Emphasizing 3D Properties in Recurrent Multi-View Aggregation for 3D Shape Retrieval

Cheng Xu, Biao Leng,[*] Cheng Zhang, Xiaochen Zhou
School of Computer Science and Engineering, Beihang University, Beijing 100191, China
{cxu, lengbiao, zcheng, zhouxiaochen}@buaa.edu.cn

Abstract

Multi-view based shape descriptors have achieved impressive performance for 3D shape retrieval. The core of view-based methods is to interpret 3D structures through 2D observations. However, most existing methods pay more attention to discriminative models and none of them necessarily incorporate the 3D properties of the objects. To resolve this problem, we propose an encoder-decoder recurrent feature aggregation network (**ERFA-Net**) to emphasize the 3D properties of 3D shapes in multi-view features aggregation. In our network, a view sequence of the shape is trained to encode a discriminative shape embedding and estimate unseen rendered views of any viewpoints. This generation task gives an effective supervision which makes the network exploit 3D properties of shapes through various 2D images. During feature aggregation, a discriminative feature representation across multiple views is effectively exploited based on LSTM network. The proposed 3D representation has following advantages against other state-of-the-art: 1) it performs robust discrimination under the existence of noise such as view missing and occlusion, because of the improvement brought by 3D properties. 2) it has strong generative capabilities, which is useful for various 3D shape tasks. We evaluate ERFA-Net on two popular 3D shape datasets, ModelNet and ShapeNetCore55, and ERFA-Net outperforms the state-of-the-art methods significantly. Extensive experiments show the effectiveness and robustness of the proposed 3D representation.

Introduction

With the rapid development of 3D scene techniques and the explosive growth of large-scale public 3D shape repositories (Chang et al. 2015), 3D shape retrieval has become more significant than ever. Among the existing methods on data-driven shape descriptors for 3D shapes, feature learning over multi-view image sequence of 3D shapes achieves the state-of-the-art performance on various 3D shape retrieval tasks (Su et al. 2015; Qi et al. 2016). Recently, many researchers have been devoted to learning a 3D representation based on deep learning techniques (Qi et al. 2017; Yi et al. 2017; Liu, Li, and Wang 2017).

Interpreting 3D structures through 2D observations of shapes is the core of the view-based algorithms. Moreover, several works have been conducted on discriminative 3D representation such as MVCNN (Su et al. 2015), fusing discriminative information of each view into a compact feature vector. These approaches, while effectively identifying different categories, have a major shortcoming: the learned representations do not significantly integrate 3D properties of shapes. Some local and crucial structures of 3D shapes like the doorknob of a door object are difficult to be captured from some rendered images. In addition, though MVCNN possesses the high discrimination power, the fact that the model is supervised only for classification tasks with 2D rendered images hinders its performance seriously in some cases such as view missing and occlusion. These cases pose challenges to the view-based 3D representation learning.

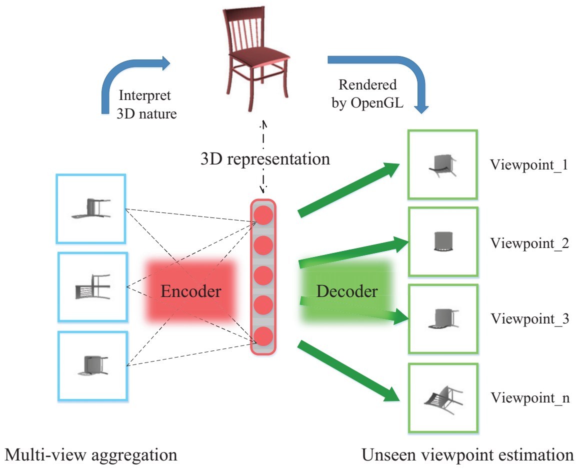


Figure 1: We propose a method for multi-view based shape descriptors through emphasizing 3D properties in multi-view features aggregation. A 3D representation is encoded with the view sequence of the 3D shape and trained to generate rendered images of unseen viewpoints.

Motivated by the recent line of work that learns to generate 3D shapes by using 3D CAD models (Dosovitskiy, Springenberg, and Brox 2015; Kulkarni et al. 2015; Choy et al. 2016; Yan et al. 2016), we introduce the generative model into the multi-view features fusion process to address

[*] Corresponding author.

these challenges. It is assumed that the 3D representation will necessarily incorporate the 3D properties, which is able to generate original 3D objects. However, the voxel representation of 3D shapes has the problems of low resolution, requiring redundant parameters and fewer 3D shapes compared to millions of images. Then, we take an equivalent alternative way that the 3D representation is trained to estimate rendered images of 3D shapes. As is shown in Fig. 1, we aggregate certain views' features to generate unseen rendered images of a 3D shape from any viewpoints.

In this paper, we propose an encoder-decoder recurrent feature aggregation network (**ERFA-Net**) to fuse multi-view features into a discriminative and generative 3D representation, aiming at incorporating 3D properties of shapes. Instead of only being supervised by the classification task, the network is trained to generate rendered images of a 3D shape from unseen viewpoints through fusing the information along a multi-view image sequence. The estimation task gives a strong supervision, which makes the network interpret 3D structures through the 2D multi-view image sequence. In order to enhance the quality of the generated images and adapt to various angles, the viewpoint transformation layer is presented in decoder part, where the different encoded viewpoints can transform the 3D representation into specific embedding space.

The aggregated scheme of multi-view features is vital for the discrimination power of the final 3D representation. In contrast to other state-of-the-art methods, our encoder part aggregates the spatial correlation information along a multi-view image sequence based on the recurrent neural network LSTM (Hochreiter and Schmidhuber 1997). In our network, shape information of different views flows along image sequence, in which the discriminative information can be captured and propagate to the deeper LSTM node. Then, with the accumulation of good features, the final aggregated representation can be more discriminative. Moreover, the randomly ordered view sequence is adopted in our method, making the proposed aggregated features robust to arbitrary rotation of 3D shapes.

We demonstrate with extensive evaluation that the encoded 3D representation has the following key advantages. Firstly, since 3D properties of shapes give a significant rise to 2D observations, our aggregated representation performs robust discrimination in some cases such as view missing and occlusion. Secondly, in addition to high discrimination, the proposed representation has strong generative capabilities. This enables us to tackle a variety of view-based 3D shape tasks. Thirdly, the feature is insensitive to arbitrary rotation of 3D shapes. Our 3D representation achieves the state-of-the-art performance on ModelNet dataset and ShapeNetCore55 dataset.

In summary, our main contributions are as follows.

- We propose an encoder-decoder recurrent feature aggregation network to build a robust discriminative and generative 3D shape representation via emphasizing 3D properties for multi-view 3D shape retrieval.
- We present a recurrent feature aggregation structure to capture the consistently discriminative features among views, exploiting spatial correlation information in the sequence based on LSTM.
- ERFA-Net significantly outperforms all the state-of-the-art methods on both ModelNet and ShapeNetCore55, and has the robustness to view missing and object occlusion.

Related Work

A large number of works (Leng et al. 2016; Girdhar et al. 2016; Fang et al. 2015; Xie et al. 2015) have been proposed to address 3D shape retrieval problem, which are coarsely divided into two categories: model-based methods and view-based methods. Compared to high dimensional model-based methods that exploit the raw 3D representations of 3D shapes, view-based methods leverage a highly informative image sequence with some desirable properties, such as regular structure, efficiency to compute and robustness to handle naive 3D representation. Moreover, owing to the success made by CNN in vision (Razavian et al. 2014; Girshick et al. 2013), the CNN-based deep representations of multiple views from 3D shapes achieve more impressive performance than model-based methods and lead the best performance on various 3D shape datasets, such as ModelNet and ShapeNetCore55. Then our work is mainly related to multi-view 3D shape retrieval approaches.

Traditionally, view-based methods mainly employed hand-crafted representations of 3D shapes. A typical example of a view-based technique is LightField Descriptor (Chen et al. 2003), placing 20 cameras on the vertices of a regular dodecahedron and representing a 3D shape with Fourier descriptors and Zernike moments. Another widely used instance is Bag-of-Words (BoW) model (Li and Perona 2005), since it has shown its superiority as natural image descriptor and reduced the computation complexity. Then BoRW is proposed through encoding regions of views with local SIFT features and clustering them with responding weights to describe a 3D shape. PANORAMA (Papadakis et al. 2010) obtained a panoramic view of a 3D shape, capturing the global shape information and improving the retrieval performance significantly. Meanwhile, (Bonaventura et al. 2015) proposed a shape descriptor of the Information Sphere and utilized mutual information-based measures for the matching. (Gao et al. 2012) represented a set of 3D shapes' rendered images with multiple hypergraphs and obtained a higher order relationship of 3D shapes.

Recently, the success made by CNN for a number of applications in vision, such as image classification (Krizhevsky, Sutskever, and Hinton 2012), object detection (Ren et al. 2017) and scene recognition (Zhou et al. 2014), has inspired many follow-on studies on view-based 3D shape retrieval. (Zhu et al. 2014) firstly used autoencoder for deep representations of 3D shapes based on projected views, exhibiting good complementarity with the local descriptor. GIFT (Bai et al. 2016) extracted view features by using CNN with GPU acceleration and adopting the inverted file to reduce computation in distance metrics. Meanwhile, a variant of CNN (Shi et al. 2015) was designed for learning a deep representation from a panoramic view of a 3D shape, adopting a row-wise max-pooling scheme

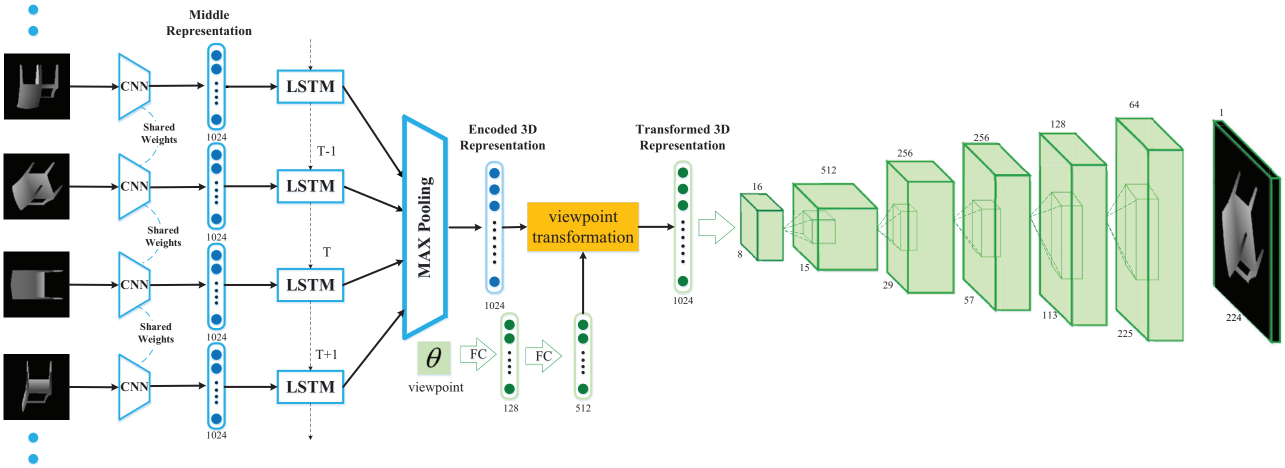


Figure 2: The detailed framework of our proposed encoder-decoder recurrent feature aggregation network. (a) The encoder (**blue**): recurrent feature aggregation network, fusing the discriminative information among a multi-view sequence into a compact 3D representation. (b) The decoder (**green**): viewpoint estimation network, processing the different viewpoints and generating rendered views correspondingly with the encoded 3D representation.

to obtain the robust alignment between different panoramic views. The aforementioned methods focused on learning a representation from every single view, which provides insufficient shape information.

Apart from a single view, many researchers attached importance to building a more discriminative representation with a highly informative multi-view image sequence of a 3D shape. In the MVCNN (Su et al. 2015), a set of CNNs was used to extract each view's deep representation and then aggregate information from multiple views into a compact shape descriptors with the element-wise maximum operation. TLC (Bai et al. 2015) was proposed for 3D shape retrieval, where each pair of views was represented in the first layer and a set of view-pair representations from different subsets were encoded into final feature vector in the second layer. (Johns, Leutenegger, and Davison 2016) overcame the limit on the fixed-length image sequence by decomposing a view sequence into a set of view pairs and then trained a CNN to build the pairwise relationship between different view pairs. However, these methods can not necessarily incorporate 3D properties of the shapes. In this paper, we propose a novel network for multi-view based shape retrieval, by learning a discriminative and generative 3D representation. During the training phase, a view sequence of the 3D shape is trained to generate the rendered images of all viewpoints. It is assumed that 3D properties of shapes can be effectively captured under the supervision where the unseen views should be estimated accurately. In addition, the information along a multi-view image sequence is recurrently aggregated based on LSTM in order to make the encoded 3D representation discriminative.

Proposed method

The goal of the proposed approach is to learn a 3D representation that is discriminative across multiple views, and generative enough to estimate rendered views of different viewpoints. These properties are significant for view-based 3D understanding tasks.

We propose a novel encoder-decoder embedding network for multi-view 3D shape retrieval. As shown in Fig. 2, this architecture consists of two parts: recurrent feature aggregation network (encoder) and viewpoint estimation network (decoder). On the one hand, the encoder network guarantees the learned 3D representation with high discrimination, where the discriminative information among multiple views is captured and aggregated automatically. On the other hand, the decoder network is trained to generate projection views of the same object from different viewpoints, which gives an effective supervision that the learned representation can necessarily incorporate the properties of 3D shapes.

Given a 3D shape x, we render an image sequence $S(x)$ from different viewpoints θ_i, the combination of azimuth θ_i^{az} and elevation θ_i^{el}, in a unit spherical coordinate system. We train the proposed approach that receives an input pair $(S(x_p), \theta_i)$, where x_p is the input 3D shape and θ_i is the target viewpoint, and then aims to estimate the target image y_i, where y_i is the rendered view of the same object x_p from the requested viewpoint θ_i. The encoder fuses the information of the image sequence $S(x)$ into a compact 3D representation D_p and then the decoder uses it to predict unseen views. The network is trained by minimizing the combination of two loss functions: the squared euclidean loss for the generated image and softmax loss $\mathcal{L}_S$ for the encoded 3D representation D_p:

$$\mathcal{L} = \mathcal{L}_S + \lambda \sum_i ||y_i - \hat{y}_i||_2^2 \qquad (1)$$

where $\hat{y}_i$ is the output of network and a scalar λ is used for balancing the two loss functions. In our experiment we set λ as 0.01.

Recurrent feature aggregation network

As illustrated in Fig. 2, the recurrent feature aggregation network contains two main parts: the extraction of features from multiple views and the aggregation of different visual feature vectors.

For the view feature extraction, a set of deep representations are extracted with CNNs for each 3D shape with a view image sequence. We use a set of CNNs to obtain different view representations, and each view is input to the CNN independently. For the image sequence $S(x_p)$, CNN is adopted to transform each projected image p_i into the feature vector v_i. The set of view features extracted from the image sequence of the 3D shape x_p can be expressed as $V(x_p) = \{v_1, v_2, ..., v_n\}$. Since the discriminative power of the fused image sequence feature heavily depends on the representations of different views, all features extracted from CNNs should be as discriminative as possible. To satisfy this requirement, we train CNN for each view and share the same weights. In terms of the CNN architecture, we adopt the GoogLeNet with Batch Normalization (Ioffe and Szegedy 2015). This 22-layer CNN has great power to learn the discriminative visual representation and fast convergence speed with batch normalization technique. We use features from the layer $pool5/7x7\_s1$ as the visual features, the dimension of which is 1024.

For the feature aggregation, due to the great power of recurrent neural networks in spatial-temporal sequence, we utilize an LSTM network to aggregate the complementary geometric correlation information in the view features sequence. In particular, an out-of-order view sequence of a 3D shape is adopted to cope with 3D shape transformation. An LSTM node includes three gates: the input gate **i**, the forget gate **f** and the output gate **o**. At time stamp t, given input v_t and previous LSTM node state h_{t-1}, the LSTM's update mechanism is as follows

$$i_t = \sigma(W_i v_t + U_i h_{t-1} + H_i c_{t-1} + b_i) \quad (2)$$

$$f_t = \sigma(W_f v_t + U_f h_{t-1} + H_i c_{t-1} + b_f) \quad (3)$$

$$c_t = f_t \otimes c_{t-1} + i_i \otimes tanh(W_c v_t + U_c h_{t-1} + b_c) \quad (4)$$

$$o_t = \sigma(W_o v_t + U_o h_{t-1} + H_o c_t + b_o) \quad (5)$$

$$h_t = o_t \otimes tanh(c_t) \quad (6)$$

where σ is the sigmoid function and we use $\otimes$ as the vector element-wise product operator. W_*, U_*, H_* denote the weight parameters and b_* is a bias vector. The LSTM network can propagate the discriminative feature information to the deeper LSTM nodes and forget the noisy information in visual features. In our experiment, we use two layers of LSTM nodes to improve the learning power of the aggregation network for large-scale 3D shape representations.

In the process of recurrent feature aggregation, information is propagated from the first part-based recurrent aggregation unit to the last one. At each time stamp t, which denotes the order of views in a sequence, the LSTM node t maintains the information of the recurrent unit $t-1$ and uses it to output the current fused feature d_t with the input feature v_t. Because the discriminative information may appear anywhere in the different views of a sequence, we adopt an element-wise max-pooling layer to combine the output features of different recurrent aggregation units, $\{d_1, d_2, ..., d_n\}$, which allows for the aggregation of information across all time steps and avoids the final representation of the view sequence biasing towards later time-steps. The final representation of image sequence D_p of the 3D shape x_p can be expressed as

$$D_p^i = max(\{d_{1,i}, d_{2,i}, ..., d_{n,i}\}) \quad (7)$$

where D_p^i denotes the i-th element of the feature vector D_p and $d_{t,i}$ is the i-th element of the view feature d_t.

Viewpoint estimation network

In order to make the aggregated representation incorporate 3D nature of objects, viewpoint estimation network is presented to predict rendered views of different viewpoints through exploiting the hidden 3D shape representation. To be specific, the encoder part transforms the input image sequence into a hidden 3D representation D_p of 3D shape x_p, and then viewpoint estimation network takes the D_p and the target viewpoint as input to output the desired rendered view.

During training, the network is presented with the encoded 3D representation and the image showing the view of the 3D shape together with the target viewpoint. 3D shapes are randomly sampled from the large-scale 3D shape database and unseen viewpoints are randomly selected.

To improve the generative ability, we propose a transformation layer that transforms encoded 3D representation according to different viewpoints θ_i, instead of directly using it to generate images. θ_i is a vector, which consists of two angles: azimuth θ_i^{az} and elevation θ_i^{el}. Firstly, the viewpoint θ_i is processed by two FC layers and the output viewpoint feature $d_{\theta_i}$is 512D. Secondly, the viewpoint feature will generate a transformed coefficient for each dimension of 3D representation D_p and the 3D representation is transformed with element-wise product operation. The transformed representation $Trans^{\theta_i}(D_p)$ can be expressed as

$$Trans^{\theta_i}(D_p) = \sigma(W \cdot d_{\theta_i}) \otimes D_p \quad (8)$$

Here, $\mathbf{W} \in \mathbb{R}^{512\times1024}$ is the parameters of the fully-connected layer. σ is the activation function and we adopt the Leaky ReLU function in this paper. We use $\otimes$ as the vector element-wise product operator.

We use the transformed 3D representation $Trans^{\theta_i}(D_p)$ to generate the image of the target viewpoint with a deconvnet architecture, which consists of 6 deconvolutional layers. We also experimented with deeper networks but did not obtain a significant improvement in performance. The deconvolution layer performs upsampling and convolution, which is opposite to the standard convolution and pooling. For the deconvolutional layers of the network, we use 3×3 filters, 2 stride and 1 padding for the first 5 deconvolutional layers and use 2×2 filters, 1 stride and 1 padding for the last deconvolutional layer.

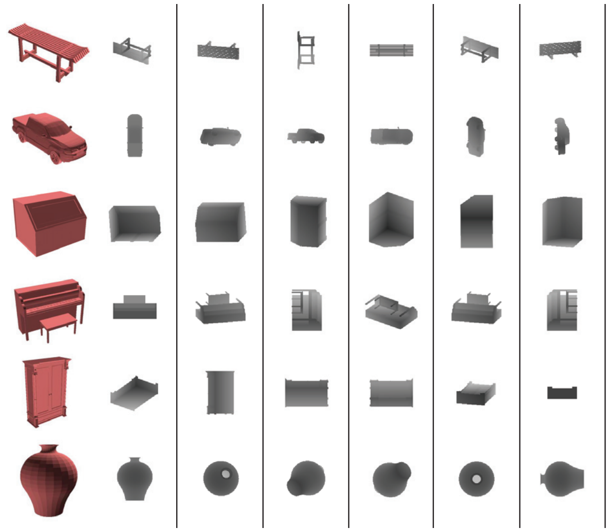


Figure 3: Some examples of rendered views of 3D shapes on ModelNet.

Experiment

In this section, we evaluate the proposed method on different 3D shape benchmarks and competitions. Moreover, we compare the performance of different methods in view missing and occlusion cases. In order to quantify the retrieval performance, we use several evaluation metrics, including mean average precision (MAP) , mean area under P-R curve (AUC) , F-measure and NDCG as defined in (Wu et al. 2015; Savva et al. 2016).

Implementation details

Each 3D shape is rendered to yield views of 256×256 pixels for depth images by placing 24 virtual cameras. The length of multi-view image sequence is 12 to produce the 3D representation, and the remaining rendered views of each 3D shape are used to be estimated.

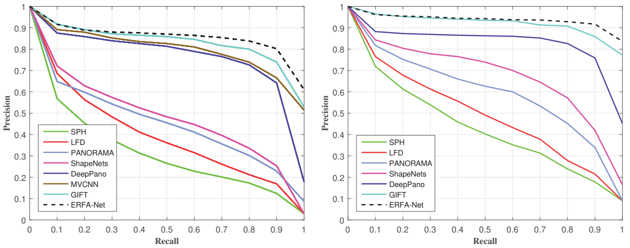
Training. For each 3D shape, we select 12 views according to specific viewpoints as the view fusion sequence and randomly choose one view from unseen rendered views as predicted image. Then a number of input tuples, including the view fusion sequence, predicted image and corresponding viewpoint, are fed to the network. For the multi-view features extracting process, the CNN is fine-tuned on specific 3D shape dataset pre-trained on the ImageNet $1k$ dataset (Deng et al. 2009) and the architecture of CNN adopts GoogLeNet with Batch Normalization, in which the input image is resized to 224×224. We obtain features from pooling layer $pool5/7 \times 7\_s1$ as middle representations. For the recurrent features aggregation, the output dimension of LSTM node is 1024.

Testing. For each 3D shape, the 12 views, according to the fixed viewpoints, are aggregated to obtain a compact 3D representation. The cosine distance is adopted for retrieval tasks.

In our experiment, we use the caffe (Jia et al. 2014) toolbox to implement and train the network. The initial learning rate is 1e-4, which is annealed by 0.5 once encountering the training loss plateaus. The weight decay is set to 5e-4 and the momentum is set to 0.9.

Table 1: The Performance Comparison With State-of-the-art Methods on ModelNet40 and ModelNet10

Methods	ModelNet40		ModelNet10	
	AUC	MAP	AUC	MAP
SPH	34.47%	33.26%	45.97%	44.05%
LFD	42.04%	40.91%	51.70%	49.82%
PANORAMA	45.00%	46.13%	60.72%	60.32%
ShapeNets	49.94%	49.23%	69.28%	68.26%
DeepPano	77.63%	76.81%	85.45%	84.18%
MVCNN	-	80.20%	-	-
GIFT	83.10%	81.94%	92.35%	91.12%
CNN+Concat	81.72%	80.49%	92.06%	91.57%
CNN+Max	80.98%	79.00%	90.16%	89.55%
ERFA-Net	**86.35%**	**85.38%**	**93.64%**	**93.24%**



(a)

(b)

Figure 4: Precision-recall curves on ModelNet40 dataset (a) and ModelNet10 dataset (b).

Retrieval on ModelNet

ModelNet Dataset is composed of 127,915 3D CAD models from 662 categories. It has two subsets, ModelNet40 dataset and ModelNet10 dataset, and we use both for evaluation. The first subset contains 12,311 models and the second contains 4,899 models. We adopt the same training and testing split mentioned in (Wu et al. 2015). We randomly select 100 unique shapes per category from the subset, where the first 80 shapes are used for training and the rest for testing. Some examples of rendered views of 3D shapes on ModelNet are shown in Fig. 3.

The retrieval results of ModelNet dataset are presented in Table 1. Our method is compared against SPH (Kazhdan, Funkhouser, and Rusinkiewicz 2003), LFD (Chen et al. 2003), Panorama (Papadakis et al. 2010), 3D ShapeNets (Wu et al. 2015), DeepPano (Shi et al. 2015), MVCNN (Su et al. 2015) and GIFT (Bai et al. 2016). In addition, several baseline methods are conducted to evaluate the proposed approach. "CNN+Concat" and "CNN+Max" use different schemes to fuse the multi-view representations. "CNN+Concat" concatenates different view-specific features into a final representation of the image sequence. "CNN+MAX" performs max pooling like

Table 2: The Performance Comparison on SHREC16 Normal Dataset

Methods	Micro			Macro			Micro + Macro		
	F-measure	MAP	NCDG	F-measure	MAP	NCDG	F-measure	MAP	NCDG
DB-FMCD-FUL-LCDR	0.472	0.728	0.875	0.203	0.596	0.806	0.338	0.662	0.841
CCMLT	0.391	0.823	0.886	0.286	0.661	0.820	0.339	0.742	0.853
ViewAggregation	0.582	0.829	0.904	0.201	0.711	0.846	0.392	0.770	0.875
MVCNN	0.764	0.873	0.899	**0.575**	0.817	0.880	0.670	0.845	0.890
GIFT	0.689	0.825	0.896	0.454	0.740	0.850	0.572	0.783	0.873
ERFA-Net	**0.776**	**0.889**	**0.919**	0.574	**0.833**	**0.898**	**0.675**	**0.861**	**0.909**

Table 3: The Performance Comparison on SHREC16 Perturbed Dataset

Methods	Micro			Macro			Micro + Macro		
	F-measure	MAP	NCDG	F-measure	MAP	NCDG	F-measure	MAP	NCDG
DB-FMCD-FUL-LCDR	0.413	0.638	0.838	0.166	0.493	0.743	0.290	0.566	0.791
CCMLT	0.246	0.600	0.776	0.163	0.478	0.695	0.205	0.539	0.736
ViewAggregation	0.534	0.749	0.865	0.182	0.579	0.767	0.358	0.664	0.816
MVCNN	0.612	0.734	0.843	0.416	0.662	0.793	0.514	0.698	0.818
GIFT	0.661	0.811	0.889	0.423	0.730	0.843	0.542	0.771	0.866
ERFA-Net	**0.713**	**0.880**	**0.914**	**0.498**	**0.834**	**0.893**	**0.606**	**0.857**	**0.904**

MVCNN. To make fair comparison, the CNN architecture adopts GoogLeNet with Batch Normalization and the experiment setup of view features extraction is the same as that of our method.

From Table 1, the proposed approach outperforms all other state-of-the-art methods remarkably. The performance of MVCNN is limited without utilizing all information of multiple views. Furthermore, shape-information deficiency of the single view may hinder the effect of GIFT. In addition, EFRA-Net performs a clear advantage over “CNN+Concat” and “CNN+Max”. It is shown that our method can effectively incorporate 3D properties to recognize 3D shapes more accurately, compared with the common fusion structures. Fig. 4 shows the comparison of the precision-recall curves of the above methods. In addition, Fig. 5 shows some generated views for 3D shapes. As we can see, the necessary 3D details are correctly estimated.

Retrieval on large-scale 3D dataset

The dataset from SHape REtrieval Contest (SHREC) 2016 is a large-scale 3D shape retrieval track. This dataset contains 51,190 3D shapes over 55 common categories, each subdivided into 204 sub-categories. In our experiment, we adopt the official training and testing split method, where the database is split into three parts, 70% shapes used for training, 10% shapes for validation data and the rest 20% for testing. Besides, two dataset versions are provided, normal dataset where all shapes are consistently aligned, and more challenging perturbed dataset where all shapes are randomly rotated. To keep the comparison fair, we take three types of results including macro, micro and mean of macro and micro, as defined in (Savva et al. 2016).

Table 2 and Table 3 present the performance comparison on the normal and perturbed datasets. Our ERFA-Net is compared to various state-of-the-art methods, including DB-FMCD-FUL-LCDR (Tatsuma and Aono 2016), CCMLT, ViewAggregation, MVCNN (Su et al. 2015) and GIFT (Bai et al. 2016). As we can see, our method achieves state-of-the-art performances on the normal and perturbed datasets.

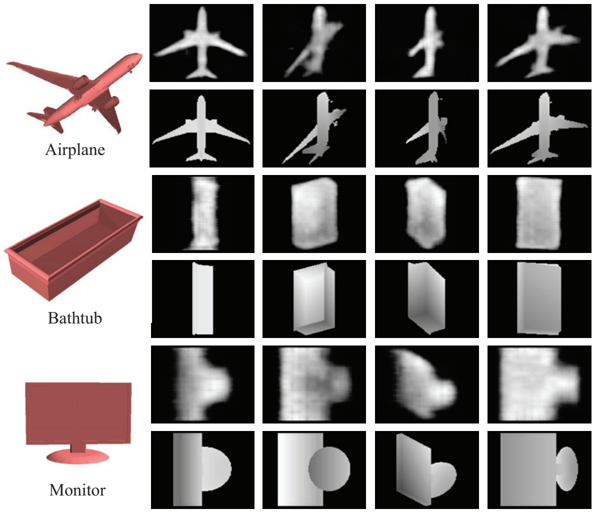


Figure 5: Predictions of the network for “airplane”, “bathtub”, “monitor” classes (top row) and the corresponding ground truth (bottom row) on the testing dataset. The network correctly estimates the views.

There is an interesting phenomenon that our approach leads to a higher improvement on the SHREC16 perturbed dataset than on the ModelNet dataset, compared to MVCNN and GIFT. The possible reasons are two-fold. On the one hand, instead of the well-distributed training and testing split in ModelNet40 or ModelNet10, a huge difference in the number of 3D shapes of different categories is presented in SHREC16 dataset, hindering the training and converging of 3D shape recognition algorithms. On the other hand, the shapes in SHREC16 perturbed dataset are randomly rotated. The performance of some methods, lacking robustness for the spatial transformation of 3D shapes, is restricted. However, our method is insensitive to 3D shapes' transformation owing to the randomly ordered view image sequences. Therefore, our method exhibits encouraging scalability and rotation invariance in large-scale 3D competition.

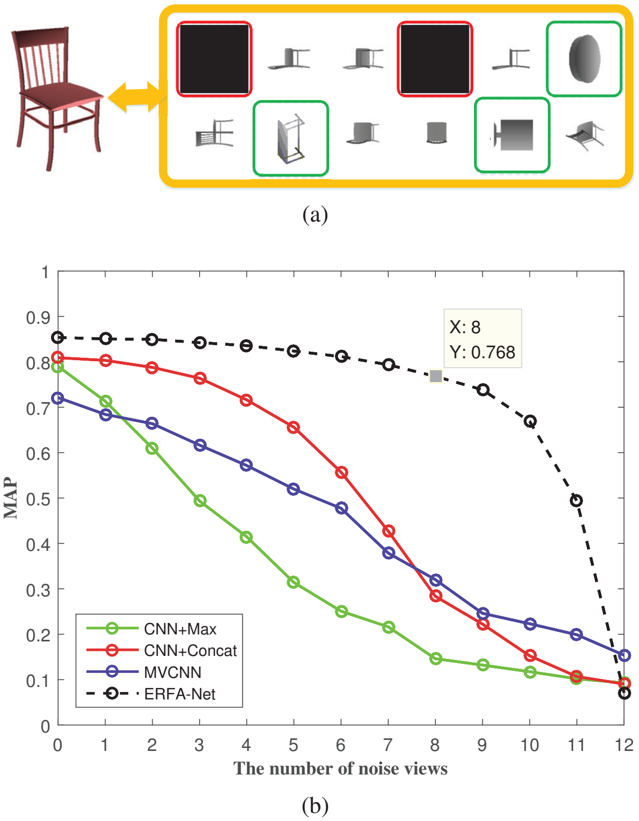
Robust discrimination of encoded representation

Robust discrimination of 3D representations is vital for multi-view 3D shape retrieval, especially for the real scenes. In practical applications like robust-operated 3D shape retrieval, view missing and occlusion caused by other objects can seriously affect the retrieval accuracy. In this experiment, we evaluate the discriminative capacity of our aggregated features on the ModelNet40 dataset by adding some noise, including view missing and occlusion, to the testing view sequences. Then the methods are trained with normal rendered images and tested with noisy view sequences. Some examples of testing images are shown in Fig. 6. For the testing view sequence, some images in the sequence are replaced by noise images, which are meaningless black background images or views of other objects.

We compare different methods with the different number of noisy views, which is shown in Fig. 6. The performances of "CNN+Max" and MVCNN decrease sharply under the noisy cases since the output of the maximum operation may be information of noisy views. Moreover, "CNN+Concat" can not capture robust 3D shape representations from redundancy information of the 3D shape multi-view sequence . As we can see, although the performances suffer from the noise, our aggregated feature still remarkably obtains the state-of-the-art results, which achieves 81.15% MAP on ModelNet40 dataset when 6 images (50%) are polluted. Compared to the noise-free image sequences, the MAP decreases slightly by 8.58% when almost 70% images (8 noise views) are contaminated. This demonstrates that the 3D properties of shapes do give a significant rise to features of 2D images, make the 3D representation maximally reduce the adverse effect of noise and have strongly robust discrimination.

Conclusion

We presented an encoder-decoder recurrent feature aggregation network that learns robust discriminative and generative 3D representations when being trained on the task of estimating views of any unseen viewpoints by exploiting multi-view sequence of 3D shapes. Supervised by generative task, the aggregated feature of the randomly ordered view sequence can effectively incorporate the 3D properties of shapes, making it robust to view missing and occlusion. Therefore, in future work, we will investigate ways to apply the proposed method to noisier real 3D scenes such as robotic-operated 3D shape recognition.


(a)

(b)

Figure 6: (a) is some examples of contaminated view sequence of the "chair" shape, including view missing (red box) and being occluding by other objects (green box). (b) shows the performance on different number of noise views. Our method performs robust discrimination in these noisy cases.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.61472023) and the project of the State Key Laboratory of Software Development Environment (No.SKLSDE-2016ZX-24).

References

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

Bai, S.; Bai, X.; Zhou, Z.; Zhang, Z.; and Latecki, L. J. 2016. Gift: A real-time and scalable 3d shape search engine. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5023–5032.

Bonaventura, X.; Guo, J.; Meng, W.; Feixas, M.; Zhang, X.; and Sbert, M. 2015. 3d shape retrieval using viewpoint informationtheoretic measures. *Computer Animation and Virtual Worlds* 26(2):147–156.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; and Su, H. 2015. Shapenet: An information-rich 3d model repository. *Computer Science*.

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

Choy, C. B.; Xu, D.; Gwak, J. Y.; Chen, K.; and Savarese, S. 2016. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision*, 628–644.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 248–255. IEEE.

Dosovitskiy, A.; Springenberg, J. T.; and Brox, T. 2015. Learning to generate chairs with convolutional neural networks. In *Computer Vision and Pattern Recognition*, 1538–1546.

Fang, Y.; Xie, J.; Dai, G.; Wang, M.; Zhu, F.; Xu, T.; and Wong, E. 2015. 3d deep shape descriptor. In *Computer Vision and Pattern Recognition*, 2319–2328.

Gao, Y.; Wang, M.; Tao, D.; Ji, R.; and Dai, Q. 2012. 3-d object retrieval and recognition with hypergraph analysis. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society* 21(9):4290–303.

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

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. 580–587.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 448–456.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, 675–678.

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

Kazhdan, M.; Funkhouser, T.; and Rusinkiewicz, S. 2003. Rotation invariant spherical harmonic representation of 3 d shape descriptors. In *Symposium on geometry processing*, volume 6, 156–164.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems*, 1097–1105.

Kulkarni, T. D.; Whitney, W.; Kohli, P.; and Tenenbaum, J. B. 2015. Deep convolutional inverse graphics network. 71(2):2539–2547.

Leng, B.; Liu, Y.; Yu, K.; Zhang, X.; and Xiong, Z. 2016. 3d object understanding with 3d convolutional neural networks. *Information Sciences* 366:188–201.

Li, F. F., and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 524–531.

Liu, Y.; Li, H.; and Wang, X. 2017. Rethinking feature discrimination and polymerization for large-scale recognition. *arXiv preprint arXiv:1710.00870*.

Papadakis, P.; Pratikakis, I.; Theoharis, T.; and Perantonis, S. 2010. Panorama: A 3d shape descriptor based on panoramic views for unsupervised 3d object retrieval. *International Journal of Computer Vision* 89(2-3):177–192.

Qi, C. R.; Su, H.; Niebner, M.; Dai, A.; Yan, M.; and Guibas, L. J. 2016. Volumetric and multi-view cnns for object classification on 3d data. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5648–5656.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proc. CVPR*.

Razavian, A. S.; Azizpour, H.; Sullivan, J.; and Carlsson, S. 2014. Cnn features off-the-shelf: An astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops*, 512–519.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2017. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(6):1137.

Savva, M.; Yu, F.; Su, H.; Aono, M.; Chen, B.; Cohen-Or, D.; Deng, W.; Su, H.; Bai, S.; and Bai, X. 2016. Shrec'16 track large-scale 3d shape retrieval from shapenet core55. In *EG 2016 workshop on 3D Object Recognition*.

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

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

Tatsuma, A., and Aono, M. 2016. Food image recognition using covariance of convolutional layer feature maps. *IEICE TRANSACTIONS on Information and Systems* 99(6):1711–1715.

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 *Computer Vision and Pattern Recognition*, 1912–1920.

Xie, J.; Fang, Y.; Zhu, F.; and Wong, E. 2015. Deepshape: Deep learned shape descriptor for 3d shape matching and retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1275–1283.

Yan, X.; Yang, J.; Yumer, E.; Guo, Y.; and Lee, H. 2016. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Advances in Neural Information Processing Systems*, 1696–1704.

Yi, L.; Su, H.; Guo, X.; and Guibas, L. 2017. SyncSpecCNN: Synchronized spectral cnn for 3d shape segmentation. In *Proc. CVPR*.

Zhou, B.; Garcia, A. L.; Xiao, J.; Torralba, A.; and Oliva, A. 2014. Learning deep features for scene recognition using places database.

Zhu, Z.; Wang, X.; Bai, S.; and Yao, C. 2014. Deep learning representation using autoencoder for 3d shape retrieval. In *International Conference on Security, Pattern Analysis, and Cybernetics*, 279–284.