View Inter-Prediction GAN: Unsupervised Representation Learning for 3D Shapes by Learning Global Shape Memories to Support Local View Predictions

Zhizhong Han,[1,2] Mingyang Shang,[1] Yu-Shen Liu,[1*] Matthias Zwicker[2]
[1]School of Software, Tsinghua University, Beijing, China
[2]Department of Computer Science, University of Maryland, College Park, USA
h312h@umd.edu, smy16@mails.tsinghua.edu.cn, liuyushen@tsinghua.edu.cn, zwicker@cs.umd.edu

Abstract

In this paper, we present a novel unsupervised representation learning approach for 3D shapes, which is an important research challenge as it avoids the manual effort required for collecting supervised data. Our method trains an RNN-based neural network architecture to solve multiple view inter-prediction tasks for each shape. Given several nearby views of a shape, we define view inter-prediction as the task of predicting the center view between the input views, and reconstructing the input views in a low-level feature space. The key idea of our approach is to implement the shape representation as a shape-specific global memory that is shared between all local view inter-predictions for each shape. Intuitively, this memory enables the system to aggregate information that is useful to better solve the view inter-prediction tasks for each shape, and to leverage the memory as a view-independent shape representation. Our approach obtains the best results using a combination of L_2 and adversarial losses for the view inter-prediction task. We show that VIP-GAN outperforms state-of-the-art methods in unsupervised 3D feature learning on three large-scale 3D shape benchmarks.

Introduction

Feature learning for 3D shapes is crucial for 3D shape analysis, including classification (Sharma, Grau, and Fritz 2016; Wu et al. 2016; Han et al. 2016; Yang et al. 2018; Achlioptas et al. 2018; Han et al. 2018), retrieval (Sharma, Grau, and Fritz 2016; Wu et al. 2016; Han et al. 2016; Yang et al. 2018; Achlioptas et al. 2018; Han et al. 2018), correspondence (Han et al. 2016; 2018) and segmentation (Qi et al. 2017a; 2017b). In recent years, supervised 3D feature learning has produced remarkable results under large-scale 3D benchmarks by training deep neural networks with supervised information (Qi et al. 2017a; 2017b), such as class labels and point correspondences. However, obtaining supervised information requires intense manual labeling effort. Therefore, unsupervised 3D feature learning with deep neural networks is an important research challenge.

Several studies have addressed this challenge (Sharma, Grau, and Fritz 2016; Wu et al. 2016; Han et al. 2016; Girdhar et al. 2016; Rezende et al. 2016; Yang et al. 2018; Achlioptas et al. 2018; Han et al. 2018) by training deep learning models using "supervised" information mined from the unsupervised scenario. This mining procedure is usually implemented using different prediction strategies, such as the prediction of a shape from itself by minimizing reconstruction error or embedded energy, the prediction of a shape from its context given by views or local shape features, or the prediction of a shape from views and itself together. These methods use multiple views to provide a holistic context of 3D shapes, and they make a single global shape prediction based on all views.

In contrast, our approach called *View Inter-Prediction GAN* (VIP-GAN) learns to make multiple local view inter-predictions among neighboring views. The view inter-prediction task is designed to mimic human perception of view-dependent patterns. That is, based on changes between neighbor views, humans can easily imagine the center view between, while the neighbor views can also be reversely imagined based on the center. As a key idea, our network architecture implements the shape representation as a shape-specific global memory whose contents are learned to support all local view inter-prediction tasks for each shape. Intuitively, the memory aggregates information over all view inter-prediction tasks, which leads to a view-independent shape representation. Our experimental results indicate that the obtained representation is highly discriminative and outperforms competing techniques on several standard shape classification benchmarks.

More specifically, VIP-GAN considers multiple views taken around a 3D shape in sequence as the context of the 3D shape, and it separates each view sequence into several overlapping sections of equal length. It then learns to predict the center view from its neighbors in each section, and the neighbors from the center. Crucially, VIP-GAN includes a memory shared by all view predictions of each shape. We show that the system uses this memory to improve its view prediction performance, in effect by learning a view independent shape representation. VIP-GAN employs an RNN-based generator with an encoder-decoder structure to implement the view inter-prediction strategy in different spaces. The encoder RNN captures the content information and spatial relationship of the neighbors to predict the center in 2D view space, while the decoder RNN predicts the neighbors in a low-level feature space according to the center predicted

*Corresponding author: Yu-Shen Liu

by the encoder. To further improve the prediction of the center, we train the generator jointly with a discriminator in an adversarial way. In summary, our significant contributions are as follows:

i) We propose VIP-GAN as a novel deep learning model to perform unsupervised 3D global feature learning through view inter-prediction with adversarial training, which leads to state-of-the-art performance in shape classification and retrieval.

ii) VIP-GAN makes it possible to mine fine-grained "supervised" information within the multi-view context of 3D shapes by imitating human perception of view-dependent patterns, which facilitates effective unsupervised 3D global feature learning.

iii) We introduce a novel implicit aggregation technique for 3D global feature learning based on RNN, which enables VIP-GAN to aggregate knowledge learned from each view prediction across a view sequence effectively.

Related work

Supervised 3D feature learning. Recently, supervised 3D feature learning is an attractive topic. With class labels, various deep learning models have been proposed to learn 3D features from different 3D raw representations, such as voxels (Wu et al. 2015), meshes (Han et al. 2018), points clouds (Qi et al. 2017a; 2017b) and views (Bai et al. 2017; Shi et al. 2015; Sfikas, Theoharis, and Pratikakis 2017; Sinha, Bai, and Ramani 2016; Su et al. 2015; Johns, Leutenegger, and Davison 2016; Kanezaki, Matsushita, and Nishida 2018), which aims to capture the mapping between 3D raw representations and class labels. The mapping is captured by spotting the distribution patterns among voxels (Wu et al. 2015), points in cloud (Qi et al. 2017a; 2017b), vertices on mesh (Han et al. 2018), or view features taken from different shapes (Bai et al. 2017; Shi et al. 2015; Sfikas, Theoharis, and Pratikakis 2017; Sinha, Bai, and Ramani 2016; Su et al. 2015; Johns, Leutenegger, and Davison 2016; Kanezaki, Matsushita, and Nishida 2018). Among these methods, multi-view based 3D feature learning methods perform the best, where pooling is widely used for view aggregation.

Unsupervised 3D feature learning. Although unsupervised 3D feature learning methods (Sharma, Grau, and Fritz 2016; Wu et al. 2016; Han et al. 2016; Girdhar et al. 2016; Rezende et al. 2016; Yang et al. 2018; Achlioptas et al. 2018; Han et al. 2018) are not always with high performance as supervised ones, their promising advantage of learning without labels still draws a lot of attention. To mine "supervised" information from unsupervised scenario, unsupervised feature learning methods usually train deep learning models by different prediction strategies, such as the prediction of a shape from itself by minimizing reconstruction error (Sharma, Grau, and Fritz 2016; Wu et al. 2016; Yang et al. 2018; Achlioptas et al. 2018) or embedded energy (Han et al. 2016), the prediction of a shape from context (Han et al. 2018), or the prediction of a shape from context and itself together (Girdhar et al. 2016; Rezende et al. 2016). These methods employ different kinds of 3D raw representations, such as voxels (Sharma, Grau, and Fritz 2016; Wu et al. 2016; Girdhar et al. 2016; Rezende et al. 2016), meshes (Han et al. 2016; 2018) or point clouds (Yang et al. 2018; Achlioptas et al. 2018), and accordingly, different kinds of context, such as spatial context of virtual words (Han et al. 2018) or views (Girdhar et al. 2016; Rezende et al. 2016), are employed. With the ideas of auto-encoder (Sharma, Grau, and Fritz 2016; Girdhar et al. 2016; Rezende et al. 2016; Yang et al. 2018; Achlioptas et al. 2018), classification (Han et al. 2018) or generative adversarial training (Wu et al. 2016; Achlioptas et al. 2018), these methods effectively learn discriminative 3D features. Different from these methods, VIP-GAN tries to learn 3D features by performing view inter-prediction to mine fine-grained "supervised" information within the multi-view context of 3D shapes, where context formed by multiple views is first explored for 3D global feature learning with adversarial training.

View synthesis and unsupervised video feature learning. View synthesis aims to generate novel views according to existing views. Deep learning based view synthesis has been drawing more and more research interests (Tatarchenko, Dosovitskiy, and Brox 2016; William, Gabriel, and David 2017). First tries teach deep learning models to predict novel views according to input views and transformation parameters (Tatarchenko, Dosovitskiy, and Brox 2016). To generate views with more detail (i.e. texture) and less geometric distortions, external image sets or geometric constraints are further employed.

Similarly, to predict the future frames in a video, the information of multiple past frames is aggregated by RNN (William, Gabriel, and David 2017). However, these methods mainly focus on the quality of generated views rather than the discriminability of learned features, where we find the view quality is not a sufficient condition for the feature discriminability in our experiments. In addition, the knowledge learned in each prediction cannot be aggregated by these methods to represent the global features. Therefore, these methods cannot be directly used for unsupervised 3D feature learning from view inter-prediction, which highlights our novelty by differentiating VIP-GAN apart from them.

VIP-GAN is also different from unsupervised video feature learning studies. Sequential views of 3D shapes are different from video frames because there is no firm starting position in view sequences. Each view could be the first view because of 3D shape rotation. This requires VIP-GAN to be invariant to the initial view position, that is, no matter which view of a 3D shape is the first, the learned feature of the shape should be the same. This is the main characteristic that makes VIP-GAN different from unsupervised video feature learning (At test stage, sensitive to the first frame of a video). Similarly, unsupervised image feature learning cannot aggregate multiple views and employ multiple view consistency as VIP-GAN.

VIP-GAN

Overview. The framework of VIP-GAN is illustrated in Fig. 1. Using multiple local view inter-predictions, VIP-GAN aims to learn a global representation or feature $\boldsymbol{F}$ of a 3D shape m from V views v_i sequentially taken around m, where $i \in [1, V]$. Note that $\boldsymbol{F}$ is learned for each shape as an F dimensional vector, effectively serving as a view-independent memory that is used in all local view inter-predictions for the shape. Hence $\boldsymbol{F}$ implicitly aggregates the knowledge learned from all sections $\boldsymbol{s}_i$ across the V views. Learning $\boldsymbol{F}$ is performed via gradient descent together with the other parameters in VIP-GAN, where $\boldsymbol{F}$ is randomly initialized. We split the set of views into V sections of equal length, where a section $\boldsymbol{s}_i$ is centered at each view v_i. We denote the center view of the section as c, and its N neighbors as n_j, where $j \in [1, N]$ ($N = 2$ in Fig. 1). In each section $\boldsymbol{s}_i$, VIP-GAN first predicts the center c in 2D space from the neighbors n_j. Conversely, it also predicts n_j in feature space from the predicted center c'.

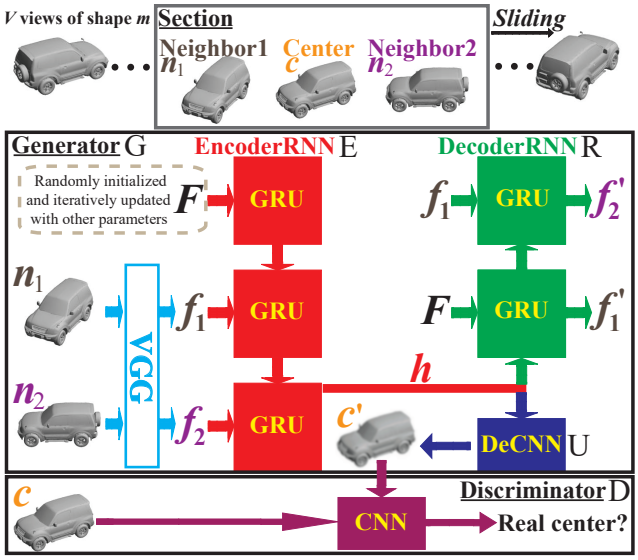

Figure 1: VIP-GAN is composed of generator G and discriminator D. The global feature $\boldsymbol{F}$ is learned in G by view inter-prediction through encoder E, decoder R and deconvolutional net U.

VIP-GAN consists of two main components, the generator G and discriminator D. The goal of the generator is to predict the center view in each section from its neighbors in image space, and the neighbors from the center in feature space. G consists of a VGG19 network, an encoder RNN E (in red), a decoder RNN R (in green) and a deconvolutional network U (in blue), where E and R are implemented by Gated Recurrent Units (GRUs). In addition, the discriminator D (in purple) is a convolutional network to distinguish whether a center view is real or not. G and D are jointly trained in an adversarial manner.

Generator G. In each section $\boldsymbol{s}_i$ of shape m, the first task of generator G is to collect a feature vector $\boldsymbol{h}_i$ that will be used to generate the predicted center view c'. For this purpose, the generator encodes the content within the neighbor views n_j and the spatial relationship among them. We extract the content of each n_j as a 4096 dimensional feature vector $\boldsymbol{f}_j$ by the last fully connected layer of a VGG19 network, where the resolution of input n_j is 224×224. We further encode the $\boldsymbol{f}_j$ with their spatial relationship using an encoder RNN E. We provide the global feature $\boldsymbol{F}$ of shape m, our learning target, at the first step of the encoder E serving as a knowledge container or memory that keeps incorporating the knowledge derived from each view prediction. Different from pooling, which is widely used as an explicit view aggregation, this implicit aggregation enables VIP-GAN to learn from more fine-grained information, such as the spatial relationship among the neighbors n_j in each section $\boldsymbol{s}_i$, and the connection between knowledge derived from different sections $\boldsymbol{s}_i$ across V views of m. Finally, at the last step of the encoder E for each section $\boldsymbol{s}_i$ we obtain a 4096 dimensional feature $\boldsymbol{h}_i$ as the hidden state, which we subsequently use to generate the predicted center c' using a deconvolutional network U.

By reshaping the 4096 dimensional $\boldsymbol{h}_i$ into 256 feature maps of size 4×4, the deconvolutional network U starts generating the predicted center c' with a resolution of 64×64 through four deconvolutional layers. The deconvolutional layers employ 256, 128, 64, and 3 kernels, respectively, and each kernel has size 3×3 and a stride of 2. In each deconvolutional layer, we use a leaky ReLu with a leaky gradient of 0.2. We utilize the L-2 loss between the predicted center view $\mathrm{U}(\boldsymbol{s}_i) = c'$ and the ground truth center c to measure the center prediction performance of G, denoted as loss L_{U},

$$L_{\mathrm{U}} = \|\mathrm{U}(\boldsymbol{s}_i) - c\|_2^2. \tag{1}$$

The second task of generator G is to reversely predict the neighbors n_j from the predicted center c' in each section $\boldsymbol{s}_i$. Different from the center view prediction task, we evaluate the prediction in feature space here. The two prediction tasks in different spaces enable VIP-GAN to more fully understand the 3D shape m. To predict both the content information within each n_j and the spatial relationship among n_j from the predicted center c', we employ a decoder RNN R with $\boldsymbol{h}_i$ as initialized hidden state that predicts the features $\boldsymbol{f}'_j$ of each neighbor view n_j step by step. Similar to the encoder E, we provide the global feature $\boldsymbol{F}$ at the first step of R, which is regarded as a reference for the following neighbor feature predictions. Then, $\boldsymbol{f}'_j$ is produced at the j-th step of R using the feature $\boldsymbol{f}_{j-1}$ of its previous counterpart as input. We predict the features $\boldsymbol{f}'_j$ in the same order as we provide the corresponding $\boldsymbol{f}_j$ to the encoder E. We measure the neighbor prediction performance of G using L-2 loss in feature space,

$$L_{\mathrm{R}} = \frac{1}{N} \sum_{j=1}^{N} \|\mathrm{R}(\boldsymbol{s}_i)_j - \boldsymbol{f}_j\|_2^2, \tag{2}$$

where $\mathrm{R}(\boldsymbol{s}_i)_j = \boldsymbol{f}'_j$ is the output at the j-th step of R. In summary, the loss of G is formed by the loss L_{U} of U and the loss L_{R} of R.

Discriminator D. In preliminary experiments, we found that the quality of predicted center views c' is not a sufficient condition to obtain a highly discriminative global feature $\boldsymbol{F}$.

Table 1: The effects of balance weights α and β on the performance of VIP-GAN under ModelNet10.

(α,β)	(1,0.05)	(3,0.05)	(5,0.05)	(3,0.1)	(3,0.01)	(3,0)	(0,0.01)	(0,0)	(0,0)C
Instance ACC	92.73	**94.05**	93.50	92.84	91.19	92.51	83.37	84.80	75.77
Class ACC	92.23	**93.71**	93.01	92.50	90.62	92.08	82.05	83.96	74.78

Table 2: The effects of parameters on VIP-GAN under ModelNet10 in terms of accuracy.

Parameters	R	D	$F(1024)$	$F(2048)$	$F(4096)$	$N(2)$	$N(6)$	$V(6)$	$V(3)$	cGan	BiDir
Instance ACC	90.53	47.80	92.29	92.51	**94.05**	93.17	93.50	92.62	92.51	89.10	93.83
Class ACC	89.88	44.49	91.73	92.03	**93.71**	92.91	93.08	92.32	92.22	88.34	93.45

For example, a complex and powerful deconvolutional network could generate c' with higher quality than our simple one introduced before, but we found that the learned feature $\boldsymbol{F}$ is much less discriminative. This phenomenon is caused by the large capacity of the more complex deconvolutional network to generate high quality view c' from any feature $\boldsymbol{h}_i$. However, this may decrease the discriminability of the learned feature $\boldsymbol{F}$. What we really want to achieve is that the quality of predicted views c' is mainly due to the discriminability of the learned feature $\boldsymbol{F}$, rather than the powerful learning ability of the deconvolutional network.

To resolve this issue, we employ discriminator D with adversarial training to facilitate our simple deconvolutional network U. Specifically, D is a CNN with five layers, including four convolutional layers and a one dimensional fully connected layer, where the resolution of input views is 64×64. Each convolutional layer contains 64, 128, 256, 512 kernels respectively, and each kernel has size 5×5 and a stride of 2, where we employ a leaky ReLu with a leaky gradient of 0.2. In the last layer of D, a sigmoid function provides the probability that the input is a real center view. Finally, the loss of D is the cross entropy of the probability produced from each $\boldsymbol{s}_i$, as defined as L_{D} in Eq. 3, where $\mathrm{D}(\mathrm{U}(\boldsymbol{s}_i))$ is the probability that D thinks the predicted center c' from $\boldsymbol{s}_i$ by U is real,

$$L_{\mathrm{D}} = \log \mathrm{D}(c) + \log(1 - \mathrm{D}(\mathrm{U}(\boldsymbol{s}_i))). \tag{3}$$

Adversarial training. Adversarial training is based on Generative Adversarial Networks (GAN) (Goodfellow et al. 2014). The predicted center c' from the generator G is passed to the discriminator D with the real center c, where D tries to learn how to distinguish whether a center is real or not. With adversarial training, the discriminator D is trained to maximize the probability when the center is real while minimizing it when the center is generated by generator G, as defined in Eq. 3. In contrast, the generator G has to be trained to fool the discriminator D. Therefore, in $\boldsymbol{s}_i$, the loss L_{D2U} for G from D is defined to make the predicted center $\mathrm{U}(\boldsymbol{s}_i)$ generated by U more real,

$$L_{\mathrm{D2U}} = \log(1 - \mathrm{D}(\mathrm{U}(\boldsymbol{s}_i))). \tag{4}$$

Finally, we define the loss function of VIP-GAN by combining the aforementioned losses as in Eq. 5, where the weights α and β are used to control the balance among them,

$$L = L_{\mathrm{U}} + \alpha L_{\mathrm{R}} + \beta L_{\mathrm{D2U}}. \tag{5}$$

Note that simultaneously with the other network parameters, we also optimize the learning target $\boldsymbol{F}$ by minimizing L using a standard gradient descent approach by iteratively updating $\boldsymbol{F}$ by $\boldsymbol{F} \leftarrow \boldsymbol{F} - \varepsilon \times \partial L / \partial \boldsymbol{F}$, where ε is the learning rate.

Modes for testing. Typically, there are two modes of unsupervised learning of features $\boldsymbol{F}$ of 3D shapes for testing, which we call the known-test mode and the unknown-test mode. In known-test mode, the test shapes are given with the training shapes at the same time, such that the features of test shapes can be learned with the features of training shapes together. In unknown-test mode, VIP-GAN is first pre-trained under training shapes. At test time, we then iteratively learn the features of test shapes by minimizing Eq. 5 with fixed pre-trained parameters of U, R and D.

Experimental results and analysis

In this section, the performance of VIP-GAN is evaluated and analyzed. First we discuss the setup of parameters involved in VIP-GAN. These parameters are tuned to demonstrate how they affect the discriminability of learned features in shape classification under ModelNet10 (Wu et al. 2015). Then, VIP-GAN is compared with state-of-the-art methods in shape classification and retrieval under ModelNet10 (Wu et al. 2015), ModelNet40 (Wu et al. 2015) and ShapeNet55 (Savva et al. 2017). All classification is conducted by a linear SVM (with default parameters in scikit-learn toolkit) under the global features learned by VIP-GAN.

Parameter setup. The balance weights α and β are important for the performance of VIP-GAN. In this experiment, we explore the effects of α and β on the performance of VIP-GAN under ModelNet10 in terms of average instance accuracy and average class accuracy, as shown in Table 1. Initially, the dimension F of global feature $\boldsymbol{F}$ is 4096, the center c gets $N = 4$ neighbors, and the $V = 12$ views of all 3D shapes under ModelNet10 are employed to train VIP-GAN in known-test mode. α and β are set to 1 and 0.05, respectively, since they make the initial values of loss L_{U}, L_{R} and L_{D2U} comparable to each other, where a normal distribution with mean of 0 and standard deviation of 0.02 is used to initialize the parameters involved in VIP-GAN.

First, the effect of α is explored by incrementally increasing α from 1 to 3 and 5. With $\alpha = 3$, best performance of VIP-GAN is achieved up to 94.05%, and the results with $\alpha = 3$ are better than the results with $\alpha = 5$. Then, the effect

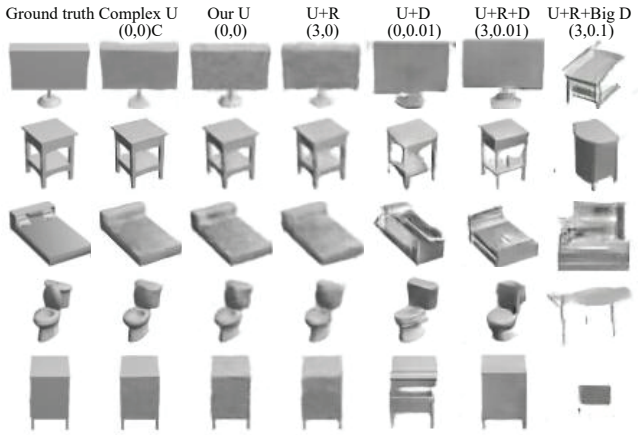

Figure 2: The predicted centers generated with different pairs of balance parameters (α, β).

of β is explored based on $\alpha = 3$ by increasing β to 0.1 and decreasing β to 0.01. These degenerated results show that the adversarial loss should not be over- or under-weighted. Subsequently, we highlight the contribution of discriminator D and decoder R to deconvolutional network U by incrementally setting α and β to 0. By setting β to 0, the results with "(3,0)" are better than the results with "(3,0.01)", but worse than the results with "(3,0.1)". This phenomenon implies that the under-weighted GAN loss is not helpful to increase the discriminability of learned features. We observe a similar phenomenon by comparing between "(0,0.01)" and "(0,0)". The comparison between "(3,0)" and "(0,0)" shows that the decoder R significantly increases the discriminability of learned features. In summary, these results show that the decoder R and the discriminator D can both improve the performance of VIP-GAN. However, R contributes more than D to U, and α is less sensitive than β.

Furthermore, as mentioned before, the quality of predicted center c' is not a sufficient condition to obtain a highly discriminative global feature $\boldsymbol{F}$. By replacing our simple U with a more complex one employed in (Dosovitskiy and Brox 2016), the quality of predicted centers becomes higher, as shown in the comparison between "(0,0)" and "(0,0)C" in Fig. 2. On the other hand, the discriminability of the learned global feature $\boldsymbol{F}$ dramatically decreases, as illustrated by the comparison between "(0,0)" and "(0,0)C" in Table 1. The reason for this is that the more complex deconvolutional network in (Dosovitskiy and Brox 2016) is too deep to facilitate effective error back propagation to train a highly discriminative global feature. To keep the network in (Dosovitskiy and Brox 2016) unchanged, the predicted views are generated in the resolution of 256×256 rather than 64×64, where the 224×224 ground truth views are padded with pixel values of 255 to enable the computation of loss L_{U}. Finally, we also highlight the importance of R and D by merely using L_{R} or L_{D2U} to train, as shown by "R" and "D" in Table 2. Compared with the importance of U as "(0,0)" in Table 1, R plays the most important role in VIP-GAN.

The predicted centers c' generated by different α and β are demonstrated in Fig. 2, where the tags marking each column are consistent with the parameters in Table. 1. According to the ground truth, the complex deconvolutional network ("(0,0)C") generates centers with higher quality than our simple ones ("(0,0)"). The comparison between "(0,0)" and "(3,0)" shows that the decoder R slightly degenerates the quality of predicted centers. In addition, the adversarial loss weighted by small β can make the predicted centers sharper, but also produce distortions, as illustrated by the comparison between "(0,0)" and "(0,0.01)", and the comparison between "(3,0)" and "(3,0.01)". The adversarial loss weighted by big β will make the loss L_{U} subtle with big distortions, as shown by "(3,0.1)".

Table 3: The comparison of classification accuracy under ModelNet10 and ModelNet40.

Methods	Supervised	MN40	MN10
MVCNN	Yes	90.10	-
MVCNN-Multi	Yes	91.40	-
ORION	Yes	-	93.80
3DDescriptorNet	Yes	-	92.40
Pairwise	Yes	90.70	92.80
GIFT	Yes	89.50	91.50
PANORAMA	Yes	90.70	91.12
VoxNet	Yes	-	92.00
VRN	Yes	91.33	93.80
RotationNet	Yes	90.65	93.84
PointNet++	Yes	91.90	-
T-L	No	74.40	-
LFD	No	75.47	79.90
Vconv-DAE	No	75.50	80.50
3DGAN	No	83.30	91.00
LGAN	No	85.70	95.30
LGAN(MN40)	No	87.27	92.18
FNet	No	88.40	94.40
FNet(MN40)	No	84.36	91.85
Our	No	**91.98**	**94.05**
Our1(SN55)	No	90.19	92.18
Our2(+SN55)	No	91.25	92.84

The effects of F, N and V are further explored in Table 2. By gradually decreasing F from 4096 to 2048 and 1024, the results are degenerated from 94.05% to 92.51% and 92.29%. To conduct this experiment with the rest of VIP-GAN unchanged, one more 4096 dimensional fully connected layer is employed before $\boldsymbol{F}$ is inputted in G. Then, the number N of neighbors in each section $\boldsymbol{s}_i$ is explored by respectively decreasing N to 2 and increasing N to 6, based on the $N = 4$ structure with our best results. Although these results are degenerated from our best results, they are still good. The degeneration is caused by that less neighbors could not provide enough discriminative information to learn while more neighbors would bring redundant discriminative information. Following this, we decrease V to 6 and 3 gradually, the results are also decreased due to the less information for learning, where N is adjusted to 2 when V is set to 3. Subsequently, we employ conditional GAN to replace the GAN structure in VIP-GAN, where the ground truth neigh-

bors are regarded as the conditions of the center. The high-level features $\boldsymbol{f}_j$ of neighbors are concatenated with the extracted feature of the center after the last convolutional layer in discriminator D, which is further followed by an extra convolutional layer and the one dimensional fully connected layer. Although the results dramatically decreased as shown by "cGan", it is still better than merely using U as listed "(0,0)" in Table 1. These results imply that GAN is better than conditional GAN for 3D global feature learning in VIP-GAN, while both the adversarial loss of GAN and conditional GAN are helpful to improve the discriminability of learned features. Moreover, we also try to train VIP-GAN by bidirectional view sequences, since human can perform the view inter-prediction from either left to right or right to left in a view sequence, as shown by the results listed as "BiDir". However, no further improvement is obtained from the doubled training samples.

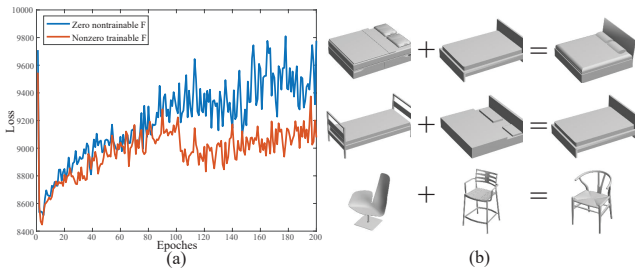

Figure 3: (a)The effectiveness of our novel implicit view aggregation is shown by the comparison between the loss with nonzero trainable $\boldsymbol{F}$ and the loss with zero non-trainable $\boldsymbol{F}$. (b)The learned global features are visualized by feature manipulation in the embedding space.

Classification. We compare VIP-GAN with the state-of-the-art methods in classification under ModelNet40 and ModelNet10. The parameters under ModelNet40 are the same ones with our best results under ModelNet10 in Table 2. The compared methods include MVCNN (Su et al. 2015), ORION (Sedaghat et al. 2017), 3DDescriptorNet (Xie et al. 2018), Pairwise (Johns, Leutenegger, and Davison 2016), GIFT (Bai et al. 2017), PANORAMA (Sfikas, Theoharis, and Pratikakis 2017), VRN (Brock et al. 2016), RotationNet (Kanezaki, Matsushita, and Nishida 2018), PointNet++ (Qi et al. 2017b), T-L (Girdhar et al. 2016), LFD, Vconv-DAE (Sharma, Grau, and Fritz 2016), 3DGAN (Wu et al. 2016), LGAN (Achlioptas et al. 2018), and FNet (Yang et al. 2018).

VIP-GAN significantly outperforms all its unsupervised competitors under ModelNet40, and some of them under ModelNet10, as shown by "Our", which is also the best result compared to eight top ranked supervised methods. For fair comparison, the result of VRN (Brock et al. 2016) is presented without ensemble learning, and the result of RotationNet(Kanezaki, Matsushita, and Nishida 2018) is presented with views taken by the default camera system orientation that is identical to the others. In addition, we try to train VIP-GAN under ShapeNet55 in unknown-test mode. Hence, we fix the parameters to extract features under ModelNet40 and ModelNet10, as shown by "Our1(SN55)". Although the results of LGAN(Achlioptas et al. 2018) and FNet(Yang et al. 2018) are better than "Our1(SN55)" under ModelNet10, it is inconclusive whether they are better than ours. This is because these methods are trained under a version of ShapeNet55 that contains more than 57,000 3D shapes, including a number of 3D point clouds. However, VIP-GAN is trained only under the 51,679 3D shapes from ShapeNet55 that are available for public download.

Table 4: The comparison of retrieval in terms of mAP under ModelNet40 and ModelNet10.

Methods	MN40	MN10
GeoImage	51.30	74.90
Pano	76.81	84.18
MVCNN	79.50	-
GIFT	81.94	91.12
RAMA	83.45	87.39
Trip	88.00	-
Our	**89.23**	**90.69**
Our1(SN55)	**87.66**	**90.09**
Our2(+SN55)	**88.87**	**90.75**

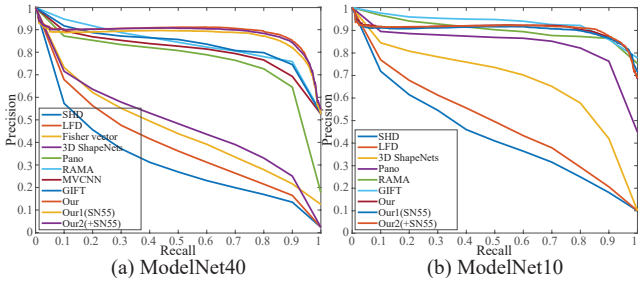

Figure 4: The comparison of PR curves for retrieval under ModelNet40 and ModelNet10.

Finally, we explore whether "Our" could be further improved by more training shapes from ShapeNet55 in known-test mode, as shown by "Our2(+SN55)". However, with the existing parameters, only comparable results are obtained. Moreover, we evaluate VIP-GAN under ShapeNet55 in known-test mode using the same parameters with our best results under ModelNet10 in Table 2, as shown in the rightmost column "Our" in Table 6. Similar to "Our2(+SN55)", with the existing parameters, only comparable results are obtained by more training shapes from ModelNet40, as shown by "Our+".

Our novel implicit view aggregation. The effect of our novel implicit view aggregation is first explored by visualization. In Fig. 3(a), we compare the training loss of our framework with a fixed, non-trainable $\boldsymbol{F}$ set to zero, and our trainable $\boldsymbol{F}$. Our approach is able to learn the characteristics of each shape to make up the missing information in each prediction, which reduces the training loss. The two losses show that the generator is getting to the Nash equilibrium.

In Fig. 3(b), we further evaluate the semantic meaning of our features by manipulating them algebraically, and visual-

Table 5: The effects of our novel implicit view aggregation under ModelNet10.

ACC	Non-trainable $\boldsymbol{F}$		Trainable $\boldsymbol{F}$		
	MaxP	MeanP	MaxP	MeanP	Our
Ins	84.58	87.22	81.72	82.49	**94.05**
Cla	83.95	87.38	80.60	81.73	**93.71**

izing the result via nearest neighbor retrieval in ModelNet10, as shown on the right. The retrieved shapes exhibit characteristics similar to both input shapes, such as the surface of the bed in the first row, and the bedhead in the second row.

Finally, we compare our implicit view aggregation with the widely used explicit view aggregation pooling under ModelNet10. Here, we use the output $\boldsymbol{h}_i$ of the encoder E as the feature of each view, and obtain the global feature of the shape by pooling all the $\boldsymbol{h}_i$ together with maxpooling and meanpooling, where each $\boldsymbol{h}_i$ is obtained with trainable $\boldsymbol{F}$ and non-trainable all zero $\boldsymbol{F}$. In Table 5, with trainable or non-trainable $\boldsymbol{F}$, our implicit view aggregation is always superior to the pooling. Without the support of trainable $\boldsymbol{F}$, the pooled features are pushed to be more discriminative than the ones with trainable $\boldsymbol{F}$ to minimize the loss, which makes the pooling results better. However, it is still not good enough to keep the loss as low as ours shown in Fig. 3(a).

Table 6: Retrieval and classification comparison in terms of Micro-averaged metrics under ShapeNetCore55.

	Micro				
Methods	P	R	F1	mAP	NDCG
Kanezaki	81.0	80.1	**79.8**	77.2	86.5
Zhou	78.6	77.3	76.7	72.2	82.7
Tatsuma	76.5	80.3	77.2	74.9	82.8
Furuya	**81.8**	68.9	71.2	66.3	76.2
Thermos	74.3	67.7	69.2	62.2	73.2
Deng	41.8	71.7	47.9	54.0	65.4
Li	53.5	25.6	28.2	19.9	33.0
Mk	79.3	21.1	25.3	19.2	27.7
Su	77.0	77.0	76.4	73.5	81.5
Bai	70.6	69.5	68.9	64.0	76.5
Taco	70.1	71.1	69.9	67.6	75.6
Our	60.0	**80.3**	61.2	**83.5**	**89.4**
Our+	60.0	**80.3**	61.2	**83.6**	**89.5**
Our accuracy			**82.97**		
Our+ accuracy			**82.51**		

Retrieval. VIP-GAN is further evaluated in shape retrieval under ModelNet40, ModelNet10 and ShapeNet55, as shown in Table 4, Table 6 and Table 7. The compared results include LFD, SHD, Fisher vector, 3D ShapeNets (Wu et al. 2015), GeoImage (Sinha, Bai, and Ramani 2016), Pano (Shi et al. 2015), MVCNN (Su et al. 2015), GIFT (Bai et al. 2017), RAMA (Sfikas, Theoharis, and Pratikakis 2017) and Trip (He et al. 2018).

In these experiments, the 3D shapes in the test set are used as queries to retrieve the rest shapes in the same set, and mean Average Precision (mAP) is used as a metric. In addition, we employ global features involved in our classification results in Table 3 and Table 6 for the retrieval experiments under the three benchmarks.

As shown in Table 4, our results of "Our" outperform all the compared results under ModelNet40, and sightly lower than the best results of 91.12 by GIFT under ModelNet10. However, it is inconclusive whether GIFT outperforms VIP-GAN, since the dataset used by GIFT is formed by randomly selecting 100 shapes from each shape class, which is much simpler than the whole benchmark that we used. In addition, with trained by more shapes from ShapeNet55, the result of "Our2" under ModelNet10 is a little bit higher than the result of "Our". Their available PR curves under ModelNet40 and ModelNet10 are also compared in Fig. 4.

In Table 6 and Table 7, the results of "Our" outperform all the compared results under ShapeNet55. Besides Taco (Cohen et al. 2018) in Table 6, the compared results without reference are from SHREC2017 shape retrieval contest (Savva et al. 2017) under ShapeNet55 with the same names, where micro-averaged and macro-averaged methods are employed to compute the metrics. Similar to "Our2" under ModelNet10, with trained by more shapes from ModelNet40, "Our+" is a little bit better than "Our".

Table 7: Retrieval comparison in terms of Macro-averaged metrics under ShapeNetCore55.

	Macro				
Methods	P	R	F1	mAP	NDCG
Kanezaki	60.2	63.9	**59.0**	58.3	65.6
Zhou	59.2	65.4	58.1	57.5	65.7
Tatsuma	51.8	60.1	51.9	49.6	55.9
Furuya	**61.8**	53.3	50.5	47.7	56.3
Thermos	52.3	49.4	48.4	41.8	50.2
Deng	12.2	66.7	16.6	33.9	40.4
Li	21.9	40.9	19.7	25.5	37.7
Mk	59.8	28.3	25.8	23.2	33.7
Su	57.1	62.5	57.5	56.6	64.0
Bai	44.4	53.1	45.4	44.7	54.8
Our	18.9	**81.2**	24.0	**69.2**	**83.7**
Our+	18.8	**81.3**	24.0	**69.9**	**84.0**

Conclusions

We proposed VIP-GAN, an approach for unsupervised 3D global feature learning by view inter-prediction that is capable of learning from fine-grained "supervised" information within the multi-view context of 3D shapes. Inspired by human perception of view-dependent patterns, VIP-GAN successfully learns more discriminative golbal features than state-of-the-art view-based methods that regard the multi-view context as a whole. With adversarial training, the global features can be learned more efficiently, which further improves their discriminability. In addition, our novel implicit aggregation enables VIP-GAN to learn within the multi-view context by effectively aggregating knowledge learned from multiple local view predictions across a view sequence. Our results show that VIP-GAN outperforms its

unsupervised counterparts, as well as some top ranked supervised methods under large-scale benchmarks in shape classification and retrieval.

Acknowledgments

Yu-Shen Liu is the corresponding author. This work was supported by National Key R&D Program of China (2018YFB0505400), the National Natural Science Foundation of China (61472202), and Swiss National Science Foundation grant (169151). We thank all anonymous reviewers for their constructive comments.

References

Achlioptas, P.; Diamanti, O.; Mitliagkas, I.; and Guibas, L. J. 2018. Learning representations and generative models for 3D point clouds. In *The International Conference on Machine Learning*, 40–49.

Bai, S.; Bai, X.; Zhou, Z.; Zhang, Z.; and Latecki, L. J. 2017. GIFT: Towards scalable 3D shape retrieval. *IEEE Transaction on Multimedia* 19(6):1257–1271.

Brock, A.; Lim, T.; Ritchie, J.; and Weston, N. 2016. Generative and discriminative voxel modeling with convolutional neural networks. In *3D deep learning workshop (NIPS)*.

Cohen, T. S.; Geiger, M.; Köhler, J.; and Welling, M. 2018. Spherical CNNs. In *International Conference on Learning Representations*.

Dosovitskiy, A., and Brox, T. 2016. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems*. 658–666.

Girdhar, R.; Fouhey, D. F.; Rodriguez, M.; and Gupta, A. 2016. Learning a predictable and generative vector representation for objects. In *Proceedings of European Conference on Computer Vision*, 484–499.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*. 2672–2680.

Han, Z.; Liu, Z.; Han, J.; Vong, C.-M.; Bu, S.; and Li, X. 2016. Unsupervised 3D local feature learning by circle convolutional restricted boltzmann machine. *IEEE Transactions on Image Processing* 25(11):5331–5344.

Han, Z.; Liu, Z.; Vong, C.; Liu, Y.-S.; Bu, S.; Han, J.; and Chen, C. 2018. Deep spatiality: Unsupervised learning of spatially-enhanced global and local 3D features by deep neural network with coupled softmax. *IEEE Transactions on Image Processing* 27(6):3049–3063.

He, X.; Zhou, Y.; Zhou, Z.; Bai, S.; and Bai, X. 2018. Triplet-center loss for multi-view 3D object retrieval. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Johns, E.; Leutenegger, S.; and Davison, A. J. 2016. Pairwise decomposition of image sequences for active multi-view recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3813–3822.

Kanezaki, A.; Matsushita, Y.; and Nishida, Y. 2018. Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3D classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, 5105–5114.

Rezende, D. J.; Eslami, S. M. A.; Mohamed, S.; Battaglia, P.; Jaderberg, M.; and Heess, N. 2016. Unsupervised learning of 3D structure from images. In *Advances in Neural Information Processing Systems*, 4997–5005.

Savva, M.; Yu, F.; Su, H.; Kanezaki, A.; Furuya, T.; Ohbuchi, R.; Zhou, Z.; Yu, R.; Bai, S.; Bai, X.; Aono, M.; Tatsuma, A.; Thermos, S.; Axenopoulos, A.; Papadopoulos, G. T.; Daras, P.; Deng, X.; Lian, Z.; Li, B.; Johan, H.; Lu, Y.; and Mk, S. 2017. SHREC'17 Large-Scale 3D Shape Retrieval from ShapeNet Core55. In *Eurographics Workshop on 3D Object Retrieval*.

Sedaghat, N.; Zolfaghari, M.; Amiri, E.; and Brox, T. 2017. Orientation-boosted voxel nets for 3D object recognition. In *British Machine Vision Conference*.

Sfikas, K.; Theoharis, T.; and Pratikakis, I. 2017. Exploiting the PANORAMA Representation for Convolutional Neural Network Classification and Retrieval. In *Eurographics Workshop on 3D Object Retrieval*, 1–7.

Sharma, A.; Grau, O.; and Fritz, M. 2016. VConv-DAE: Deep volumetric shape learning without object labels. In *Proceedings of European Conference on Computer Vision*, 236–250.

Shi, B.; Bai, S.; Zhou, Z.; and Bai, X. 2015. Deeppano: Deep panoramic representation for 3D shape recognition. *IEEE Signal Processing Letters* 22(12):2339–2343.

Sinha, A.; Bai, J.; and Ramani, K. 2016. Deep learning 3D shape surfaces using geometry images. In *European Conference on Computer Vision*, 223–240.

Su, H.; Maji, S.; Kalogerakis, E.; and Learned-Miller, E. G. 2015. Multi-view convolutional neural networks for 3D shape recognition. In *International Conference on Computer Vision*, 945–953.

Tatarchenko, M.; Dosovitskiy, A.; and Brox, T. 2016. Multi-view 3D models from single images with a convolutional network. In *European Conference on Computer Vision*, volume 9911, 322–337.

William, L.; Gabriel, K.; and David, C. 2017. Deep predictive coding networks for video prediction and unsupervised learning. In *International Conference on Learning Representations*.

Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; and Xiao, J. 2015. 3D ShapeNets: A deep representation for volumetric shapes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1912–1920.

Wu, J.; Zhang, C.; Xue, T.; Freeman, B.; and Tenenbaum, J. 2016. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In *Advances in Neural Information Processing Systems*. 82–90.

Xie, J.; Zheng, Z.; Gao, R.; Wang, W.; Zhu, S.-C.; and Wu, Y. N. 2018. Learning descriptor networks for 3D shape synthesis and analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Yang, Y.; Feng, C.; Shen, Y.; and Tian, D. 2018. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *IEEE Conference on Computer Vision and Pattern Recognition*.