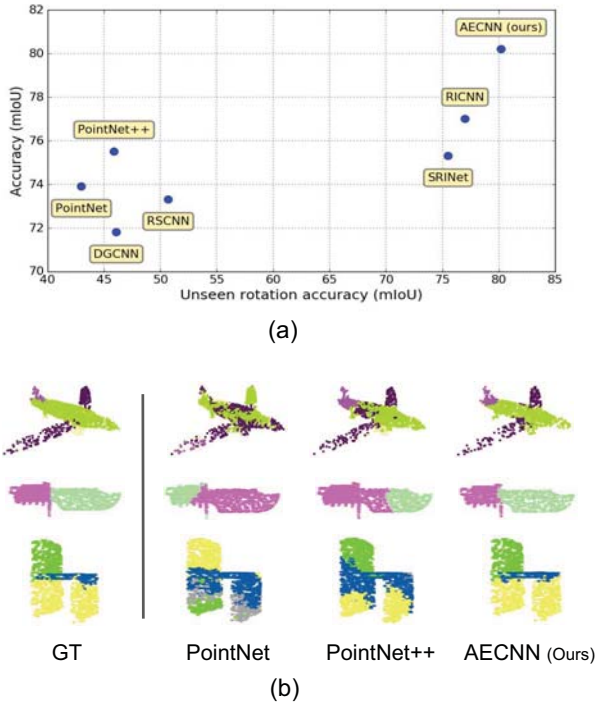


Figure 1. An comparison of the performance of state-of-the-art techniques to the method developed in this paper while performing part segmentation on rotated point clouds from the ShapeNet dataset. We report results on models of arbitrary rotation during testing while they trained with only rotation along vertical direction (x axis) and on models of arbitrary rotation during both training and testing (y axis) (subfigure a). This degradation in performance can be seen on the quality of part segmentation in unseen views (subfigure b, where different colors correspond to different part categories).

can preserve rotational symmetry by directly converting point clouds into a spherical voxel grid and then extracting rotation-equivariant features [2, 18, 30]. Unfortunately this conversion still sacrifices resolution that can adversely affect point cloud analysis. To remedy this loss of information, others have proposed to represent point clouds relative to a Local Reference Frame (LRF), which is determined by a local subset of a point cloud [1, 32]. Each point in a neighborhood of a constructed LRF is represented with respect to that LRF and a local feature is learned for each point set. Subsequently these local features are fused together to define global features. These LRF-based learned representations are invariant to rotation; however, as we illustrate in this paper, the accuracy of techniques utilizing LRF-based representation are only marginally better than those utilizing an absolute coordinate based representation for point cloud classification tasks. This is in part because the learned local features for a pair of points are not aligned before they are fused together.

To address the limitations of existing approaches, this paper proposes a novel 3D representation of point clouds that is invariant under rotation and introduces a new neural network architecture, the Aligned Edge Convolutional Neural Network (AECNN), to utilize this representation. As in prior work, we leverage the notion of LRFs to ensure that different orientations of a point cloud are mapped into the same representation. Each point in a neighborhood of a constructed LRF is represented with respect to that LRF before subsequent processing. This ensures that the model is able to learn internal geometric relationships between points rather than learning geometric relationships that are a function of the absolute coordinates of the points that may change after rotation. Our proposed AECNN architecture processes these local internal features and aligns them with local internal features drawn from other LRFs before fusing them together in a hierarchical fashion to define global features. Importantly, in contrast to prior work that utilizes a spherical coordinate system [1] or non-orthogonal basis [22], we construct a basis for the LRFs that is orthonormal. This ensures that feature alignment can be computed in a straightforward manner which makes the hierarchical fusion of local features tenable.

The contributions of this paper are three-fold: First, we propose a novel representation of points clouds that is invariant to arbitrary rotations. Second, we propose a novel alignment strategy to align neighboring features within distinct LRFs. This makes it feasible to perform feature fusion within a hierarchical network, which makes a reasonable feature fusion between local and global features. Finally, we illustrate that our propose representation is robust to rotation and achieves state-of-the-art results in both point cloud classification and segmentation tasks.

2. Related work

This section reviews the various techniques that have been applied to represent point clouds.

View-Based and Volumetric Methods. A variety of methods have represented 3D shape as a sequence of 2D images since they can leverage existing algorithms from 2D vision [5, 4, 11]. These view-based methods typically project a 3D shape onto 2D planes from different views. These different images are then processed by CNNs. Though these methods achieve good performance at classification tasks using off-the-shelf architectures and pre-trained model [3, 20, 6, 21], the projection of 3D shape onto 2D planes sacrifices 3D geometric structures that are critical during point cloud analysis.

Converting point clouds into 3D voxels can preserve some of this geometric information. The higher the resolution of quantization, the more geometric information is preserved. These converted point clouds can then leverage existing CNNs with 3D kernels [12, 16, 28]. Since points

are only sampled from the surface of objects, regular quantization can waste valuable resolution on the empty space within or outside objects. Better partitioning methods, such as KD-tree [7] and Oct-tree [25, 23] have been proposed to address this limitation. In contrast to these methods, our approach works directly with point clouds without requiring any conversion.

Point Set Learning. Point set learning methods take raw point clouds as input. The pioneering work in this area is PointNet [15], which independently transforms each point and outputs a global feature vector describing the input point cloud by aggregation using a max pooling layer. Unfortunately, PointNet is unable to learn local structure over increasing scales, which is important for high-level learning tasks. Extensions that utilize hierarchical structure [17, 8], graph network [26, 31], or relation-aware features [10] have been proposed to preserve local geometric structure during processing. Other extensions that rethink the convolution operation to better accommodate point cloud processing have also been proposed [27, 9]. However, the representation that is learned by these approaches changes when the point clouds in the training set are rotated. As a result, these representations perform poorly when utilized during classification or segmentation tasks on rotated versions of point clouds that were not included during training.

Rotation Learning. Various methods have been proposed to either learn rotation-invariant or rotation-equivariant representations. For instance, spatial transformer networks (STNs) have been proposed to learn rotation-invariant representations [15]. The STN learns a transformation matrix to align input point clouds without requiring that the alignment place the point clouds in some fixed ground-truth orientation. As a result, the transformation matrix is not guaranteed to align objects to a consistent orientation which restricts its utility. Other approaches achieve rotation invariance by relying on LRFs [1, 32, 22]. However, the difficulty of aligning the features learned with respect to LRFs, as described earlier, has limited the potential expressive capabilities of these techniques. Since designing a model that is invariant to rotation is difficult, a variety of methods have attempted to achieve rotation-equivariance using spherical convolutions [2, 18, 30]. Spherical convolutions require the spherical representation of a point cloud. Unfortunately projecting 3D point clouds into 2D sphere results in a loss of information.

3. Learning Rotation-Invariant Representation of Point Clouds

This section introduces our proposed rotation-invariant representation (RIR) of point clouds using LRFs, our proposed aligned edge convolution designed for RIR, and the proposed hierarchical network architecture for classification

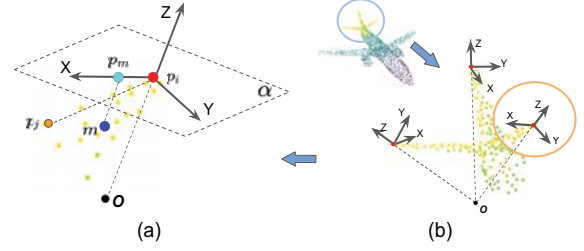


Figure 2. An illustration of aligned edge convolution and the LRFs that help define it. To construct the RIR, one takes the reference point p_i and the k -nearest points to it (subfigure a). The LRF is determined by p_i and the anchor point m which is defined as the barycenter of the k -nearest points to p_i . The coordinates of k -nearest points are described with respect to the LRF. Note that LRFs may not be aligned due to independence of the local neighborhood of points that define each LRF (subfigure b).

and segmentation tasks.

3.1. Rotation-Invariant Representation

To construct a representation that is invariant to rotation, we represent a point's coordinates relative to a LRF. Each LRF is defined using three orthonormal basis vectors and is designed to be dependent on local geometry, as is depicted in Fig. 2 (a). To define this LRF, suppose we are given a reference point p_i in a point cloud and a set of neighboring points $\{p_1, \dots, p_k\}$ to p_i in the point cloud whose coordinates are all described with respect to a coordinate system with global origin o . Note, we describe how to select these reference points in Section 3.3, and neighboring points are those within a certain radius to the reference point. Next, define an anchor point m as the barycenter of the neighboring points:

$$m = \frac{1}{k} \sum_{j=1}^k p_j, \quad (1)$$

and the plane α that is orthogonal to $\overrightarrow{op_i}$ and intersects with p_i . Using these definitions, we can define the projection of m onto the plane α :

$$\overrightarrow{p_i p_m} = \overrightarrow{om} - \frac{\overrightarrow{op_i}}{|\overrightarrow{op_i}|} \cdot \langle \overrightarrow{om}, \frac{\overrightarrow{op_i}}{|\overrightarrow{op_i}|} \rangle. \quad (2)$$

With this definition, we can construct the following coordinate axes for the LRF:

$$\vec{x} := \frac{\overrightarrow{p_i p_m}}{|\overrightarrow{p_i p_m}|}, \quad \vec{z} := \frac{\overrightarrow{op_i}}{|\overrightarrow{op_i}|}, \quad \vec{y} := \vec{z} \times \vec{x}. \quad (3)$$

Note that z axis is defined as the direction from global origin o pointing at p_i ; the x axis is defined as the direction from p_i pointing at p_m ; and the y axis is defined as the direction of cross product of z and x axis. The origin of LRF is at the reference point p_i . We assume that the global origin o is known, and in our case we use the center of point clouds.

We introduce rotation invariance by representing the set of neighboring points relative to their LRF:

$$t_j^i = (\langle p_{ij}, \vec{x} \rangle, \langle p_{ij}, \vec{y} \rangle, \langle p_{ij}, \vec{z} \rangle) \quad (4)$$

where $p_{ij} = p_j - p_i$ for each $p_j \in \{p_1, \dots, p_k\}$. Note t_j^i is the RIR for the point p_j relative to the LRF at point p_i . We then use a PointNet structure to capture the geometry within the neighboring points using the RIR:

$$f(\{t_1^i, \dots, t_k^i\}) = \text{MAX}(\{h(t_1^i), \dots, h(t_k^i)\}) \quad (5)$$

where f is a learning function, which takes a point set as input, and outputs a feature vector representing input point clouds, h is a feature transformation function and is approximated by a multi-layer perceptron (MLP). Note, the max pooling layer aggregates information.

3.2. Aligned Edge Convolution

To capture the geometric relationship between points in a point cloud, the notion of edge convolution via DGCNN has been developed [26]. To understand how edge convolution works, suppose we are given a point cloud with n points, denoted by $P = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^3$, along with F -dimensional features corresponding to each point, denoted by $X = \{x_1, \dots, x_n\} \subseteq \mathbb{R}^F$. Suppose we construct the k -nearest neighbor (NN) graph (V, E) in the feature space, where $V = 1, \dots, n$ and $E \subseteq V \times V$ are vertices and edges, then the edge convolution output at i -th vertex is given by:

$$x_i' = \text{MAX}_{j:(i,j) \in E} g(x_i, x_j - x_i) \quad (6)$$

where g is a MLP.

Note, edge convolution is essentially performing feature fusion. It fuses the global information captured by x_i with local neighborhood information captured by $x_i - x_j$. To be able to do this, x_i and x_j must be learned in the same coordinate system. Unfortunately, it is nonviable to directly apply edge convolutions to features in our case. This is because in our case, x_i and x_j are learned relative to two different LRFs, and the LRFs of x_i and x_j may not be aligned, as is shown in the Fig. 2 (b). As a result, applying edge convolution directly on our learned features may create inconsistent features.

To resolve this problem, we propose aligning x_j into the LRF of x_i before performing feature fusion. We call our approach, which is depicted in Fig. 3, Aligned Edge Convolution (Aligned EdgeConv). To construct our approach, we begin by understanding the relationship between different LRFs which can be described using a rotation R and translation T . To construct this rotation and translation, suppose the basis of the LRF for each feature is denoted by $E = \{e_1, \dots, e_n\} \subseteq \mathbb{R}^{3 \times 3}$. Then the rotation matrix and translation vector can be computed by:

$$R_j = e_i \cdot e_j^{-1} = e_i \cdot e_j^\top \quad (7)$$

$$T_j = t_j^i \quad (8)$$

where e_j is an orthogonal matrix defined in (3) and t_j is defined in (4).

R and T describe the relationship between LRFs, so we use them to transform x_j into the LRF of x_i . Though it is easy to invert a rotation and translation in 3D, extending it to the high dimensional feature space that x_j lives in would be challenging. One option to resolve this problem is to apply an approach similar to the STN proposed in the PointNet wherein one predicts a transformation matrix from R and T and applies it to x_j :

$$\hat{x}_j = \phi(R_j, T_j) \cdot x_j \quad (9)$$

where ϕ is a MLP and outputs an $F \times F$ matrix compatible with x_j . Typically a regularization term is added to the loss during training to constrain the feature transformation matrix to be close to an orthogonal matrix. Another option is to take R , T and x_j as inputs and directly output a transformed feature:

$$\hat{x}_j = \phi(R_j, T_j, x_j) \quad (10)$$

In this paper, we utilize the second option. As we show in Section 4.5, option one requires more graphics processing units (GPUs) memory and has more parameters. Therefore we update the (6) by:

$$x_i' = \text{MAX}_{j:(i,j) \in E} q(x_i, \hat{x}_j - x_i). \quad (11)$$

Similar to PointNet++ [17], we also include the RIR t_j^i in the edge convolution to maintain more information. So the aligned edge convolution is given by

$$x_i' = \text{MAX}_{j:(i,j) \in E} q(x_i, \hat{x}_j - x_i, t_j^i). \quad (12)$$

3.3. Network Architecture

Our proposed network architecture that takes raw point clouds as input and learns a representation is depicted in Fig. 3. Our architecture is inspired by techniques that perform local-to-global learning which has been successfully applied to 2D images [33] and has been shown to effectively extract contextual information. We exploit a hierarchical structure to learn both local and global representation. Our approach captures larger and larger local regions using two set abstraction (SA) blocks, proposed by PointNet++ [17], and the global features are constructed by aggregating outputs from the last SA block using max pooling.

The two SA blocks each have distinct structures; however, they share the same processing pipeline: sampling, grouping, and processing. The structure of the first SA block (SAFirst) is illustrated in the green block in Fig. 3. Given the input point clouds, we use farthest point sampling (FPS) to subsample the point cloud while preserving

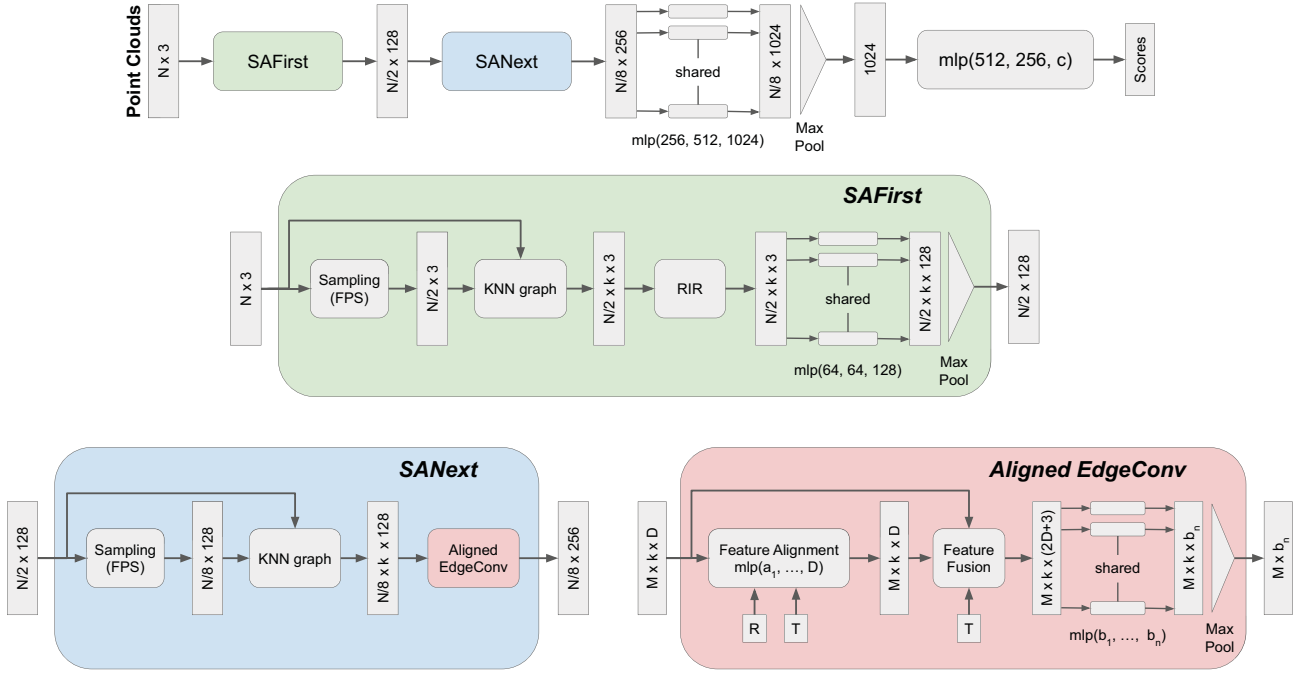


Figure 3. An illustration of the deep hierarchical architecture proposed in this paper to learn a rotationally invariant representation of point clouds for classification. The network takes N points as inputs and uses two SA blocks to hierarchically learn a representation of larger and larger regions. The final part aggregates features from the last SA block and outputs a feature vector encoding the input point set. The SAFirst block samples reference points and builds a k -NN graph in Euclidean space. It converts points into the RIR and transforms them using a shared-weights PointNet structure that outputs feature vectors encoding information for each reference point’s neighbors. The SANext block applies a similar process to the SAFirst block except for building the k -NN graph in feature space and extracting features using Aligned EdgeConv. The feature alignment module within Aligned EdgeConv aligns local features to the LRF of the reference point using the rotation matrix R and the translation vector T that are defined in the Section 3.1. Then the aligned local features are fused with the feature at the reference point along with translation vector T . Note mlp in the figure is abbreviated for MLP. The output is classification scores for c classes.

its geometric structure. The constructed points serve as the reference points during the construction of a k -NN graph. The k -NN graph is computed in Euclidean space, and it is used to generate the RIR described in Section 3.1. A shared-weights PointNet structure then processes the RIRs for each local set of points and outputs a feature vector describing the set of points near each reference point.

The structure of SANext is shown within the blue block in Fig. 3. The sampling and grouping strategy in SANext is identical to the one in SAFirst except that one quarter the number of points are selected as reference points and the k -NN graph is dynamically updated and computed in the feature space, which has been shown to be more beneficial and have larger receptive fields than a fixed graph version [26]. Then the proposed aligned edge convolution extracts features encoding larger local regions than the previous SA block. The rotation matrix R and translation vector T , which can be derived from basis and positions of LRFs, are fed into a feature alignment module to align local features to the frame of reference point. Essentially, the

SANext is the building block that can be used to iteratively capture larger and larger local regions.

Note for the segmentation task, we require a feature for each point. We adopt a similar strategy to PointNet++ [17], which propagates features from subsampled points to original points. Specifically, the interpolated features from the previous layer are concatenated with skip linked features output from SA blocks. The interpolation is done via the inverse distance weighted average based on k -NN. Importantly, the proposed feature alignment idea is also integrated in this feature propagation pipeline.

4. Experiments

This section describes how we implement the network described in Section 3.3 and how we validate its utility. First, we evaluate our method on a shape classification task (Sec 4.2) and part segmentation task (Sec 4.3). Second, we evaluate the design of our LRFs (Sec 4.4) and illustrate the effectiveness of the proposed aligned edge convolution (Sec 4.5).

4.1. Implementation details

We implement our network in PyTorch. All experiments are run on a single NVIDIA Titan-X GPU. During optimization, we use the Adam optimizer with batch size 32. Models are trained for 250 epochs. The learning rate starts with $1e-3$ and scale by 0.2 every 100 epochs.

In some experiments, we augment the dataset with arbitrary rotations. However, it is impossible to cover all rotations in 3D space. Similar to ClusterNet [1], we uniformly sample possible rotations. Each rotation is characterized by a rotation axis v and a rotation angle θ that is given by:

$$R = I + (\sin \theta)K + (1 - \cos \theta)K^2 \quad (13)$$

where K denotes the cross-product matrix for the rotation axis v which has a unit length and I is the identity matrix. In the experiments, we sample 3-dimensional vectors from a normal distribution and normalize v to be a unit vector.

We follow the approach presented in [2] to perform experiments in three different settings: 1) training and testing with rotation along the vertical direction (Y/Y), 2) training with rotation along vertical direction and testing with arbitrary rotation (Y/AR) and 3) performing arbitrary rotation during training and testing (AR/AR). The last two settings in particular are used to evaluate the generalization ability of the model under unseen rotations.

4.2. Shape Classification

One of the primary point clouds analysis tasks is to recognize the category of point clouds. This task requires a model to learn a global representation.

Dataset. We evaluate our model on ModelNet40, which is a shape classification benchmark [28]. It provides 12,311 CAD models from 40 object categories, and there are 9,843 models for training and 2,468 models for testing. We use their corresponding point clouds provided by PointNet [15], which contain 1024 points in each point clouds. During training we augment the point clouds with random scaling in the range $[-0.66, 1.5]$ and random translation in the range $[-0.2, 0.2]$ as in [7]. During testing, we perform ten voting tests while randomly sampling 1024 points and average the predictions.

Point clouds classification. We report the results of our model and compare it with other approaches in the Table 1. Three different training and testing setting are performed, which are introduced in Section 4.1. All approaches, except for the last five which are specially designed for rotation learning, perform well in the Y/Y setting, but experience a significant drop in accuracy when evaluated on unseen rotation as shown in the Y/AR setting. We conclude that these approaches only generalize well to rotations that they are trained on. However, our proposed method performs equally well across all three settings. Every evaluated approach has lower accuracy in the AR/AR setting than the

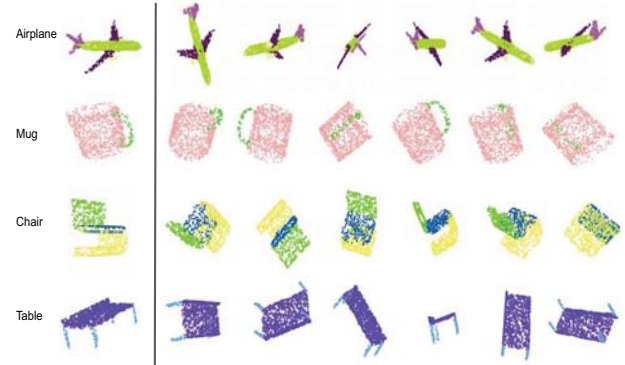


Figure 4. Qualitative results of our proposed method on part segmentation task on the ShapeNet dataset. From top to bottom, segmentation results of different categories are shown. From left to right, we show ground truth label and results when the input point clouds are arbitrarily rotated during testing. Different colors correspond to different part categories. Our model is robust to arbitrary rotations of the input point clouds.

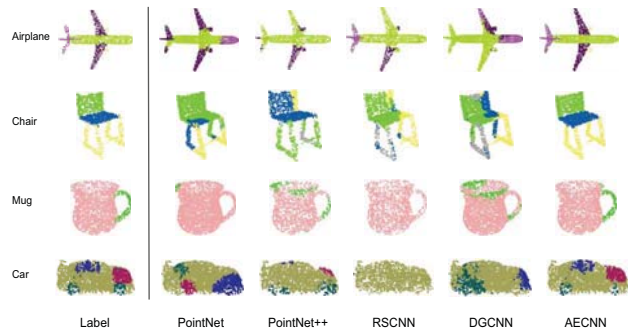


Figure 5. Qualitative results of part segmentation on ShapeNet. The Y/AR setting is adopted for all models. From top to bottom, segmentation results of different categories are shown. From left to right, we show ground truth label and the results from different approaches. Different colors correspond to different part categories. Our method achieves state-of-the-art performance while other approaches fail to generalize to unseen rotation.

Y/Y setting, except for RI-CNN, SRINet, ClusterNet, and our proposed method. This is most likely due to the difficulty of mapping identical objects in different poses to a similar feature space. This requires larger model complexity and is difficult to address with just dataset augmentation. RI-CNN (86.4%), SRINet (87.0%) and ClusterNet (87.1%) are specially designed for rotation invariance. Our model has superior performance in the setting of Y/AR and AR/AR (91.0%), which means our proposed method generalizes well to unseen rotations.

4.3. Part Segmentation

The part segmentation task requires assigning each point in a point cloud a category label. Since this is a point-wise

Method	Inputs	Input size	Y/Y	Y/AR	AR/AR
MVCNN 80x [20]	view	80×224^2	90.2	81.5	86.0
Spherical CNN [2]	voxel	2×64^2	88.9	76.9	86.9
PointNet [15]	point	1024×3	88.5	21.8	83.6
PointNet++ [17]	point	1024×3	89.3	31.7	84.9
RS-CNN [10]	point	1024×3	89.6	24.7	85.2
DG-CNN [26]	point	1024×3	91.7	31.5	88.0
RI-CNN [32]	point	1024×3	86.5	86.4	86.4
SRINet [22]	point	1024×3	87.0	87.0	87.0
ClusterNet [1]	point	1024×3	87.1	87.1	87.1
SF-CNN [18]	point	1024×3	92.3	84.8	90.1
Ours	point	1024×3	91.0	91.0	91.0

Table 1. Classification results on ModelNet40 dataset. We report the accuracy (%) in three different settings: training and testing with rotation along the vertical direction (Y/Y), training with rotation along vertical direction and testing with arbitrary rotation (Y/AR), and performing arbitrary rotation during training and testing (AR/AR). Though our model is only the third best performer in the Y/Y setting it is the top model in each of the other categories. In particular our proposed model has superior performance in the Y/AR and AR/AR, which means that it can generalize well to unseen rotations.

Method	Input size	Y/Y	Y/AR	AR/AR
PointNet [15]	2048×3	79.3	43.0	73.9
PointNet++ [17]	2048×3	80.6	45.9	75.5
DG-CNN [26]	2048×3	79.2	46.1	71.8
RS-CNN [10]	2048×3	80.0	50.7	73.3
RI-CNN [32]	2048×3	-	75.3	75.5
SRINet [22]	2048×3	77.0	77.0	77.0
Ours	2048×3	80.2	80.2	80.2

Table 2. Part segmentation results on ShapeNet dataset. Point coordinates are taken as inputs, and mIoU across all classes is reported in three different settings including Y/Y, Y/AR and AR/AR. Our model outperforms all approaches except PointNet++ in Y/Y setting. Our model has superior performance in the Y/AR and AR/AR settings, which means that it can generalize well to unseen rotations.

classification task, part segmentation is typically more challenging than classification.

Dataset. We evaluate our model on ShapeNet part dataset [29], which contains 16,881 shapes from 16 categories and annotated with 50 parts in total. We split the dataset into training, validation and test sets following the convention in PointNet++ [17]. 2048 points are randomly picked on the shape of objects. We concatenate the one-hot encoding of the object label to the last feature layer in the model as in [17]. During evaluation, mean inter-over-union (mIoU) that are averaged across all classes is reported.

3D part segmentation. We report the result of our model and compare it with other approaches in Table 2 and (subfigure a) in the Fig. 1. Results align well with performance in the classification task. In addition, our method outperforms other approaches in all three settings, except for PointNet++ which is slightly better than our proposed method in the Y/Y setting. The consistent performance of

Method	EdgeConv	AEConv1	AEConv2	AEConv3
Acc.	89.6	90.2	48.5	91.0
Para.	1.94M	2.14M	-	1.99M
FLOPs	4170M	6393M	-	4841M

Table 3. Ablation study on aligned edge convolution and comparison with original edge convolution. Accuracy, number of parameters and FLOPs per sample are reported. No further experiments were done on AEConv2 due to its poor accuracy. The number of neighbors is 48.

our method in all three settings demonstrates good generalization to unseen rotations. Qualitative results of part segments are illustrated in Fig. 4. The comparison results with other approaches in the Y/AR setting are also visualized in Fig. 5.

4.4. LRF Analysis

We achieve rotation invariance by expressing points coordinates respect to LRFs. Note that the LRF is hand-crafted rather than learned from raw data. As a result, one may be concerned about the effectiveness of the designed LRF [19]. Currently, common ways of designing LRFs use the eigenvectors of the covariance matrix of the local point set [13, 24] or rely on the surface normal as the reference axis [14]. However, computing eigenvectors for all points in some local neighborhood of points is time-consuming and estimating an accurate normal from point clouds is still challenging. In this study, we consider several alternative ways to compute LRFs and illustrate their effects on our models’ performance.

The definition of LRF is introduced in the Section 3.1. Note that given a reference point, the only variation in the definition of the LRF arises from the x axis. Because the

Searching		Grouping		# neighbors				Acc.
knn	ball	Mean	Max. D	10	16	32	48	
	✓	✓					✓	90.4
	✓		✓				✓	90.3
✓			✓				✓	90.3
✓		✓					✓	91.0
✓		✓		✓				89.6
✓		✓			✓			90.3
✓		✓				✓		90.8

Table 4. Ablation study on LRFs. Models are evaluated on ModelNet40 dataset in the Y/AR setting. Three aspects which effect LRFs are studied: searching methods, grouping ways and the number of neighbors.

Radii	(0.1, 0.2)	(0.1, 0.4)	(0.2, 0.2)	(0.2, 0.4)
Acc.	89.1	89.3	90.3	90.4

Table 5. Ablation study on radius in ball query. Models are evaluated on ModelNet40 dataset in the Y/AR setting. Model’s performance is marginally sensitive to radius if ball query is used. (r1, r2) indicates the radius in the first SABlock and the second SABlock respectively.

z axis is defined as the direction from the global origin to the reference point, and the y axis is defined by the cross product of the z and x axis. Recall that the x axis is associated with the anchor point m , which is shown in the Fig. 2. Here we investigate three different aspects that influence the determination of the anchor point: searching methods, ways of grouping the data, and the number of neighbors. We perform experiments on shape classification, and Table 4 shows the results.

We compare two searching methods: ball query and k-NN search. Ball query finds all points within a certain radius to the reference point, but only up to k points are considered in the experiments. k-NN finds a fixed k nearest points to the reference point. As is shown in the first four rows of Table 4, given the same grouping method models with k-NN search have equal or higher accuracy than ball query. However the performance of the ball query method is marginally sensitive to the size of the radius chosen, which needs to be assigned manually, as is shown in Table 5.

We study two ways of determining the anchor point: we define the anchor point as the mean of neighboring points or as the point with largest projected distance to reference point on plane α shown in the Fig. 2 (b). From rows three and four in Table 4, we conclude that anchor points computed from mean of neighbors is preferred (91.0%) over anchor points with largest projection distance (90.3%). The last four rows of Table 4 illustrate the performance of our model with different numbers of neighboring points. We find performance drops with decreasing of k and the model with 48 nearest points achieves the best performance

(91.0%). Further increasing the number of nearest points leads to extra computation burden.

4.5. Aligned EdgeConv Analysis

The proposed AECNN is specially designed to learn a rotation-invariant representation. Recall that we align features before doing feature fusion within the edge convolution and call it aligned edge convolution (AEConv). This study compares the original edge convolution which does not perform feature alignment [26] with our proposed aligned edge convolution. We also experiment with different strategies for doing alignment. The results are shown in Table 3.

We report three different strategies for alignment: transforming the source feature by a transformation matrix in the feature space (AEConv1), which is defined in (9); taking source feature x_j , LRF e_i of source point, LRF e_j of reference point and translation T as inputs to predict the aligned feature (AEConv2); taking source x_j along with rotation matrix R and translation T as inputs to predict the aligned feature (AEConv3), which is defined in (10). Our proposed feature alignment idea is verified by comparison between the first column and the last columns, where aligned edge convolution has higher accuracy (91.0 %) than the original edge convolution (89.6 %) which has no alignment process. AEConv1 (90.2 %) also outperforms edge convolution, but it loses slightly to AEConv3. Due to limited GPU memory, we need to reduce the number of learning kernels within the SAFirst block of AEConv1, so that it can be fed into a single GPU during training. Even in this case, AEConv1 still has more parameters (2.14M) to learn than AEConv3 (1.99M) and more FLOPs per sample during the test (6393M) than AEConv3 (4841M). Additionally, AEConv2 is not able to converge.

5. Conclusions

This work proposes AECNN, or the Aligned Edge Convolutional Neural Network, which addresses the challenges of learning rotationally-invariant representations for point clouds. Rotation invariance is achieved by representing points’ coordinates relative to local reference frame. The proposed AECNN architecture is designed to better extract and fuse information from local and global features. In this way, the AECNN architecture is able to generalize well to unseen rotation. Extensive experiments are performed in classification and segmentation and demonstrate effectiveness of AECNN.

6. Acknowledgement

This work was supported by a grant from the Ford Motor Company via the Ford–University of Michigan Alliance under award N028603.

References

- [1] C. Chen, G. Li, R. Xu, T. Chen, M. Wang, and L. Lin. Clusternet: Deep hierarchical cluster network with rigorously rotation-invariant representation for point cloud analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4994–5002, 2019.
- [2] C. Esteves, C. Allen-Blanchette, A. Makadia, and K. Daniilidis. Learning so (3) equivariant representations with spherical cnns. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–68, 2018.
- [3] Y. Feng, Z. Zhang, X. Zhao, R. Ji, and Y. Gao. Gvcnn: Group-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 264–272, 2018.
- [4] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [6] A. Kanazaki, Y. Matsushita, and Y. Nishida. Rotationnet: Joint object categorization and pose estimation using multi-views from unsupervised viewpoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5010–5019, 2018.
- [7] R. Klovov and V. 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.
- [8] J. Li, B. M. Chen, and G. Hee Lee. So-net: Self-organizing network for point cloud analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9397–9406, 2018.
- [9] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen. Pointcnn: Convolution on x-transformed points. In *Advances in Neural Information Processing Systems*, pages 820–830, 2018.
- [10] Y. Liu, B. Fan, S. Xiang, and C. 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.
- [11] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [12] D. Maturana and S. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928. IEEE, 2015.
- [13] A. Mian, M. Bennamoun, and R. Owens. On the repeatability and quality of keypoints for local feature-based 3d object retrieval from cluttered scenes. *International Journal of Computer Vision*, 89(2-3):348–361, 2010.
- [14] A. Petrelli and L. Di Stefano. On the repeatability of the local reference frame for partial shape matching. In *2011 International Conference on Computer Vision*, pages 2244–2251. IEEE, 2011.
- [15] C. R. Qi, H. Su, K. Mo, and L. 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.
- [16] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2016.
- [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*, pages 5099–5108, 2017.
- [18] Y. Rao, J. Lu, and J. Zhou. Spherical fractal convolutional neural networks for point cloud recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 452–460, 2019.
- [19] R. Spezialetti, S. Salti, and L. D. Stefano. Learning an effective equivariant 3d descriptor without supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6401–6410, 2019.
- [20] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015.
- [21] J.-C. Su, M. Gadelha, R. Wang, and S. Maji. A deeper look at 3d shape classifiers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.
- [22] X. Sun, Z. Lian, and J. Xiao. Srinet: Learning strictly rotation-invariant representations for point cloud classification and segmentation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 980–988, 2019.
- [23] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2088–2096, 2017.
- [24] F. Tombari, S. Salti, and L. Di Stefano. Unique signatures of histograms for local surface description. In *European conference on computer vision*, pages 356–369. Springer, 2010.
- [25] P.-S. Wang, Y. Liu, Y.-X. Guo, C.-Y. Sun, and X. Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions on Graphics (TOG)*, 36(4):72, 2017.
- [26] 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 Transactions on Graphics (TOG)*, 38(5):146, 2019.
- [27] W. Wu, Z. Qi, and L. 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.
- [28] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. 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.

- [29] L. Yi, V. G. Kim, D. Ceylan, I. Shen, M. Yan, H. Su, C. Lu, Q. Huang, A. Sheffer, L. Guibas, et al. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (TOG)*, 35(6):210, 2016.
- [30] Y. You, Y. Lou, Q. Liu, L. Ma, W. Wang, Y. Tai, and C. Lu. Prin: Pointwise rotation-invariant network. *arXiv preprint arXiv:1811.09361*, 2018.
- [31] Y. Zhang and M. 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.
- [32] Z. Zhang, B.-S. Hua, D. W. Rosen, and S.-K. Yeung. Rotation invariant convolutions for 3d point clouds deep learning. In *2019 International Conference on 3D Vision (3DV)*, pages 204–213. IEEE, 2019.
- [33] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.