Siamese CNN-BiLSTM Architecture for 3D Shape Representation Learning

Guoxian Dai[1,2,4], **Jin Xie**[1,2], **Yi Fang**[1,2,3,*]
[1] NYU Multimedia and Visual Computing Lab
[2] Dept. of ECE, NYU Abu Dhabi, UAE
[3] Dept. of ECE, NYU Tandon School of Engineering, USA
[4] Dept. of CSE, NYU Tandon School of Engineering, USA
guoxian.dai@nyu.edu, jin.xie@nyu.edu, yfang@nyu.edu

Abstract

Learning a 3D shape representation from a collection of its rendered 2D images has been extensively studied. However, existing view-based techniques have not yet fully exploited the information among all the views of projections. In this paper, by employing recurrent neural network to efficiently capture features across different views, we propose a siamese CNN-BiLSTM network for 3D shape representation learning. The proposed method minimizes a discriminative loss function to learn a deep nonlinear transformation, mapping 3D shapes from the original space into a nonlinear feature space. In the transformed space, the distance of 3D shapes with the same label is minimized, otherwise the distance is maximized to a large margin. Specifically, the 3D shapes are first projected into a group of 2D images from different views. Then convolutional neural network (CNN) is adopted to extract features from different view images, followed by a bidirectional long short-term memory (LSTM) to aggregate information across different views. Finally, we construct the whole CNN-BiLSTM network into a siamese structure with contrastive loss function. Our proposed method is evaluated on two benchmarks, ModelNet40 and SHREC 2014, demonstrating superiority over the state-of-the-art methods.

1 Introduction

Recently, due to the advancement of both computer hardware and software techniques, large amounts of 3D models are widely available free on-line, such as 3D Warehouse, Cults 3D, and Pinshape, as well as in our daily lives, such as 3D printing, gaming, and cartoon. And such massive 3D models make the problem of 3D shape retrieval become a vital issue in computer vision. The model-based retrieval is a straightforward way to search for desired 3D shapes, which is given a query 3D model and returns similar models. Generally, each shape is represented as a high dimensional vector, called shape descriptor. The similarities of different models are determined by the distances of their corresponding shape descriptors.

The key problem for model-based 3D shape retrieval is how to extract robust and deformation-invariant shape descriptor. The problem is quite challenging because 3D shapes have complex geometric structures as well as all kinds of variations, such as noise and deformation. The shape descriptor should correctly represent 3D shape regardless of all kinds of variations. In the past few decades, lots of methods were proposed for model-based 3D shape retrieval [Chen *et al.*, 2003; Litman *et al.*, 2014; Xie *et al.*, 2016]. And the existing methods could be roughly classified into three groups: 1) diffusion-based methods characterize 3D shape by employing heat diffusion on the meshed surface, which could not handle large deformations. 2) projection-based methods render 3D shape into a number of different view images. However, they either need exhaustively comparing each pair of views for different shapes, which is quite time-consuming, or simply conduct max-pooling to get one compact descriptor, which will lose quite a lot of information from other views. 3) voxelization-based methods convert 3D shape into 3D grid for 3D convolutional operations, however due to GPU memory limitations, the voxelized 3D shape could only have very low resolutions for learning.

In this work, we revisit the projection-based methods. There are generally three steps for the projection-based methods: 1) multi-view rendering, projecting 3D shape into a group of 2D images; 2) view-based feature extraction. Classical 2D image features could be extracted, such as SIFT. In addition, due to the great success of deep convolutional neural network (CNN) [Krizhevsky *et al.*, 2012; He *et al.*, 2015] in feature learning recently, deep CNN could also be applied to extract robust visual features from the rendered images [Su *et al.*, 2015; Bai *et al.*, 2016]; 3) view-based feature comparison, the key step for projection-based methods. One could exhaustively compare each pair of views between two shapes and choose the distance of the most similar pair as the similarity for two shapes. However, the main problem for exhaustive comparison is that the computation cost increases drastically, as the number of views increases. In addition, one could also fuse features from different views to get one compact shape

*Corresponding author

descriptor for comparison. The most straightforward way for fusion is pooling, such as max-pooling and average-pooling. However, both are problematic. max-pooling will lose lots of information from other views, and average-pooling is only a linear combination of different view, which could not comprehensively characterize the 3D shape.

In this paper, by employing convolutional neural network for visual feature extraction and recurrent neural network for aggregating information across different views, we propose a siamese CNN-BiLSTM network for 3D shape representation learning. Our proposed method minimizes a discriminative loss function to learn a deep nonlinear transformation, mapping 3D shapes from the original space into a nonlinear feature space. In the transformed space, the distances of shapes with the same label are encouraged to be as small as possible, otherwise the distances are maximized to a large margin. Specifically, the 3D shape is first projected into a group of 2D images, which are treated as a sequence. Second, we use CNN to extract visual features from different view images, followed by a bidirectional recurrent neural network (RNN), particularly, long short-term memory (LSTM), to aggregate information across different views. Then, the outputs of all BiLSTM cells are passed through an average-pooling across different views to form one compact representation. Finally, we construct the whole CNN-BiLSTM network into a siamese structure with a discriminative loss function to minimize the positive pairwise (with the same label) distance and maximize the negative pairwise (with different labels) distance of the deep learned shape descriptor. In addition, we verify our proposed method on two large-scale benchmarks, ModelNet40 [Wu *et al.*, 2015] and SHREC 2014 [Li and Lu, et al., 2014], and the experimental results on both benchmarks demonstrate the effectiveness of our proposed method.

The main contribution of our proposed method is that we employ BiLSTM to capture the geometric information of 3D shape across different views. Compared to previous projection-based methods, either simply max-pooling across different views [Su *et al.*, 2015], which will lose quite a lot of information, or exhaustively comparing among different views [Bai *et al.*, 2016], which will involve lots of computation, our proposed method is more efficient to capture information among different views. Instead of using one directional LSTM, which only considers the dependencies of previous views, we adopt BiLSTM, considering the dependencies of both previous and future views, which could generate a more comprehensive representation for 3D model. In addition, most of the existing works about LSTM focused on sequence prediction and classification tasks, we argue that LSTM is also effective for representation learning by constructing the CNN-BiLSTM network into a siamese structure. The proposed siamese CNN-BiLSTM network minimizes a discriminative loss function to learn a deep nonlinear transformation, so that the deep learned shape descriptors of 3D models from the same class are guaranteed to be as similar as possible, otherwise as dissimilar as possible. Thus, our deep learned shape descriptor could more robustly characterize 3D shape, insensitive to deformations, noises or defects, *etc*.

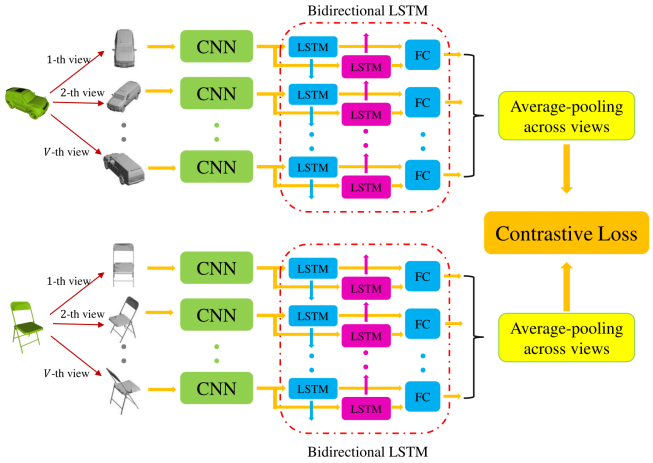

Figure 1: Detailed framework of our proposed method. Each input shape is rendered on V different views. All the V projected images are treated as a sequence. They are first passed through CNN to extract visual features. Then bidirectional LSTM is adopted to aggregate information across all the views, from both the forward and backward directions. FC denotes fully connected layer, summing information from both directions. The outputs from all the BiLSTM cells are passed through an average-pooling across different views. Finally, we construct the CNN-BiLSTM network into a siamese structure with the contrastive loss function.

2 Related Work

The related works are introduced from two aspects, model-based 3D shape retrieval and metric learning. Next we will discuss the representative works accordingly.

3D shape retrieval Existing works about model-based 3D shape retrieval could roughly be divided into three categories, diffusion-based, projection-based and voxelization-based methods.

Diffusion-based method The geometric structure of 3D shape could be characterized by heat diffusion or probability distribution on the meshed surface, such as heat kernel signature (HKS) [Sun *et al.*, 2009], and scale-invariant heat kernel signature (SIHKS) [Bronstein and Kokkinos, 2010]. [Litman *et al.*, 2014] applied supervised dictionary learning with HKS and SIHKS to describe 3D shape. On top of the distribution of HKS, [Xie *et al.*, 2016] adopted deep neural network to learn deformation-invariant shape descriptor. Besides, [Bu *et al.*, 2014] adopted deep belief network with BoW of HKS to learn high-level shape descriptors.

Projection-based method Apart from diffusion-based method, 3D shapes could also be projected into a group of images from different views, thus classical 2D image features or learning techniques could be applied to describe 3D shapes [Su *et al.*, 2015; Bai *et al.*, 2016]. [Bai *et al.*, 2016] optimized the view-based methods and proposed a real-time retrieval engine. In addition, [Su *et al.*, 2015] proposed a multi-view CNN for 3D shape recognition, by using deep CNN to extract visual features from different view images and employing max-pooling across views to learn one compact shape descriptor.

Voxelization-based method Both diffusion and projection based methods could not directly deal with 3D shape, they either need to extract raw features first, or project 3D shape into 2D images. Recently, lots of researchers are seeking to directly use CNN on 3D shape [Wu *et al.*, 2015; Qi *et al.*, 2016; Maturana and Scherer, 2015]. [Wu *et al.*, 2015] voxelized the 3D shape into 3D grids and trained a generative model for 3D shape recognition using convolutional deep belief network. Similarly, [Maturana and Scherer, 2015] proposed a supervised 3D CNN on voxelized 3D representation, which could perform 3D object recognition in real-time. [Qi *et al.*, 2016] made comparisons between voxelization-based CNN and multi-view projection-based CNN, and proposed two improved versions of volumetric CNN.

Metric learning Most of the traditional metric learning methods focused on learning a linear transformation to map samples from the original space into a new feature space, which may not be capable to deal with samples on the nonlinear manifold. Inspired by the great success of deep learning [Krizhevsky *et al.*, 2012; He *et al.*, 2015], deep metric learning was proposed recently, which learns a much more powerful deep nonlinear transformation [Chopra *et al.*, 2005; Hu *et al.*, 2014; Li and Tang, 2015; Hoffer and Ailon, 2015]. [Chopra *et al.*, 2005] proposed to use a siamese network to learn similarity metric for face verification. [Hu *et al.*, 2014] adopted similar structure by imposing a marginal distance between similar pair and non-similar pair. Different from above methods with pairwise training examples, [Hoffer and Ailon, 2015] used a triplet structure and achieved better performance. Instead of partially selecting pairwise samples like [Chopra *et al.*, 2005; Hoffer and Ailon, 2015], [Oh Song *et al.*, 2016] considered all the possible pairs in a minibatch for metric learning.

3 Method

In this paper, we propose a siamese CNN-BiLSTM network for 3D shape representation learning. Fig. 1 shows the detailed framework of our proposed method. First, each 3D shape is rendered on multiple different views based on predefined virtual cameras. Then, the projected images are treated as a sequence for training an end-to-end siamese CNN-BiLSTM network. CNN is used to extract visual features from different view images, then bidirectional LSTM is employed to efficiently capture informations across different views from both forward and backward directions. The outputs of all the BiLSTM cells are passed through an average-pooling across different views to form one compact presentation. Finally, we construct the CNN-BiLSTM network into a siamese structure with the contrastive loss. The proposed method minimizes the contrastive loss to learn a deep nonlinear transformation, mapping 3D shapes from the original space into a new feature space. In the transformed space, the distance for positive pairwise examples is minimized, and the distance for negative pairwise examples is maximized to a large margin.

3.1 Multiple-view Rendering

Since it is very difficult to directly conduct convolutional operation on the meshed surface of 3D model, we first project 3D shape into a group of 2D images. The key problem of multi-view rendering is how to set up camera locations. Generally, more number of views could characterize 3D shape better, however, on the other hand, it also involves more computation cost. Thus, there is trade-off between the number of rendered views and computation cost. Fortunately, most of the modern on-line repositories, such as 3D ShapeNet [Wu *et al.*, 2015] and 3D Warehouse, follow one important assumption that all the models are stored upright as default. Such default assumption makes it much easier to set up the camera locations. We follow the same camera setting in [Su *et al.*, 2015], 12 cameras are put in the horizontal plane by every 30°, pointing to the centroid of the model. In addition, before rendering, all the 3D models are rescaled to the same uniform size to avoid scale-variation. Based on the upright assumption, 12 cameras could capture most information of the 3D model. The rendered images could form a sequence, according to the order of camera locations, from 0° to 359°, step by 30°.

3.2 Siamese CNN-BiLSTM

The main problem for projection-based methods is how to efficiently aggregate information among different views. Traditional methods are either simply max-pooling across different views [Su *et al.*, 2015], losing lots of information, or exhaustively comparing each pair of views between two shapes with high computation complexities. To this end, we propose a siamese CNN-BiLSTM network for 3D shape representation learning, which could efficiently aggregate information across different views meanwhile guarantee the deep learned shape descriptor as discriminative as possible. Specifically, we first use deep CNN to extract visual features from the rendered images, then adopt bidirectional recurrent neural network, particularly BiLSTM, to aggregate information from each view of the rendered images. In addition, the outputs of all BiLSTM cells are passed through an average-pooling across different views to form one compact representation. Finally, we construct the CNN-BiLSTM network into a siamese structure with the contrastive loss. Through minimizing the contrastive loss, our proposed method could minimize the intra-class distance and maximize the inter-class of the deep learned shape descriptors.

CNN For the CNN part network, we use AlexNet [Krizhevsky *et al.*, 2012], which includes 5 convolution layers and 3 fully connected layers. The AlexNet is pre-trained on ImageNet [Russakovsky *et al.*, 2015] and we fine-tune AlexNet with the rendered image data as a single image classification task. The label of the rendered image is the same as that of its source shape. During training, all the image data are shuffled and treated independently without considering sequential information. After fine-tuning, we remove the last two fully connected layers and connect the remaining parts with LSTM cells.

BiLSTM LSTM is an improved variant of RNN to solve the long term dependency problem. Fig. 2 shows the structure of LSTM cell in our proposed method. Compared to the traditional RNN, LSTM added three control gates, namely, input gate which controls whether to consider the current input signal and previous hidden state, forget gate which controls whether to forget the previous memory cell state, output gate which controls whether to transfer the memory cell to hidden state. All the three control gates use sigmoid transfer function $\sigma(x) = 1/(1+e^{-1})$ to rescale the signal into $[0, 1]$. While the modulate gate uses hyperbolic tangent function $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$ to rescale the signal into $[-1, 1]$. By employing those control gates, LSTM could explicitly control when to forget previous hidden state or input the current signal.

Fig. 1 shows the structure of BiLSTM in our proposed method. For each time step, the BiLSTM considers the dependency of not only the previous steps, but also the future steps. During the rendering part, each shape is rendered on V different views, and the visual feature extracted from the v-th view is denoted as $x_v \in \mathbb{R}^{N\times 1}$, where N is the size of visual feature. All the features of V different views could form an input sequence, $x = \{x_1, x_2, \cdots, x_V\}$, which is passed through the LSTM cell array. The BiLSTM cell updated at the v-th view depends on both the forward and backward directions. The forward equation is shown in Eq. 1. And the backward equation could be derived similarly by replacing $\rightarrow$ with $\leftarrow$.

$$\begin{aligned}
\overrightarrow{i}_v &= \sigma(\overrightarrow{W}_i[x_v, \overrightarrow{h}_{v-1}] + \overrightarrow{b}_i) \\
\overrightarrow{f}_v &= \sigma(\overrightarrow{W}_f[x_v, \overrightarrow{h}_{v-1}] + \overrightarrow{b}_f) \\
\overrightarrow{o}_v &= \sigma(\overrightarrow{W}_o[x_v, \overrightarrow{h}_{v-1}] + \overrightarrow{b}_o) \\
\overrightarrow{g}_v &= \tanh(\overrightarrow{W}_c[x_v, \overrightarrow{h}_{v-1}] + \overrightarrow{b}_c) \\
\overrightarrow{c}_v &= \overrightarrow{f}_v \odot \overrightarrow{c}_{v-1} + \overrightarrow{i}_v \odot \overrightarrow{g}_v \\
\overrightarrow{h}_v &= \overrightarrow{o}_v \odot \tanh(\overrightarrow{c}_v)
\end{aligned} \tag{1}$$

where $\rightarrow$ and $\leftarrow$ denote the forward and backward directions, respectively. i_v, f_v, o_v, g_v and h_v denote the input gate, forget gate, output gate, modulate gate and hidden state at the v-th view, respectively. While parameters W_i, W_f, W_o and W_c denote the weight matrix for those above gates, respectively. $\odot$ denotes the element-wise multiplication. The overall output $\mathcal{H}_v$ of BiLSTM cell at the v-th view would be the sum of both forward direction $\overrightarrow{h}_v$ and backward direction $\overleftarrow{h}_v$, formulated as follows,

$$\mathcal{H}_v = \overrightarrow{W}_h \overrightarrow{h}_v + \overleftarrow{W}_h \overleftarrow{h}_v + b_h. \tag{2}$$

By employing BiLSTM, our proposed method considers the dependency of each view upon all the other views, which could more comprehensively characterize the 3D shape. The outputs of all BiLSTM cells are passed through an average-pooling across different views to form one compact representation.

Contrastive loss In order to guarantee our deep learned shape descriptor as discriminative as possible, we construct the CNN-BiLSTM network into a siamese structure with the contrastive loss function. Through minimizing the contrastive loss function, our proposed method learns a deep nonlinear transformation, which maps 3D shape from the original space into a new feature space. In the transformed space, the distances of the positive pairs are minimized, and the distances of negative pairs are maximized to a large margin.

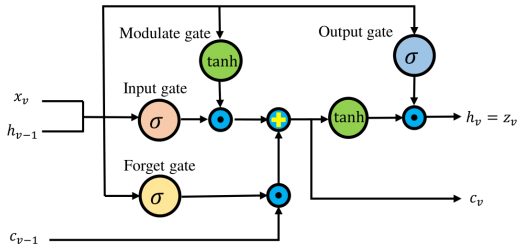

Figure 2: A diagram of LSTM cell in our work.

We denote the training set of 3D shapes as $S = \{(x_1, y_1), (x_2, y_2), \cdots, (x_i, y_i), \cdots\}$, where x_i denotes the i-th 3D shape in S and y_i denotes its corresponding label. Assume the overall transfer function for the CNN-BiLSTM network is $f : x \rightarrow f(x)$. Given a pair of 3D shapes x_i and x_j passing through the proposed CNN-BiLSTM network, their Euclidean distance $d(x_i, x_j)$ in the transformed space can be defined as follows,

$$d(x_i, x_j) = \|f(x_i) - f(x_j)\|_2^2. \tag{3}$$

In order to learn a more discriminative shape descriptor in the transformed space, the distance of the 3D shapes from the same class should be encouraged as small as possible, while the distance of the 3D shapes from different classes should be encouraged as large as possible. The overall contrastive loss function $L(x_i, x_j)$ is formulated as follows,

$$L(x_i, x_j) = Y d(x_i, x_j) + (1-Y)\max\{0, h - d(x_i, x_j)\} \tag{4}$$

where $Y = 1$, if $y_i = y_j$, otherwise, $Y = 0$. The first term aims to minimize the distance of 3D shapes from the same class, while the second term is a hinge loss, which aims to maximize the distance of 3D shapes from different classes to a large margin h.

The whole siamese network could be trained with back propagation using stochastic gradient descent. After the network is trained, each 3D shape could be forwarded through the CNN-BiLSTM network to generate one compact representation, which is used to compare the similarities among different shapes.

4 Experimental Results

Our proposed method is evaluated on two large-scale benchmarks, Princeton ModelNet [Wu *et al.*, 2015] and SHREC 2014 [Li and Lu, et al., 2014]. We first compare the experimental results between siamese CNN and siamese CNN-BiLSTM to demonstrate the effectiveness of our proposed method, that BiLSTM could efficiently aggregate information across different views. In addition, we also compare our proposed method with the state-of-the-art methods to demonstrate the superiority of our proposed method. The evaluations are based on standard criterion, such as precision-recall

curve for visualizing the retrieval performance, and other quantitative measurements, including nearest neighbor (NN), first tier (FT), second tier (ST), discounted cumulative gain (DCG) and mean average precision (mAP).

4.1 Implementation Details

In this subsection, we briefly introduce the implementation details. For the multi-view rendering part, we choose 12 different views at the horizontal plane towards to the centroid of the 3D shape, step by every 30°. For the CNN part, we use AlexNet. The network is first fine-tuned with single image classification task by setting each rendered image with a class label. After fine-tuning, the last two fully connected layers are removed, and the left parts are connected to BiLSTM. The size of the hidden state for LSTM is set to 512. The outputs of all BiLSTM cells are passed through an average-pooling to generate one compact representation, which is followed by two fully connected layers with the sizes of 200, 200. The whole CNN-BiLSTM network is constructed into a siamese structure with the contrastive loss function and the margin h is set to 10. In addition, our proposed method is implemented using Caffe with a single GPU, Nvidia Tesla K80. The batch size is set to 360, which is 30 sequences; the weight decay rate is set to 0.0005; the momentum rate is set to 0.1; the base learning rate is set to 0.001.

4.2 Retrieval on ModelNet40

The ModelNet benchmark [Wu *et al.*, 2015] includes more than 15000 3D models, which are classified into 662 categories. A subset of 40-comment classes, ModelNet40, is manually cleaned up by deleting the irrelevant objects. ModelNet40 contains 12311 models and the number of models for each class is not equal. For our experiments, we follow the same training and testing splitting in [Wu *et al.*, 2015; Su *et al.*, 2015], by selecting 100 shapes for each class, 80 for training and 20 for testing.

For retrieval, we use the cosine similarity to compare the similarities of the deep learned descriptors for shapes from the testing set, which provides slightly better performance compared to the $L2$ distance. Given each query shape from the testing set, the similar models from the remaining test data are returned according to the similarity ranking order.

To demonstrate the effectiveness of our proposed method, we first compare the experimental results between the proposed siamese CNN-BiLSTM with two baseline methods, siamese CNN and MVCNN [Su *et al.*, 2015]. For siamese CNN, we remove the LSTM parts and the outputs of CNN are directly passed through an average-pooling, then construct it into a siamese structure. For MVCNN [Su *et al.*, 2015], we reimplemented it with AlexNet [Krizhevsky *et al.*, 2012] as a fair comparison, which is also used in our proposed method. And the max-pooling is located at "*fc6*" layer.

Fig. 3 shows the precision-recall curve of MVCNN, siamese CNN and siamese CNN-BiLSTM on ModelNet40. The higher of the curve indicates better performance. As we can see in Fig. 3, the proposed siamese CNN-BiLSTM could outperform the two baseline methods. Without LSTM, the siamese CNN could already achieve reasonable performance with mAP=0.760. By employing BiLSTM, the proposed siamese CNN-BiLSTM could further increase the mAP to 0.833, with the gain of 0.07. The experimental results demonstrate that BiLSTM could effectively aggregate information across different views to generate a more compact shape descriptor. The MVCNN simply conducts max-pooling across different views, which will lose lots of information and achieve inferior performance compared with the proposed method. The mAP is only 0.754 (reimplemented).

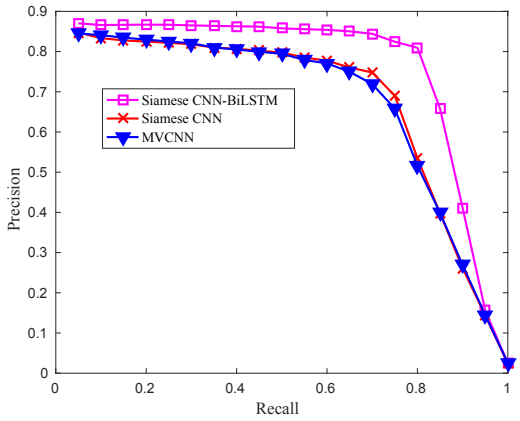

Figure 3: Precision-recall curve of siamese CNN-BiLSTM, siamese CNN and MVCNN on ModelNet40.

In addition, we also compare our proposed method with the state-of-the-art methods, such as light field descriptor (LFD) [Chen *et al.*, 2003], spherical harmonic descriptor (SHD) [Kazhdan *et al.*, 2003], 3D shapeNet [Wu *et al.*, 2015], MVCNN [Su *et al.*, 2015] and GIFT [Bai *et al.*, 2016]. LFD [Chen *et al.*, 2003] rendered 3D shape into a serial of different views, and extract hand-craft features to exhaustively search for the best alignment. Instead of exhaustively searching, [Su *et al.*, 2015] used deep convoluational neural network to extract features from different views and conducted max-pooling across different views to get one shape descriptor. [Bai *et al.*, 2016] also used CNN to extract features from different views, and proposed a fast approach for multi-view matching.

The comparison results are shown in Table. 2. As we can see in Table. 2, our proposed method could outperform the state-of-the-art methods. It is noted that our proposed method doesn't need any post-processing, such as metric learning to improve performance like [Su *et al.*, 2015], nor exhaustively comparing each pair of views for matching different shapes since LSTM could efficiently aggregate information across different views to produce one compact global shape descriptor for retrieval.

Fig. 4 shows some retrieved examples on ModelNet40. The query shapes are listed at the left first column, which include 8 categories, namely airplane, guitar, person, lamp, cup, bed, book shelf and vase. The top 12 retrieved shapes are listed on the right side, based on their retrieved ranking order. All the mistaken objects are marked with red box. As we can see from Fig. 4, for airplane, guitar, person and lamp, all the top 12 retrieved models are correct. While for cup, bed, book shelf and vase, our proposed method provides sev-

Methods	mAP
LFD [Chen *et al.*, 2003]	0.409
SHD [Kazhdan *et al.*, 2003]	0.333
ShapeNet [Wu *et al.*, 2015]	0.492
MVCNN [Su *et al.*, 2015]	0.802
GIFT [Bai *et al.*, 2016]	0.819
Siamese CNN-BiLSTM	**0.833**

Table 1: Performance comparison with the state-of-the-art methods on ModelNet40.

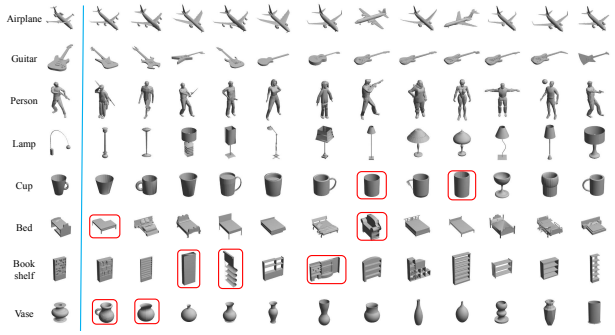

Figure 4: Illustration of the retrieved examples on ModelNet40. The query models are listed at the left first column, and the retrieved shapes are listed on the right side by their ranking order.

eral irrelevant objects. The mistakes are reasonable, due to the appearance-similarities among those 3D models. For example, the book shelf and door look very similar; the vase and cup also look like each other very much.

4.3 Retrieval on SHREC 2014

In this subsection, we evaluate our proposed method on SHREC 2014 dataset [Li and Lu, et al., 2014]. The models of SHREC 2014 come from various benchmarks, such as the Princeton Shape Benchmark (PSB) [Shilane *et al.*, 2004] and the Toyohashi Shape Benchmark (TSB) [Tatsuma *et al.*, 2012]. There are 8987 shapes overall in SHREC 2014, which are classified into 171 categories. The number of shapes for each class is not equal, for example, there are more than 300 shapes for airplane, however, only 2 shapes for alarm clock. It is noted that the whole dataset of SHREC 2014 benchmark is used for testing. As for our experiments, we randomly split the shapes of each group into two halves equally as training and testing. To avoid variation, we repeat the experiments for 3 times to get the averaged result.

Fig. 5 shows the precision-recall curve of our proposed method against the state-of-the-art methods, including DASD (DA)[Xie *et al.*, 2016], KVLAD (KV) [Li and Lu, et al., 2014], DBNAA_DERE (DB) [Li and Lu, et al., 2014], MR-D1SIFT (MR) [Li and Lu, et al., 2014] and ZFDR (ZF) [Li and Lu, et al., 2014]. As we can see from Fig. 5, before the recall value reaches around 0.85, the precision value of our proposed method stays higher than other methods, and then it drops below other methods. Generally, people are more curious about the earlier retrieved objects than latter ones, which indicates our proposed method has better retrieval performance. Except for precision-recall curve, We also calculate standard evaluation criterion against the state-of-the-art methods, including NN, FT, ST, DCG and mAP, as shown in Table. 2. Except for NN, our proposed method could outperform the state-of-the-art methods in all the other criterion with a large margin.

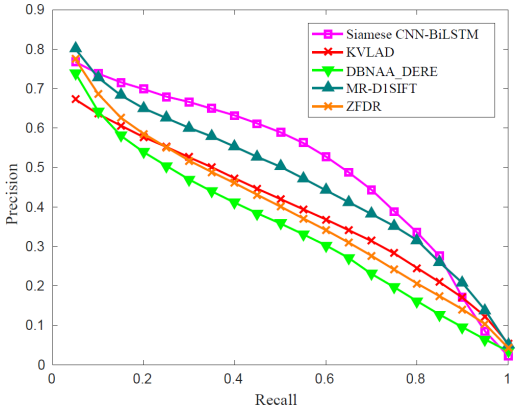

Figure 5: Performance comparison of precision-recall curves on SHREC 2014.

methods	NN	FT	ST	DCG	mAP
KV	0.605	0.413	0.546	0.746	0.396
DB	0.817	0.355	0.464	0.731	0.344
MR	0.856	0.465	0.578	0.792	0.464
ZF	0.838	0.386	0.501	0.757	0.387
LC	0.864	0.528	0.661	0.823	0.541
DA	**0.897**	0.401	0.503	0.790	-
Proposed	0.812	**0.617**	**0.730**	**0.831**	**0.644**

Table 2: Comparison with the state-of-the-art methods on SHREC 2014.

5 Conclusions

In this work, by employing deep CNN for visual feature extraction and RNN for capturing feature across different views, we develop a siamese CNN-BiLSTM network for 3D shape representation learning. Specifically, we first render 3D shape into a group of 2D images, which could form a sequence according to the rendered order. Then CNN is adopted to extract visual features from different view images, followed by a bidirectional recurrent neural network, particularly LSTM, to efficiently aggregate information from different view images. The outputs of all BiLSTM cells are passed through an average-pooling to form one compact representation. Finally, we construct the CNN-BiLSTM network into a siamese structure with the contrastive loss function. Through minimizing the contrastive loss, our proposed method could minimize the intra-class variations and maximize inter-class variations of the deep learned shape descriptors. Finally, our proposed method is evaluated on two large-scale benchmarks, ModelNet40 and SHREC 2014. The experimental results demonstrate the superiority of our proposed method over the state-of-the-art methods.

References

[Bai *et al.*, 2016] Song Bai, Xiang Bai, Zhichao Zhou, Zhaoxiang Zhang, and Longin Jan Latecki. Gift: A real-time and scalable 3d shape search engine. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[Bronstein and Kokkinos, 2010] Michael M Bronstein and Iasonas Kokkinos. Scale-invariant heat kernel signatures for non-rigid shape recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1704–1711. IEEE, 2010.

[Bu *et al.*, 2014] Shuhui Bu, Zhenbao Liu, Junwei Han, Jun Wu, and Rongrong Ji. Learning high-level feature by deep belief networks for 3-d model retrieval and recognition. *IEEE Transactions on Multimedia*, 16(8):2154–2167, 2014.

[Chen *et al.*, 2003] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. On visual similarity based 3d model retrieval. In *Computer graphics forum*, volume 22, pages 223–232. Wiley Online Library, 2003.

[Chopra *et al.*, 2005] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 539–546. IEEE, 2005.

[He *et al.*, 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[Hoffer and Ailon, 2015] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015.

[Hu *et al.*, 2014] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1875–1882, 2014.

[Kazhdan *et al.*, 2003] Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. Rotation invariant spherical harmonic representation of 3 d shape descriptors. In *Symposium on geometry processing*, volume 6, pages 156–164, 2003.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[Li and Lu, et al., 2014] Bo Li and Yijuan Lu, et al. SHREC'14 track: Large scale comprehensive retrieval track benchmark. In *Eurographics Workshop on 3D Object Retrieval, Strasbourg, France*, 2014.

[Li and Tang, 2015] Zechao Li and Jinhui Tang. Weakly supervised deep metric learning for community-contributed image retrieval. *IEEE Transactions on Multimedia*, 17(11):1989–1999, 2015.

[Litman *et al.*, 2014] Roee Litman, Alex Bronstein, Michael Bronstein, and Umberto Castellani. Supervised learning of bag-of-features shape descriptors using sparse coding. In *Computer Graphics Forum*, volume 33, pages 127–136. Wiley Online Library, 2014.

[Maturana and Scherer, 2015] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 922–928. IEEE, 2015.

[Oh Song *et al.*, 2016] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4004–4012, 2016.

[Qi *et al.*, 2016] Charles R Qi, Hao Su, Matthias Niessner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. *arXiv preprint arXiv:1604.03265*, 2016.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[Shilane *et al.*, 2004] Philip Shilane, Patrick Min, Michael Kazhdan, and Thomas Funkhouser. The princeton shape benchmark. In *Shape modeling applications, 2004. Proceedings*, pages 167–178. IEEE, 2004.

[Su *et al.*, 2015] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 945–953, 2015.

[Sun *et al.*, 2009] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, volume 28, pages 1383–1392. Wiley Online Library, 2009.

[Tatsuma *et al.*, 2012] Atsushi Tatsuma, Hitoshi Koyanagi, and Masaki Aono. A large-scale shape benchmark for 3d object retrieval: Toyohashi shape benchmark. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, pages 1–10. IEEE, 2012.

[Wu *et al.*, 2015] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[Xie *et al.*, 2016] 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*, PP(99):1–1, 2016.