Unsupervised Deep Shape Descriptor with Point Distribution Learning

Yi Shi* Mengchen Xu* Shuaihang Yuan Yi Fang†
NYU Multimedia and Visual Computing Lab
New York University Abu Dhabi, Abu Dhabi, UAE
New York University, New York, USA
{yishi,jl8566,sy2366,yfang}@nyu.edu

Abstract

Deep learning models have achieved great success in supervised shape descriptor learning for 3D shape retrieval, classification, and correspondence. However, the unsupervised shape descriptor calculated via deep learning is less studied than that of supervised ones due to the design challenges of unsupervised neural network architecture. This paper proposes a novel probabilistic framework for the learning of unsupervised deep shape descriptors with point distribution learning. In our approach, we firstly associate each point with a Gaussian, and the point clouds are modeled as the distribution of the points. We then use deep neural networks (DNNs) to model a maximum likelihood estimation process that is traditionally solved with an iterative Expectation-Maximization (EM) process. Our key novelty is that "training" these DNNs with unsupervised self-correspondence L2 distance loss will elegantly reveal the statistically significant deep shape descriptor representation for the distribution of the point clouds. We have conducted experiments over various 3D datasets. Qualitative and quantitative comparisons demonstrate that our proposed method achieves superior classification performance over existing unsupervised 3D shape descriptors. In addition, we verified the following attractive properties of our shape descriptor through experiments: multi-scale shape representation, robustness to shape rotation, and robustness to noise.

1. Introduction

With recent advancements in range sensors and imaging technologies, 3D geometric data has been applied in a variety of applications[2, 32, 22, 46, 45]. It is therefore of great interest to develop methods that can automatically analyze a large amount of 3D geometric data for different tasks (e.g. 3D object recognition, classification, correspondence and retrieval) [4, 5, 3, 11, 40, 39, 37, 18, 38, 10, 9, 25]. To that end, a lot of efforts have been made, many of which focused on building robust 3D shape representations. However, the irregular property and the structural variation of 3D objects (i.e. 3D human models with different poses, 3D car models with various design patterns) pose great challenges on the task of learning a high-quality 3D shape descriptor.

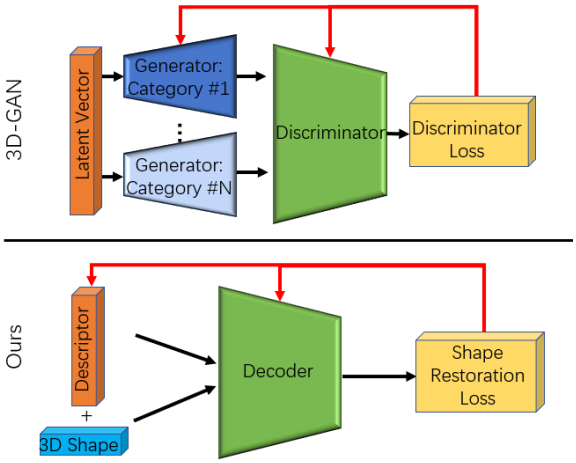

Figure 1. The illustration of our proposed unsupervised descriptor learning method with 3D-GAN [37]. Comparing with 3D-GAN, Our proposed method does not require the generator and the discriminator while our framework extracts the shape descriptor only using a decoder.

The promising performance of Deep Neural Networks (DNNs) in dealing with 2D images motivates 3D computer vision researchers to transform the 3D geometric data to Voxel grids in a regular format so that the 3D data can be fed to a deep net architecture [38] for further processing. While the extension of the deep learning to volumetric shapes is conceptually simple, the computational cost of 3D convolution severely limits the storage and computational efficiency in processing 3D geometric data for object recognition [30].

*Equal contribution
†Corresponding author

Given that the most common approaches for descriptor extraction are conducted in supervised learning, generative models such as 3DGAN [37] treads an uncharted territory where 3D shape descriptors can be generated in an unsupervised manner. Despite its good performance over various tasks, it still relies on 3D convolutional layers where the geometric information loss is inevitable given the necessity of the volumetric transformation. It also suffers from the restriction of its encoder and discriminator structure, since they are only specialized in feature extraction when faced with the shape categories encountered during training. A drop in the performance when processing out-of-category data is thus expected.

In this paper, we propose a novel zero-shot unsupervised approach for learning an instance-level 3D shape descriptor. Firstly, we claim that a point cloud instance can be described by probability distributions where each point is represented by a Gaussian. With the aim of obtaining the distribution of points given a shape instance, in each epoch, we synthesize a new point cloud instance corresponding to the input point cloud. To elaborate on the synthesizing process, each point in the new point cloud is sampled from the 3D space of a multivariate Gaussian distribution defined with a fixed standard deviation σ around the original point. Here we denote this process of point cloud synthesis as 3D Gaussian sampling. An encoder-free network is then leveraged to model a Maximum Likelihood Estimation process where the network learns to predict the parameters of the distribution while a shape descriptor is optimized to describe the geometric information. The L2 distance between the original 3D point cloud instance and the predicted instance will then be calculated to train the general decoder and shape descriptors for all training data. In order to generate the shape descriptor for any point cloud data, the exact same operation during the decoder training phase will be performed, except that the decoder will be fixed using the weights learned during the training phase.

Our unsupervised encoder-free model is more versatile compared to other generative structures in two aspects: 1) It avoids the design of specific 3D feature encoder for irregular non-grid point cloud; 2) It frees our model from the limitation of the fixed encoder weights optimized for categories encountered during training. It enhances feature learning for unseen out-of-category data. Furthermore, with our auto-decoder network (mostly MLP structure), compared to complex GAN structures in [19, 37], our approach is able to generate the 3D shape descriptor efficiently in an unsupervised manner. Our contributions are summarized as followed:

- A probabilistic representation that models the point clouds using Gaussian distributions. The shape descriptor of a point cloud instance can be revealed by solving the parameters of the distributions via a DNN.
- An unsupervised shape descriptor learning mechanism from which shape descriptors that are robust to rotation and noise can be generated.
- A novel multi-scale shape descriptor fusion technique that is able to represent an instance in a coarse-to-fine manner which enhances the performance of our descriptor in various tasks.

2. Related Works

2.1. Hand-Crafted 3D shape Descriptors

3D shape descriptor is a succinct and compact representation of 3D objects that capture the geometric essence of a 3D object. Some existing shape descriptors have been developed to describe the 3D objects [20, 38, 30]. The earlier D2 shape distribution, statistical moments, Fourier descriptor, Light Field Descriptor, Eigenvalue Descriptor have been proposed to describe the 3D shape, particularly for rigid 3D objects. The Spin Image [44] was developed based on the dense collection of 3D points and surface normals. There are also feature histogram [44] and signatures of histogram shape descriptors developed based on the distribution of a type of statistical geometric properties. The efforts on robust 3D shape features are further developed by heat diffusion geometry. A global shape descriptor, named temperature distribution (TD) descriptor, is developed based on HKS information at a single scale to represent the entire shape [14]. Hand-crafted shape descriptors are often not robust enough to deal with structural variations and incompleteness present in 3D real-world models and are often not able to be generalized to data of different modality.

2.2. Shape feature learning

The bag-of-features (BOF) is first introduced to learn to extract a frequency histogram of geometric words for shape retrieval [12, 13, 21]. To learn global features, [15] adopted auto-encoder with the distribution of HKS learns a deformation-invariant shape descriptor. Recent development in deep learning motivates researchers to learn a 3D shape descriptor from a large-scale dataset using deep neural networks. However, to feed the 3D geometric data to neural networks, the 3D geometric data are often transformed into 3D Voxel grids or a collection of 2D projection images from different views.

The volumetric representation plays an important role in the computer graphics community since the 1980s. It provides a uniform, simple and robust description to synthetic and measured objects and founds the basis of volume graphics [23]. In other words, a voxel is an extension of a pixel, and the binary volume is an extension of binary image. Recently, many researchers begin to develop 3D CNN on volumetric shapes. [38] voxelized the 3D shape into 3D grids

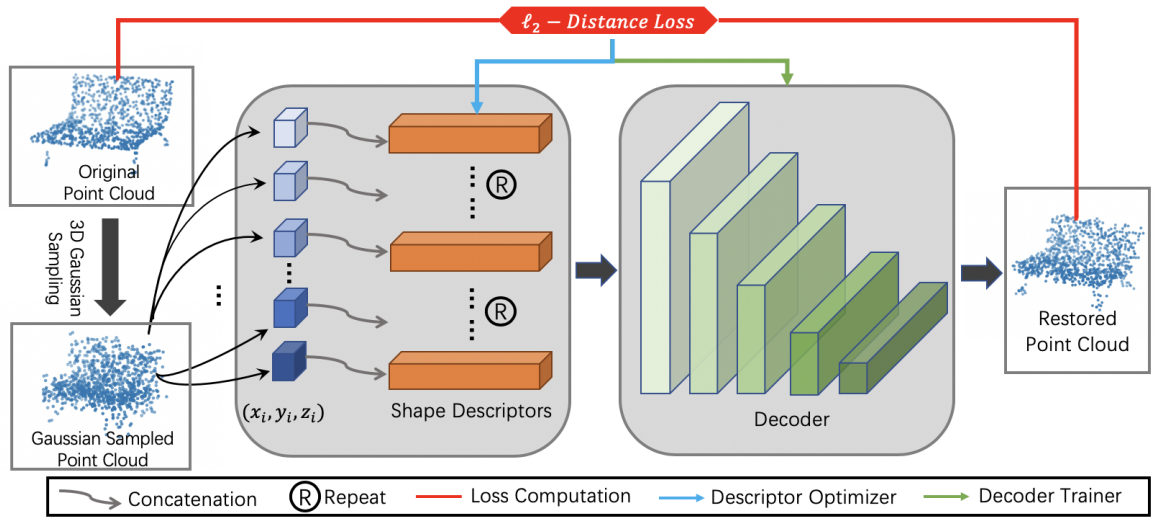

Figure 2. The training process. We concatenate a randomly initialized vector z to each point of the sampled instance shape. The L2 point distance loss between the decoded point set and the original point set will be calculated. During the decoder training phase, the loss will be back-propagated to update the shape descriptors and decoder simultaneously. During the descriptor generation phase, the loss will only be used to update the shape descriptors.

and train a generative model for 3D shape recognition using convolutional deep belief networks. Similarly, [30] proposed a real-time 3D supervised learning architecture on volumetric 3D shapes. Apart from supervised CNN, [37] generate 3D objects from a probabilistic space by leveraging advances in volumetric CNNs and GANs, and the unsupervised features can be widely used in 3D object recognition. [33] proposed a 3D convolutional auto-encoder for recognizing 3D shape.

Apart from the direct 3D representation, the 3D shapes can also be projected to 2D space. [34] proposed a multi-view CNN for 3D shape recognition by using CNN to extract visual features from images with different views and employing max-pooling across views to learn one compact shape descriptor. The LFD [8] extract features from the light fields rendered from cameras on a sphere exhaustively to improve the robustness against rotations. [6] proposed a coding framework for constructing a compact descriptor based on a set of 2D views in the format of depth buffer rendered from each 3D object.

2.3. Variational Auto-encoder

A Variational Auto-encoder (VAE) [24] is another popular generative framework that learns a probability distribution $P(X)$ from a set of observations $\mathcal{X}$. Suppose we have a random variable X which represents a distribution. The VAE maps the distribution using an encoder with a prior distribution of z $P_\phi(z)$, and a decoder $P_\theta(X|z)$ tries to capture the distribution of X given z. The encoder and decoder are trained simultaneously to maximize a lower bound on the log-likelihood of $\mathcal{X}$. After training, Both the encoder mapping $Q_\phi(z|X)$ and the decoder mapping $P_\theta(X|z)$ are acquired. Instance generation and shape descriptor acquisition can then be achieved with the trained models.

2.4. Adversial Networks based methods

[17, 29] have shown the effectiveness of Generative Adversarial Networks (GAN), where a pair of neural networks jointly learn together by pursuing competing goals, the generator learns the mapping from a latent space to data distribution, while the discriminator learns to distinguish from ground truth data and the generated data. [37] extends the usage of GAN from 2D image to voxelized 3D grid. Its category specialized generator learns a mapping from a low-dimensional probabilistic space to 3D shape while its discriminator provides a powerful 3D shape descriptor. It can generate shape descriptors without supervision.

3. Methods

3.1. Problem Formulation

Consider M point cloud instances in the Euclidean space $S = \{X_i\}_{i=1,2,...M}$ where each point cloud instance is a cluster of N points $X_i = \{x_{ji}\}_{j=1,2,...N}$ and $x_{ji} \in \mathbb{R}^3$.

3.2. Point Distribution Representation

In most applications, point clouds are either acquired via scanning sensors or sampled on the surface of a mesh file. Point clouds are thus born with a noisy and random nature.

In pursue of a robust representation, it's common to represent a point cloud with probability distributions.

Aimed to simplify our probability distribution representation in a succinct yet expressive form, the distribution for a point $x_{ji} \in \mathbb{R}^3$ can be viewed as a multivariate Gaussian distribution parameterized with mean $\mu_{ji} \in \mathbb{R}^3$ that equals to the point coordinate and a constant covariance $\sigma^2 I$, where $I \in \mathbb{R}^{3\times3}$ is an identity matrix. The point distribution is thus symmetric about the corresponding point coordinate and affected by a refinement parameter σ. Its effect will be covered in Section 3.4.

To describe the distribution of point x_{ji} in an instance X_i, a new point x'_{ji} is synthesized by sampling from the Gaussian distribution centered around the corresponding original point during each epoch. After a number of epochs, a point distribution is expected to be well-presented given adequate samples have been observed. This Gaussian sampling procedure can also be viewed as an effective data augmentation technique which increases the diversity of data significantly. It is one of the crucial factors that empower our zero-shot learning model when faced with categories with few instances. Noted that each point distribution in X_i is independent, the shape distribution of a sampled instance X'_i can then be modeled naturally as the product of point distributions.

$$P(X'_i|\theta) = \prod_{j=0}^{N} p(x'_{ji}|\theta) \tag{1}$$

in which $p(x_{ji}|\theta) = p(x_{ji}|\mu_{ji}) = \frac{1}{(2\pi\sigma^2)^3} e^{-\frac{1}{2}\left\|\frac{x-\mu_{ji}}{\sigma}\right\|^2}$. Considered that each point contributes to the geometric information of the entire point cloud instance, in order to evaluate a point cloud, we align the independent point distributions with the likelihood of an instance. Since our goal is to acquire a descriptor for an instance, we hope adequate geometric information can be learned by solving the parameters of distributions. Intuitively, such a problem is solved in a direct approach with Maximum Likelihood Estimation (MLE).

$$\begin{aligned} \theta^{\mathbf{optimal}} &= \arg\max_{\theta} P(X'_i|\theta) \\ &= \arg\max_{\theta} \sum_{j=0}^{N} \log(p(x'_{ji}|\theta)) \end{aligned} \tag{2}$$

where θ stands for distribution parameters that include all mean vectors of point distributions. Traditionally, an iterative optimization method, the Expectation-Maximization (EM) algorithm is used here. However, the accuracy and time cost of EM is hard to guarantee given its iterative nature. More importantly, we are unable to connect point distributions with a descriptor latent distribution from which the shape descriptors can be obtained. In contrast, we leverage DNNs to model such an MLE process in an unsupervised manner.

3.3. Maximum Likelihood Estimation with Deep Neural Networks

Unlike common approaches in recent proposed generative models [24, 37] that establishes a mapping from a possibility distribution to another distribution via encoding and decoding a hidden latent vector, our approach directly optimizes a hidden latent vector determines the mapping from a distribution to another distribution without the encoding-decoding process. It frees us from the restriction of encoder weights that are specialized with the training data, thus guaranteeing a better generalization ability on unseen data and categories.

In essence, our model reveals the deep shape descriptor parameters through optimizing a latent encoding for restoration from distribution described by Gaussian sampled points to the original point. In the proposed model, a point cloud instance can be thought of as a distribution of distributions. More specifically, the distributions of points as a whole describe the distribution of shape, and the shape descriptor of the instance itself is thought of as a sample from a distribution of all shapes descriptors. For an original shape X_i, a synthesized shape X'_i is created by Gaussian sampling in an iteration. The point coordinates in X'_i are $\{x'_{ji}\}_{j=1,2,...N}$. Each individual shape encountered is assigned a descriptor Z_i. The prior distribution $p(Z_i)$ of the corresponding shape descriptor Z_i is set as a Gaussian with zero-mean. The posterior distribution of a descriptor Z_i given a synthesized point cloud instance X'_i with the point distribution parameters θ can be formulated as:

$$P(Z_i|X_i{}', \theta) = p(Z_i) \prod_{j=0}^{N} p(x'_{ji}|Z_i, \theta) \tag{3}$$

where $p(x'_{ji}|Z_i, \theta)$ describes the independent point probability distribution of a Gaussian sample x'_{ji} given the shape descriptor and the original coordinate. As a large number of synthesized point generated, $P(Z_i|X'_i, \theta) \to P(Z_i|X_i, \theta)$ as $X'_i \to X_i$, thus allowing $P(Z_i|X'_i, \theta)$ to be approximated to $P(Z_i|X_i, \theta)$.

From the perspective of the pipeline, we split the descriptor learning into two major phases, decoder training phase and descriptor optimizing phase. Shape descriptors are treated as learnable parameters. Each descriptor is initialized from a zero-mean Gaussian and concatenated with point coordinates. The decoder $\mathcal{D}$ with weights θ then receives the latent-enhanced coordinates and guesses the mean of the corresponding point distribution as shown in Figure 2. An unsupervised self-correspondence loss will be calculated and optimize the learnable variables in the

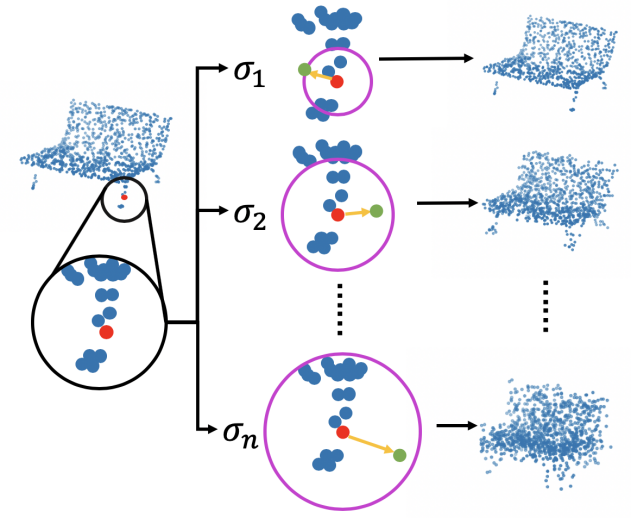

Figure 3. Instances are sampled with standard deviation of different scales in purpose of generating coarse-to-fine descriptors.

model. During the decoder training phase, we aim to optimize the weights θ of the decoder $\mathcal{D}$.

$$\theta^{\mathbf{optimal}}, \mathbf{Z}^{\mathbf{optimal}} = \arg\min_{\theta, \mathbf{Z}} \sum_{i=0}^{M} \sum_{j=0}^{N} \mathcal{L}(\mathcal{D}_\theta(x'_{ji}, Z_i), x_{ji}) \tag{4}$$

where $\mathbf{Z} = \{Z_i\}_{i=1,\ldots,M}$. During the shape descriptor optimizing phase, all previously learned shape descriptors are discarded and the learned decoder weights are fixed. In each iteration, a single instance performs the exact same sampling sequence as in the decoder learning phase. Only the corresponding shape descriptor of the instance is optimized.

$$\mathbf{Z}^{\mathbf{optimal}} = \arg\min_{\mathbf{Z}} \sum_{i=0}^{M} \sum_{j=0}^{N} \mathcal{L}(\mathcal{D}_\theta(x'_{ji}, Z_i), x_{ji}) \tag{5}$$

Euclidean Distance is chosen as our loss function in both phase since it is a more strict evaluation than other similarity metrics. For $A = \{a_i\}_{i=0}^{N}$ and $B = \{b_i\}_{i=0}^{N}$,

$$\mathcal{L}(\mathcal{A}, \mathcal{B}) = \sum_{j=0}^{N} ||a_j - b_j||_2 \tag{6}$$

3.4. Multi-scale Feature Descriptor

In our setting, the only variable that affects the sampling process is the standard deviation σ. The larger σ is, the more likely the synthesized points are far away from the original location. To avoid disrupting the geometric information in an instance or sampling similar point multiple times, we determine a suitable global σ value for a dataset by calculating the mean of the distance between a point and its tenth-nearest-neighbor. We demonstrate in Table 1 that our simple approach improves on state-of-the-art results.

However, it is still difficult for a network to learn shape features that contain information from local to global scales using samples synthesized by a single fixed σ. If σ is set too small compared to the average distance between points, the point distributions are "squeezed" into the centroids. A negligible loss can be achieved by simply predicting the input point coordinates as the mean of point distribution. In consequence, both the decoder and the shape descriptor are not effectively optimized. On the other hand, a σ that is set too high disrupts the shape distribution. The delicate geometric details where the distances between points are smaller than σ are utterly destroyed. The synthesized instances from similar categories will be indistinguishable. In order to overcome the obstacle, a multi-scale feature fusion is conducted here to take advantage of shape descriptors learned under distinctive σ settings to represent coarse-to-fine shape information.

Instead of generating only one descriptor using synthesized data from Gaussian with a fixed standard deviation, we calculated N sets of shape descriptors $\{Z_i\}_{i=1}^{M}$, each generated by data synthesized with a different standard deviation σ. For each point in the point cloud, the distance between itself and its tenth-nearest-neighbor is computed and a histogram is plotted. Empirically, three different sets of standard deviation $\{\sigma_i\}_{i=1}^{3}$ are chosen where σ_1 equals to the mean of the lower 20% tenth-neighbor distance, σ_2 equals to the global mean, and σ_1 equals to the average distance of the upper 80%. Finally, we concatenate the descriptors of an instance as $[Z_1,\ Z_2,\ Z_3]$ to form a new shape descriptor for the corresponding point cloud set. As proved in the experiment, it achieves the best performance in the task of classification than other configurations.

4. Experiments

In this section, we first introduce our experimental settings including datasets and detailed network architectures. Then we prove that our proposed shape descriptors can be applied to 3D shape recognition tasks, and verify the improvements made by the multi-scale feature representation. In addition, we demonstrate the reconstruction results and explore the effects of different σ values on the outcome of the previous experiments. At last, we conduct quantitative experiments to validate the extra properties of our descriptors.

4.1. Experiment settings

Dataset: We mainly use shapes from two 3D datasets in our experiments [7, 43]. Our general decoder is trained only on the seven major categories in ShapeNet. The main classification task is performed on the shape descriptors

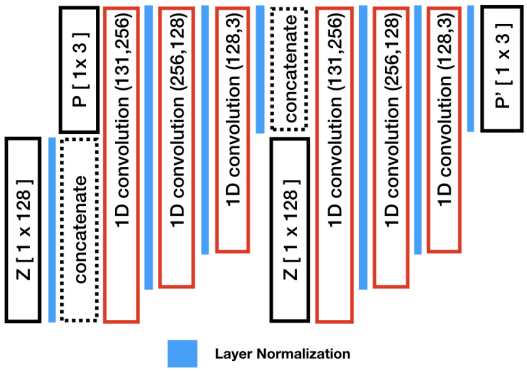

Figure 4. Detailed structure of our decoder model

generated on the entire ModelNet40. A 3D point cloud reconstruction experiment is tested on ShapeNet. We also prove the shape descriptors generated by our model are robust to rotation and noise by carrying out experiments on ShapeNet.

Architecture: We set the size of all shape descriptors as 128. As shown in Figure 4, a randomly initialized shape descriptor z is concatenated with the coordinate of each point and sent into a decoder network. The network architecture of the decoder consists of multiple 1D convolutions and linear layers. Between each three 1D convolution layers, the output is concatenated again with the shape descriptor z. The batch size during training is set as 64. During the descriptor optimization phase, each descriptor should only be conditioned by the loss of its own rather than an averaged loss within a batch, which inevitably leads to a batch size of "one". Layer Normalization layers are thus chosen over Batch Normalization in our network structure. Therefore, only instance-level information is leveraged in our process of descriptor generation and it allows our model to generate per-instance shape descriptors in an online manner. In comparison, recent methods such as [27] require trans-instance comparisons and the performance relies deeply on the size of the target dataset. During the decoder training phase, the model is trained for 200 epochs. During the descriptor optimization phase, each descriptor is trained for 100 epochs.

4.2. Unsupervised Shape Descriptor Evaluation

Extracting shape descriptors with supervised learning is the common approach that yields excellent performance. Our model, however, learns a shape descriptor for a 3D shape in an unsupervised manner. After obtaining shape descriptors, we evaluate them by performing classification tasks. Following the idea of proving out-of-category capability as [37], the decoder is trained on the subset composed of seven major categories from ShapeNet. For an accurate comparison, we also follow the same evaluation setting on the entire ModelNet40 benchmark as in [37]. We then evaluate the generated feature descriptors by training an MLP as a classifier. There are unsupervised approaches [19, 1, 42] where the networks are trained using the entire 55 categories from the ShapeNet55 that contains 57,000 shapes in total. Due to different data settings in the experiment, a direct comparison based on accuracy might not be most appropriate.

Table 1 shows the performance comparison between our proposed approach and the state-of-the-art supervised and unsupervised methods. Our unsupervised shape representation outperforms the 3DGAN by scoring 84.7% on ModelNet40. Considered that most of the categories from ModelNet40 are completely novel to our model, it demonstrates a great out-of-category generalization capability.

Supervision	Method	Accuracy
Supervised	MVCNN[35]	90.1%
	PointNet++[31]	90.7%
	PointCNN[26]	92.2%
	DGCNN[36]	92.2%
	Point2Sequence[28]	92.6%
Unsupervised	T-L Network[16]	74.4%
	VConv-DAE[33]	75.5%
	3D-DescripNet[41]	83.8%
	3D-GAN[37]	83.3%
	Ours	84.7%

Table 1. Classification evaluation on ModelNet40.

4.3. Multi-scaled Representation

To verify the effectiveness of our coarse-to-fine multi-scale descriptor, we compare its performance in classification with the shape descriptors optimized with a single σ value. As is described in Section 3.4, the multi-scale descriptor is obtained by concatenating multiple sets of shape descriptors each trained with a different σ value.

Experiment Setup: In this section, we conduct the evaluation within ShapeNet. However, only data belongs to the seven major categories are used during training. Three sets of shape descriptors and their corresponding decoders are trained with different σ values. To determine suitable σ values, we evaluate the unit density of point clouds sampled with 2048 points from the shape instances. We normalize the point clouds by feature scaling and calculate the distance from each point to its tenth nearest neighbor. Given the distance distribution calculated within the dataset, we select three σ values 0.04, 0.08, and 0.12, with $\sigma = 0.08$ as the average tenth-nearest-neighbor distance of all instances. For each instance, the shape descriptors

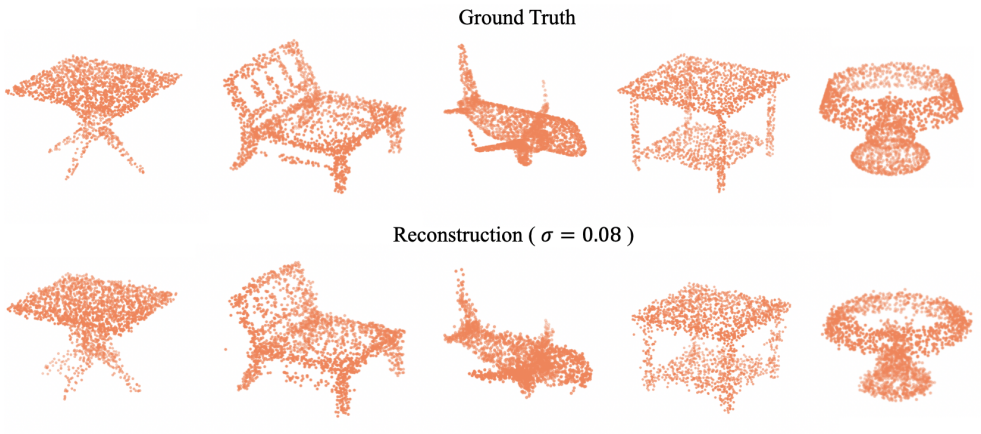

Figure 5. Reconstruction results of shape descriptors with $\sigma = 0.08$ on instances from ShapeNet

optimized using different σ settings are then concatenated.

Results: As is shown in Table 2, the multi-scale shape descriptor Z_{concat} outperforms all of the single-valued shape descriptors. It proves that our shape descriptor can be enhanced by multi-scale feature fusion with different σ values. It is also worth noticing that Z_2 with $\sigma = 0.08$ achieves the highest accuracy, which indicates that our approach of selecting the optimal σ value fits our feature learning model well.

Shape Descriptor	σ	Accuracy
Z_1	0.04	91.4%
Z_2	0.08	**94.3%**
Z_3	0.12	92.9%
Z_{concat}	multi-scale	**96.2%**

Table 2. The classification evaluation on ShapeNet. The shape descriptors Z_1, Z_2, Z_3 trained with single σ values are outperformed by a multi-scale descriptor Z_{concat}

4.4. Reconstruction

Our approach allows reconstructing point cloud from its sampled instance with the corresponding optimized shape descriptor. We examine the reconstruction results under different σ settings and observe the results on instances from different categories.

Experiment Setup: The reconstruction experiments are conducted under the same data settings explained in Section 4.3, we perform multiple experiments using instances sampled in normal distributions with standard deviation σ of 0.04 and 0.08.

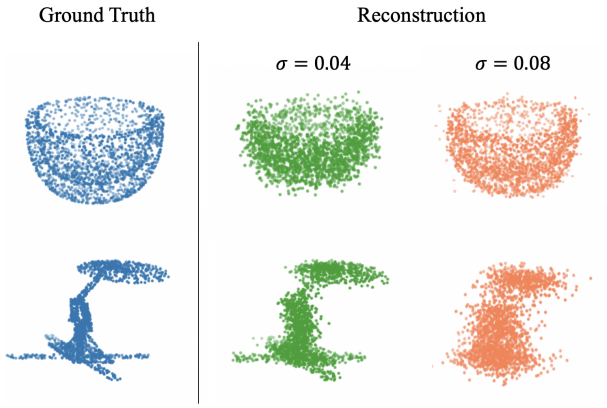

Figure 6. Comparison between the shape reconstruction results with $\sigma = 0.04$ and $\sigma = 0.08$ on different categories. The bowl is a representative of point clouds with lower unit density, while the lamp represents point clouds with higher unit density.

Results: As shown in Figure 5, our shape descriptor is able to achieve great reconstruction performance on out-of-category instances if σ is set to 0.08. In Figure 6, items with complicated parts are reconstructed well under a relative small σ setting but fails with a large σ value. Interestingly, a bowl with few details does not have a satisfying reconstruction using the same decoder with a small σ while that with a large σ is more successful. The relative sensitivity of reconstruction qualities with respect to different choices of σ is the main reason that we introduce the coarse-to-fine multi-scale descriptor in Section 3.4.

4.5. Robustness to Rotations

Given that the proposed shape descriptors are learned from the mapping from the distribution of points to the corresponding point origin, our approach is expected to be more robust to rotations than extracting geometric

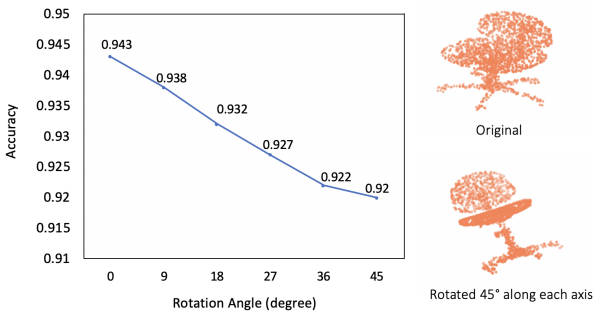

Figure 7. Rotation-Accuracy plot

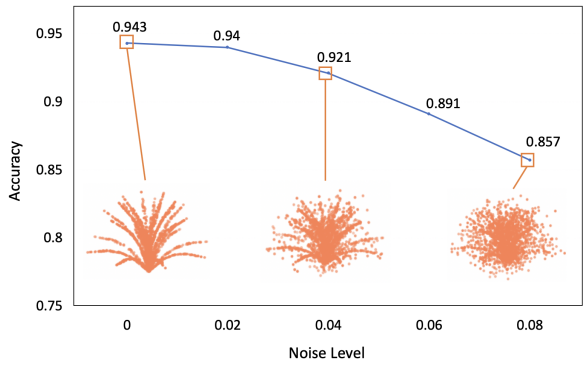

Figure 8. Noise-Accuracy plot, noise level stands for the standard deviation of a Gaussian distribution which is used to generate the translation vector to be applied on the original instances

information directly from coordinates. This test intends to demonstrate the prowess of our model for classifying rotated shapes at different angles.

Experiment Setup: We apply rotation along each axis from 0 to 45 degrees and do several experiments at an interval of 9 degrees. The rotated shape with 45 degrees along each axis is shown in Figure 7. Descriptors of the rotated instances are generated using the same decoder trained with the seven major categories in Section 4.2. For each level of rotation, we leveraged the decoder that has been trained on the seven major categories, and repeat the evaluation on the rotated shapes from the test set.

Results: The quantitative experimental results are plotted in Figure 7. As shown in the evaluation, our approach achieves an impressive level of performance when the rotation angle is within 20 degrees since the accuracy maintains above 93.0%. When the shapes are rotated with 45 degrees, the accuracy drops by 2.3% from the best performance with no rotation applied. Since rotation with 45 degrees at each axis has significantly changed the orientation of shapes, we can conclude that our model is robust to rotations.

4.6. Resistance to Noise

A good shape descriptor should also maintain consistency in classification performance over shapes perturbed with a certain amount of noise. In this part, we conduct experiments on noisy point clouds and assess the noise resistance quality by classifying shapes at different noise levels. Our decoder is capable of overcoming the noise affected data given enough samples with small random perturbation would not alter the distribution in a significant way.

Experiment Setup: Descriptors corresponding to the noise-perturbed test instances are generated using the same decoder trained with the seven major categories in Section 4.2. During the process, we perturb the point clouds by various levels of noise. For each point in the instance, we randomly move it by a translation vector that follows a Gaussian distribution with the mean of its coordinate and the standard deviation of 0.00, 0.02, 0.04, 0.06, and 0.08. We define the noise level as the standard deviation of the Gaussian distribution.

Results: The quantitative experimental results are plotted in Figure 8. The performance of our descriptors is relatively consistent when the standard deviation of the Gaussian that represents the noise level is less than 0.04. With a greater level of noise applied to the shape, the categorical information is severely damaged and hardly recognizable even by humans. At this point, the classification becomes less meaningful, so we do not apply σ values greater than 0.08. As shown in Figure 8, we can conclude that our shape descriptor is reasonably resistant to Gaussian noise perturbations.

5. Conclusion

In this paper, we introduced an alternative unsupervised method for calculating instance-level shape descriptors through modeling the Maximum Likelihood Estimation process with an encoder-free network. We proposed a multi-scale descriptor fusion technique that can represent an instance in a coarse-to-fine manner which enhances the overall performance. In addition, we proved with experiments that our descriptors have outstanding properties of noise resistance and rotation invariance.

6. Acknowledgement

We would like to thank all reviewers for their insightful suggestions and efforts towards improving our manuscript. This work is partially supported by NYUAD Research Enhancement Fund (No.RE132).

References

[1] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. J. Guibas. Learning representations and generative models for 3d point clouds. *arXiv preprint arXiv:1707.02392v3*, 2018. 6

[2] A. Albarelli, E. Rodolà, F. Bergamasco, and A. Torsello. A non-cooperative game for 3d object recognition in cluttered scenes. In *International Conference on 3d Imaging, Modeling, Processing, Visualization and Transmission*, pages 252–259, 2011. 1

[3] S. Bai, X. Bai, Q. Tian, and L. J. Latecki. Regularized diffusion process on bidirectional context for object retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2018. 1

[4] S. Bai, S. Sun, X. Bai, Z. Zhang, and Q. Tian. Improving context-sensitive similarity via smooth neighborhood for object retrieval. *Pattern Recognition*, 83:353 – 364, 2018. 1

[5] S. Bai, Z. Zhou, J. Wang, X. Bai, L. J. Latecki, and Q. Tian. Automatic ensemble diffusion for 3d shape and image retrieval. *IEEE Transactions on Image Processing*, PP(99):1–1. 1

[6] X. Bai, S. Bai, Z. Zhu, and L. J. Latecki. 3d shape matching via two layer coding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(12):2361–2373, 2015. 3

[7] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], 2015. 5

[8] D. Chen, X. Tian, Y. Shen, and M. Ouhyoung. On visual similarity based 3d model retrieval. In *Computer Graphics Forum*, pages 223–232, 2003. 3

[9] J. Chen, L. Wang, L. Xiang, and F. Yi. Arbicon-net: Arbitrary continuous geometric transformation networks for image registration. *Thirty-third Conference on Neural Information Processing Systems (NeurIPS)*, December 2019. 1

[10] J. Chen and F. Yi. Deep cross-modality adaptation via semantics preserving adversarial learning for sketch-based 3d shape retrieval. *European Conference on Computer Vision (ECCV)*, 2018. 1

[11] G. Dai, J. Xie, and Y. Fang. Deep correlated holistic metric learning for sketch-based 3d shape retrieval. *IEEE Transactions on Image Processing (TIP)*, 2018. 1

[12] M. Eitz, J. Hays, and M. Alexa. How do humans sketch objects? *ACM Trans. Graph.*, 31(4):44–1, 2012. 2

[13] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE transactions on visualization and computer graphics*, 17(11):1624–1636, 2011. 2

[14] Y. Fang, M. Sun, and K. Ramani. Temperature distribution descriptor for robust 3d shape retrieval. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 9–16. IEEE, 2011. 2

[15] Y. Fang, J. Xie, G. Dai, M. Wang, F. Zhu, T. Xu, and E. Wong. 3d deep shape descriptor. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2319–2328, 2015. 2

[16] R. Girdhar, D. F. Fouhey, M. Rodriguez, and A. Gupta. Learning a predictable and generative vector representation for objects. *CoRR*, abs/1603.08637, 2016. 6

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3

[18] Z. H. M.-Y. L. S. B. B. H. Guandao Yang, Xun Huang. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Computer Vision and Pattern Recognition*, pages 252–259, 2019. 1

[19] Z. Han, M. Shang, Y. Liu, and M. Zwicker. View inter-prediction gan: Unsupervised representation learning for 3d shapes by learning global shape memories to support local view predictions. *AAAI*, 2019. 2, 6

[20] V. Hegde and R. Zadeh. Fusionnet: 3d object classification using multiple data representations. 2016. 2

[21] R. Hu and J. Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Computer Vision and Image Understanding*, 117(7):790–806, 2013. 2

[22] S. Katz, G. Leifman, and A. Tal. Mesh segmentation using feature point and core extraction. *Visual Computer*, 21(8-10):649–658, 2005. 1

[23] A. Kaufman, D. Cohen, and R. Yagel. Volume graphics. *Computer*, 26(7):51–64, 1993. 2

[24] D. Kingma and M. Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 3, 4

[25] X. Li, L. Wang, and Y. Fang. Pc-net: Unsupervised point correspondence learning with neural networks. *International Conference on 3D Vision (3DV)*, 2019. 1

[26] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen. Pointcnn: Convolution on x-transformed points. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 820–830. Curran Associates, Inc., 2018. 6

[27] Z. Z. Ling Zhang. Unsupervised feature learning for point cloud by contrasting and clustering with graph convolutional neural network. *3DV*, 04 2019. 6

[28] X. Liu, Z. Han, Y.-S. Liu, and M. Zwicker. Point2sequence: Learning the shape representation of 3d point clouds with an attention-based sequence to sequence network. 2019. 6

[29] D. Lowd and C. Meek. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 641–647. ACM, 2005. 3

[30] D. Maturana and S. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *Ieee/rsj International Conference on Intelligent Robots and Systems*, pages 922–928, 2015. 1, 2, 3

[31] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems 30*, pages 5099–5108, 2017. 6

[32] E. Rodola, A. M. Bronstein, A. Albarelli, F. Bergamasco, and A. Torsello. A game-theoretic approach to deformable shape matching. In *The IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 182–189. IEEE, 2012. 1

[33] A. Sharma, O. Grau, and M. Fritz. Vconv-dae: Deep volumetric shape learning without object labels. In *European Conference on Computer Vision*, pages 236–250, 2016. 3, 6

[34] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. December 2015. 3

[35] H. Su, C. R. Qi, Y. Li, and L. J. Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 6

[36] 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)*, 2019. 6

[37] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems*, pages 82–90, 2016. 1, 2, 3, 4, 6

[38] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1912–1920, 2015. 1, 2

[39] J. Xie, G. Dai, F. Zhu, E. Wong, and Y. Fang. Deepshape: Deep-learned shape descriptor for 3d shape retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016. 1

[40] J. Xie and Y. Fang. Learning barycentric representations of 3d shapes for sketch-based 3d shape retrieval. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[41] J. Xie, Z. Zheng, R. Gao, W. Wang, S. Zhu, and Y. Wu. Learning descriptor networks for 3d shape synthesis and analysis. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[42] Y. Yang, C. Feng, Y. Shen, and D. Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, 2018. 6

[43] A. K. F. Y. L. Z. X. T. Z. Wu, S. Song and J. Xiao. 3D ShapeNets: A Deep Representation for Volumetric Shapes. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2015. 5

[44] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[45] F. Zhu and Y. Fang. Heat diffusion long-short term memory learning for 3d shape analysis. *European Conference on Computer Vision (ECCV)*, 2016. 1

[46] J. Zhu and Y. Fang. Learning object-specific distance from a monocular image. *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1