Guoxian Dai

Personal Information

Status:	**PH.D. STUDENT**
Program:	**Computer Science and Engineering**
School:	**Tandon School of Engineering, New York University**
Period:	**From 2014-09 to 2018-12**

Biography

I was a Ph.D. student at New York University and advised by Professor YiFang. During my Ph.D. period, I was a research assistant in NYU Multimedia and Visual Computing (MMVC) Lab. I am broadly interested in 3DComputer Vision and Deep Learning. Now I am a researchscientistatX-motors.ai (CA,USA).

Research Project: Learning with hand-crafted features

Description

Existing hand-crafted features could characterize 3D shape up to certain variations, such as isometric deformation, scale-variation. However, they could not handle large deformations. In this chapter, we mainly study the problem of learning more roust and deformation-invariant shape descriptor on top of the hand-crafted features. we developed a discriminative deformation-invariant 3D shape descriptor via many-to-one encoder on top of scale-invariant heat kernel signature (SIHKS).

Method

Our method mainly includes three steps: 1) local feature extraction. We extract scale-invariant heat kernel signature (SIHKS) to describe each vertex of the shape. SIHKS has attractive geometric properties, include invariance to both isometric transformation and scale change. 2) LLC based shape descriptor (LSD). Since local features focus on capturing the local geometric structures of 3D model, which could not directly describe the whole 3D shape. Thus we convert the local features into a global shape descriptor using locality-constrained linear coding (LLC). Compared to bag-of-words (BoW) coding scheme, LLC has lower reconstruction error. 3) LSD was further improved via many-to-one encoder (MOencoder) to get a more discriminative deformation-invariant 3D shape descriptor.

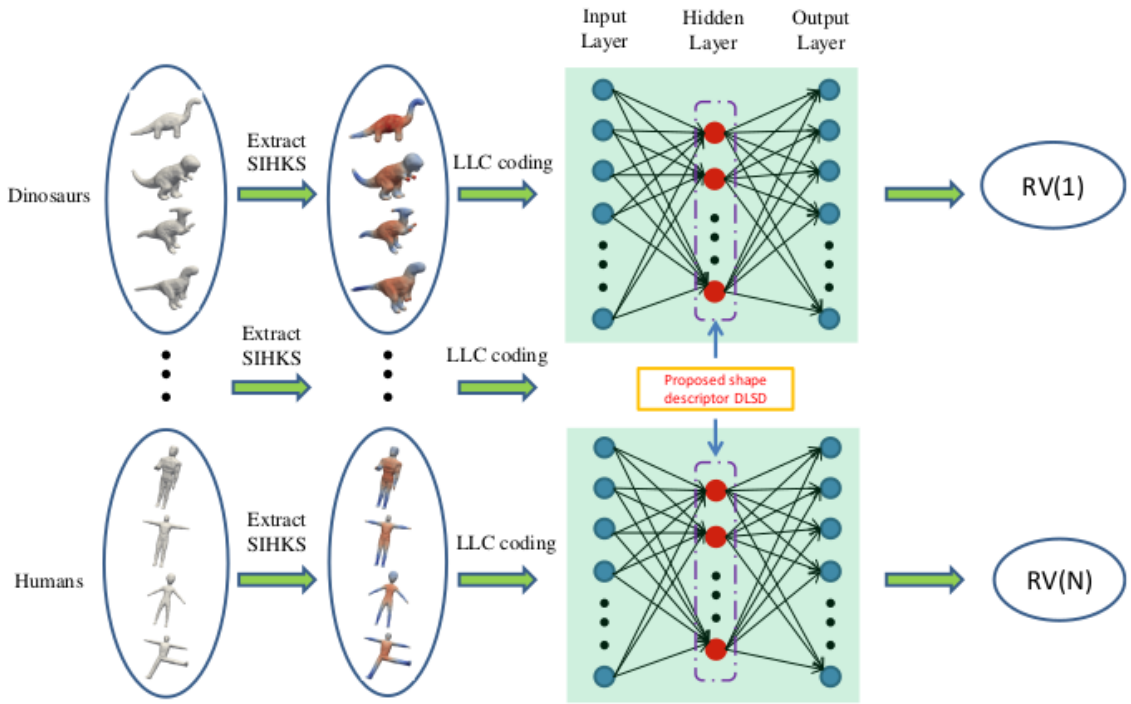

Figure 1: Detailed framework of our proposed method. The proposed method mainly includes three components, namely, local feature extraction, specifically, SIHKS; constructing global shape descriptor using LLC; learning discriminative deformation-invariant shape descriptor via MOencoder.

Results

To evaluate the performance of our proposed shape descriptor, we conduct shape retrieval experiments on three well-known benchmarks, including McGill, SHREC'10, and SHREC'14 human. In most cases, our proposed shape descriptor outperforms state-of-the-art methods.

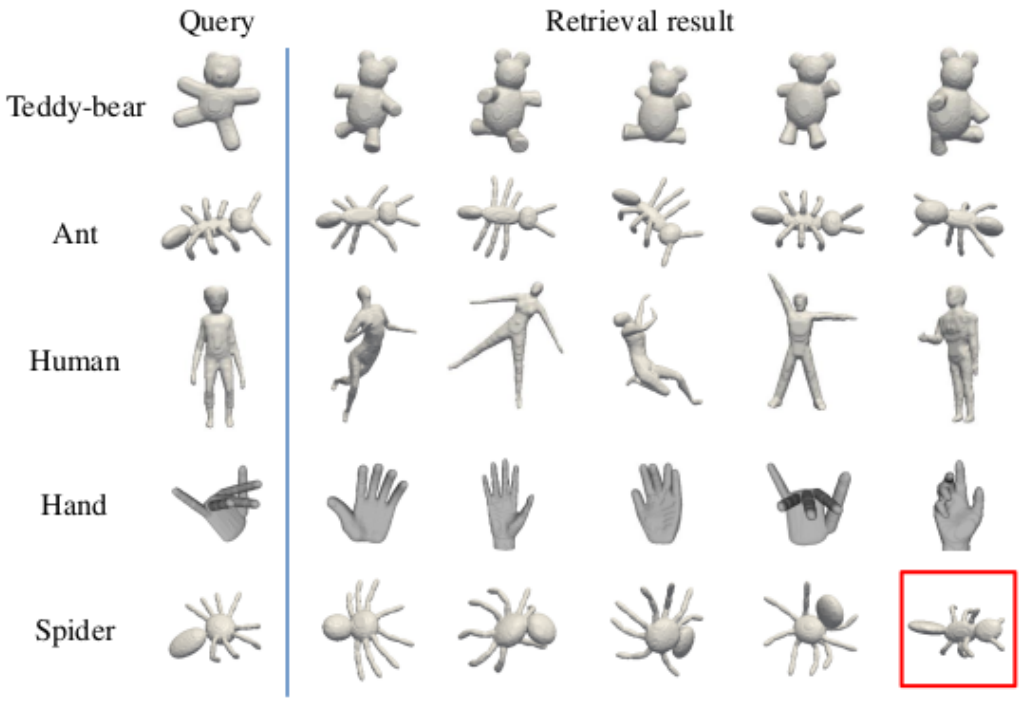

Figure 2: Retrieval examples on McGill dataset. The queries are listed on the leftmost column, while the retrieved objects are listed on the right side, according to their retrieved ranking order. The wrong results are marked with red box.

Method	mAP
VQ [19]	0.891
Unsupervised [84]	0.891
Supervised [84]	0.921
LSD	0.84
DLSD (proposed)	**0.96**

Figure 3: mAP comparison on SHREC 2010 dataset.

Method	Synthetic	Real
ISPM [78, 79]	0.902	0.258
DBN [96]	0.842	0.304
HAPT [44]	0.817	0.637
VQ [19]	0.813	0.514
Unsupervised [84]	0.842	0.523
Supervised [84]	0.954	**0.791**
LSD	0.800	0.355
DLSD (proposed)	**0.99**	0.754

Figure 4: mAP comparison of different methods on SHREC'14 human dataset.

Research Project: Learning with projected images

Description

Instead of learning on top of hand-crafted features, 3D shapes could also be projected into 2D images from different views, and standard CNN models could be applied for learning robust shape descriptor. In this chapter, we mainly study the problem of learning discriminative shape descriptor with its projected 2D images. We propose a siamese CNN-BiLSTM network for 3D shape representation learning. Our proposed method is evaluated on two large-scale benchmarks, Princeton ModelNet and SHREC 2014. The experimental results demonstrate the superiority of our proposed method over the state-of-the-art methods.

Method

First, each 3D shape is rendered on multiple different views based on pre-defined 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 information 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 shapes from the same class is minimized, and the distance for shapes from different classes is maximized to a large margin.

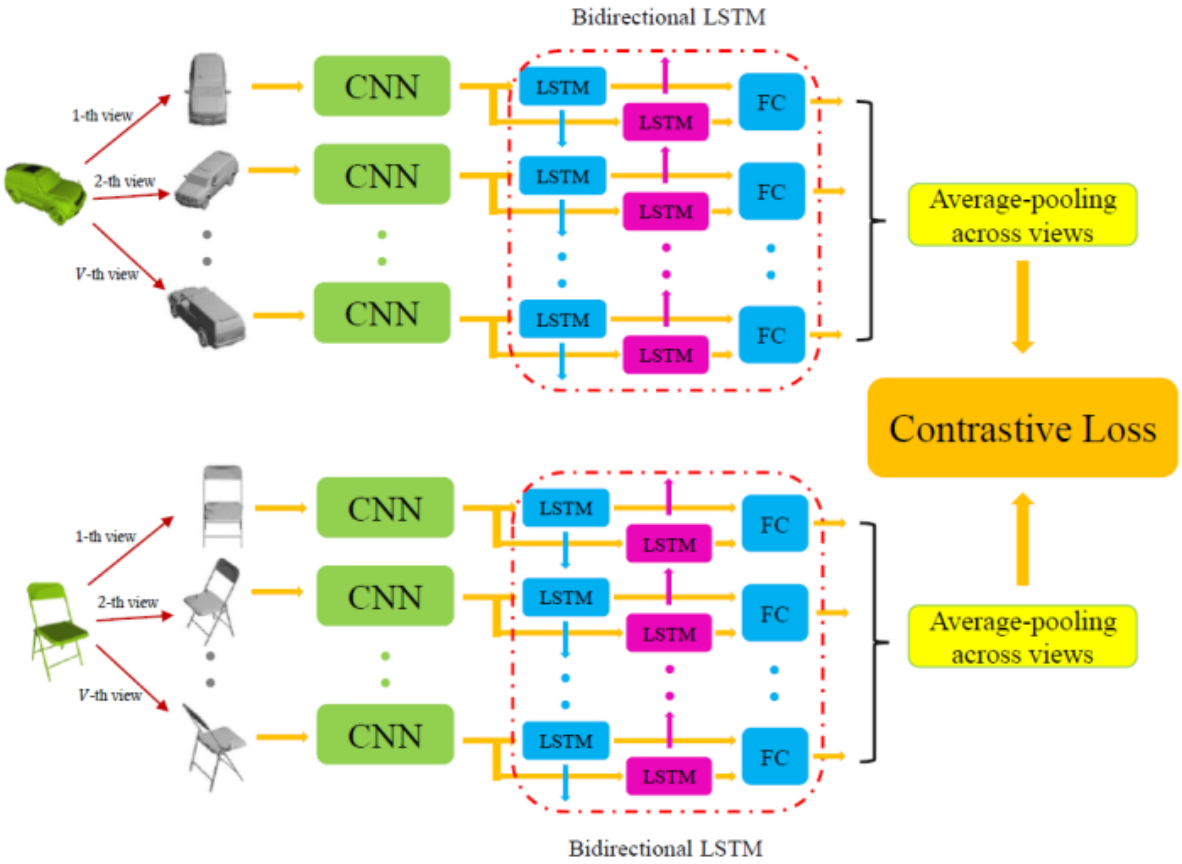

Figure 5: Detailed framework of our proposed method. Each input shape is rendered on V different views. All the V projected images are treated as a sequence. They are first passed through CNN to extract visual features from each view. Then bidirectional LSTM is adopted to aggregate information across 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.

Results

Our proposed method is evaluated on two large-scale benchmarks, Princeton ModelNet and SHREC 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 experimental results demonstrate the superiority of our proposed method over the state-of-the-art methods.

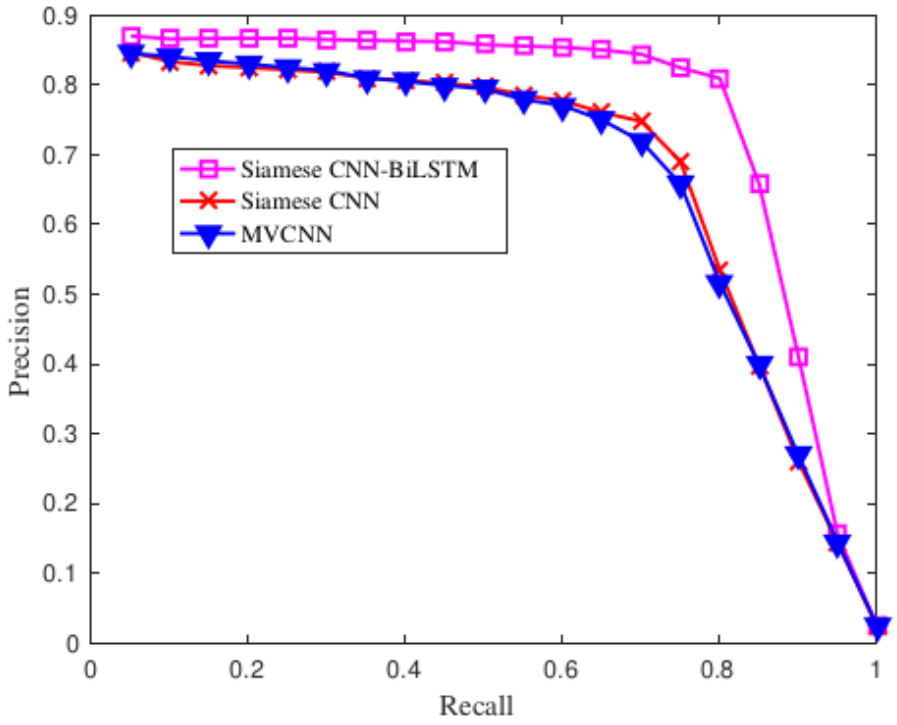

Figure 6: Precision-recall curves of siamese CNN-BiLSTM, siamese CNN and MVCNN on ModelNet40.

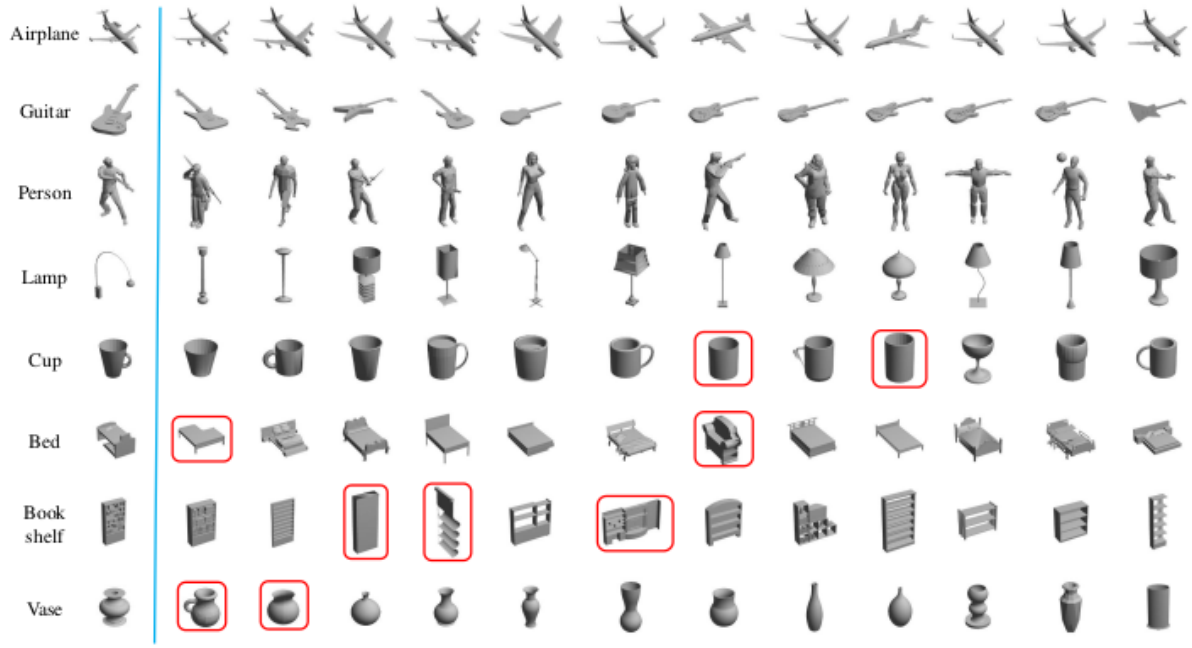

Figure 7: Illustration of the retrieved examples on ModelNet40. The query models are listed at the leftmost column, and the retrieved shapes are listed on the right side by their ranking order.

Methods	mAP
LFD [23]	0.409
SHD [65]	0.333
ShapeNet [145]	0.492
MVCNN [127]	0.802
GIFT [6]	0.819
Siamese CNN-BiLSTM	**0.833**

Figure 8: Comparison with the state-of-the-art methods on ModelNet40.

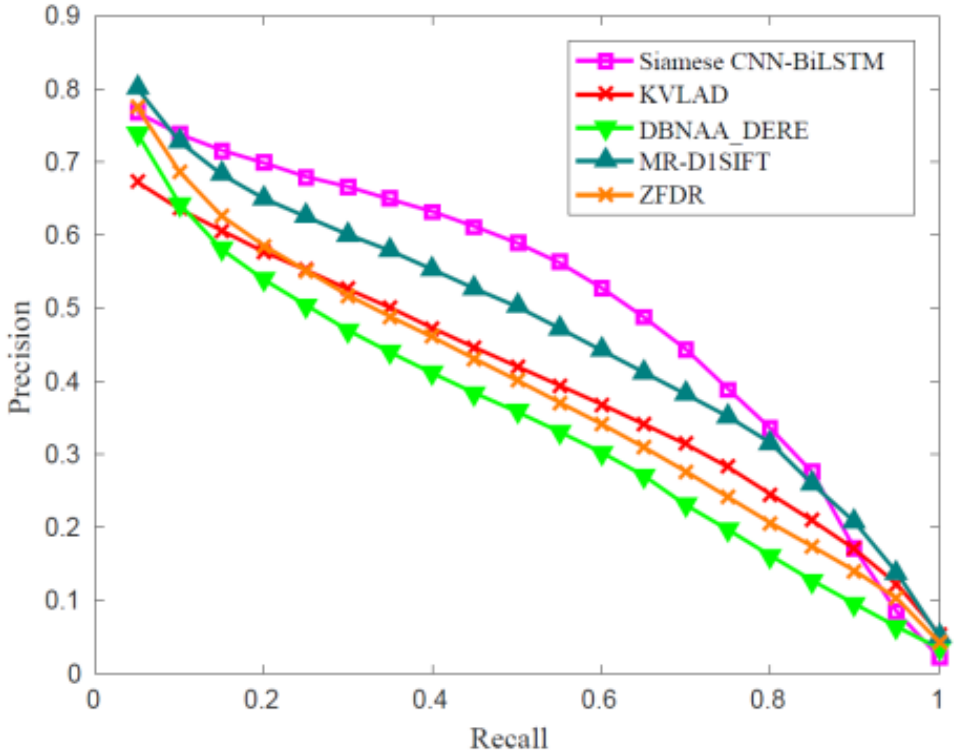

Figure 9: Performance comparison of precision-recall curves on SHREC 2014.

methods	NN	FT	ST	E	DCG	mAP
KV [77]	0.605	0.413	0.546	0.214	0.746	0.396
DB [77]	0.817	0.355	0.464	0.188	0.731	0.344
MR [77]	0.856	0.465	0.578	0.234	0.792	0.464
ZF [77]	0.838	0.386	0.501	0.209	0.757	0.387
LC [77]	0.864	0.528	0.661	0.255	0.823	0.541
DA [146]	**0.897**	0.401	0.503	0.243	0.790	-
Proposed	0.812	**0.617**	**0.730**	**0.358**	**0.831**	**0.644**

Figure 10: Comparison with the state-of-the-art methods on SHREC 2014.

Research Project: Adversarial embedding

Description

In order to learn more generic and discriminative shape descriptor, we study the problem of shape embedding with semantic information. Specifically, we propose an adversarial embedding process, which maps 3D shape from the geometric space into the vector representations of their class label with an adversarial loss. The vector representations of class labels are generated by word2vec, which has nice properties, the vector representations of semantic-similar words are closer to each other than those dissimilar words. Our proposed method is evaluated on two large-scale benchmarks, ModelNet40 and SHREC 2014. And the experimental results demonstrate the superiority of the proposed method over state-of-the-art methods.

Method

Specifically, the 3D models are embedded to the vector representations of the semantic class labels with an adversarial loss. A generator is trained to minimize the distance between the generated shape descriptor and the vector representation of class label, meanwhile fool the discriminator. On the other hand, a discriminator is trained to distinguish between the generated shape descriptor and the ground truth vector representation. The vector presentations of the class labels are generated from word2vec. Consequently, shapes from the same class are mapped to the same vector representation, so that their distances are minimized in the embedding space. The distances of shapes from different classes are inherited from the distances of their corresponding vector representations, which makes the semantic-similar classes stay closer than the semantic-dissimilar classes in the embedding space.

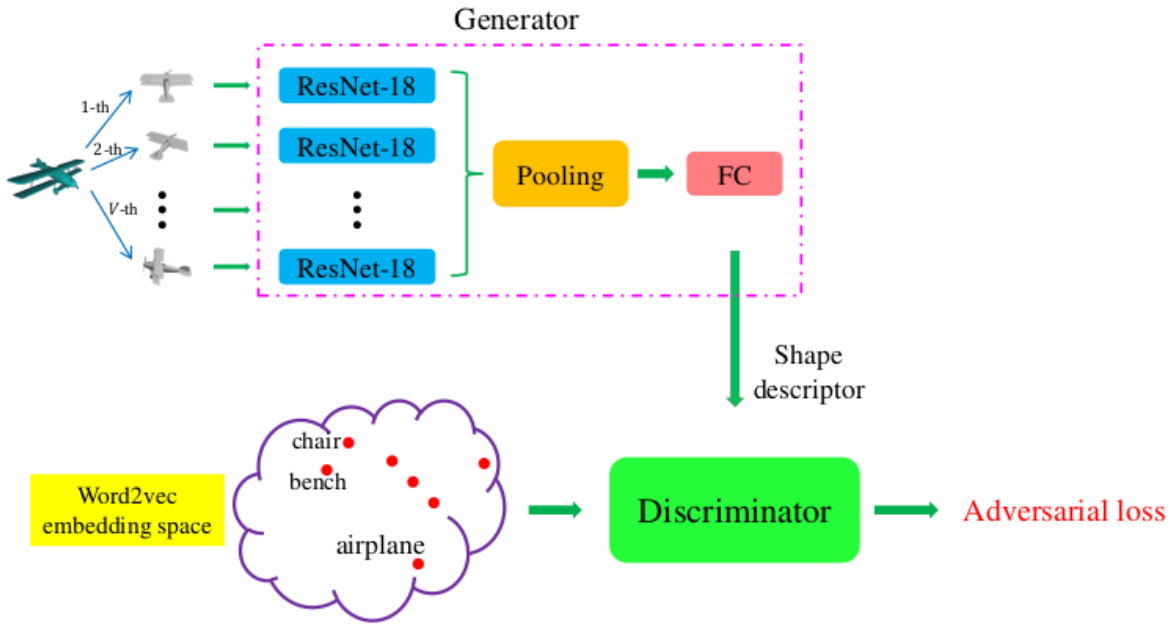

Figure 11: The detailed framework of the proposed method. Each shape is ren- dered on V different views, and then ResNet-18 is used to extract visual features from the rendered images. The multiview features are passed through a pooling layer across different views to form a global representation. "FC" denotes fully con- nected layer. Additional fully connected layers are followed to form the final shape descriptor. The proposed adversarial embedding is trained with an adversarial loss.

Results

Our proposed method is verified on two large-scale benchmarks, ModelNet40 and SHREC 2014. And the evaluation is based on standard criterion. Precision-recall curve is used to visualize the retrieval performance of the proposed method, meanwhile nearest neighbor (NN), first tier (FT), second tier (ST), discounted cumulated gain (DCG) and mean average precision (mAP) are used to quantitatively measure the retrieval performance of the proposed method. In addition, we also compare the proposed method with state-of-the-art methods. The experimental results demonstrate the superiority of the proposed method over other methods on both benchmarks.

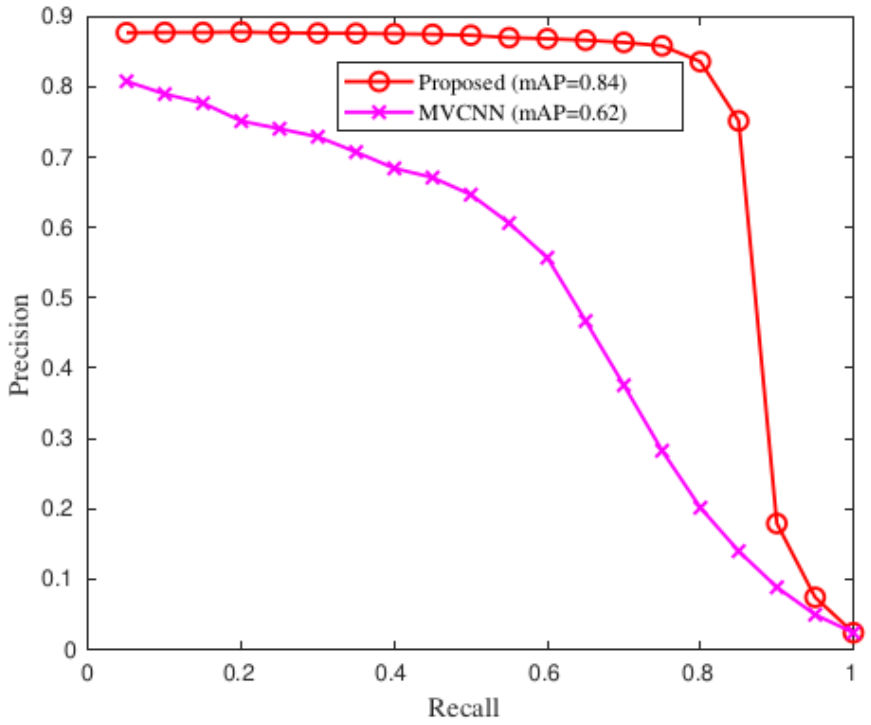

Figure 12: Precision-recall curves of the proposed method and MVCNN on Mod- elNet40.

Method	NN	FT	ST	DCG	mAP
Direct embedding	0.860	0.798	0.861	0.902	0.814
Adversarial embedding	**0.875**	**0.828**	**0.880**	**0.915**	**0.840**

Figure 13: Performance comparison between the proposed method and direct em- bedding with Euclidean loss.

Method	mAP
LFD [23]	0.409
SHD [65]	0.333
ShapeNet [144]	0.492
MVCNN [127]	0.802
GIFT [7]	0.819
Direct embedding	0.814
Proposed	**0.840**

Figure 14: Examples of fragment alignment results using our learned 3D local descriptor. For comparison, we also provide the alignment results using the descriptors generated from the state-of-the-art 3DMatch.

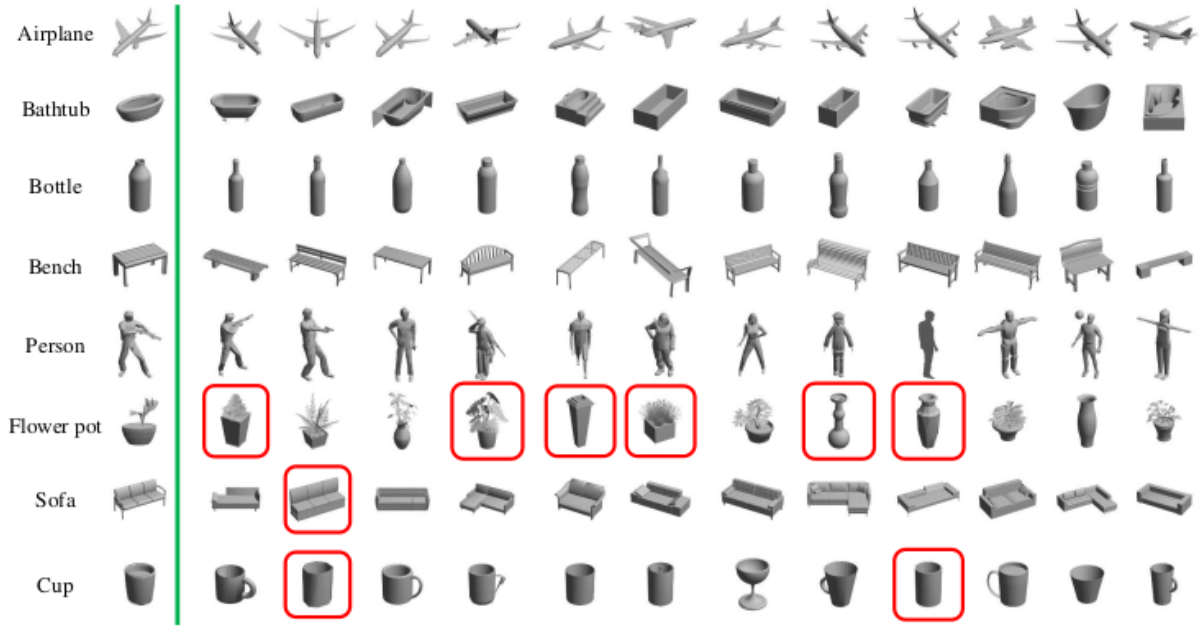

Figure 15: Performance comparison with state-of-the-art methods on ModelNet40.

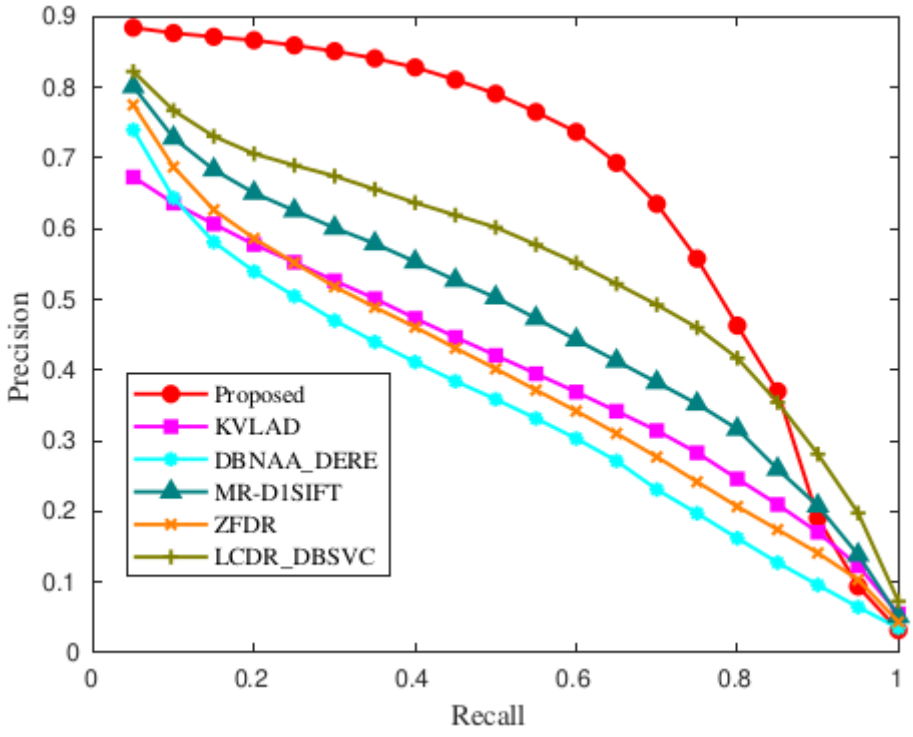

Figure 16: Retrieval examples of the proposed method on ModelNet40. All the query objects are listed at the leftmost column. And the top 12 retrieved models are listed on the right side based on the ranking order. In addition, the mistakenly returned objects are marked with red boxes.

Research Project: Jointly learning 3D shape with sketch

Description

In this project, we mainly focus on the problem of cross-domain learning, which jointly learns 3D shape with sketch into a unified representation for cross-domain recognition or retrieval. The main challenges for the cross-domain learning task are that 3D shapes and sketches are from two different modalities, and the huge modality gap makes it very difficult to learn a unified representation for both modalities. We proposed a novel deep correlated holistic metric learning method to mitigate the discrepancies between 3D shape with sketch. Our proposed method is evaluated on SHREC 2013, 2014 and 2016, and the experimental results demonstrate superiority over the state-of-the-art methods.

Method

We first extract low-level features for both sketches and 3D shapes. For 2D sketch, we use pre-trained AlexNet to extract features; for 3D sketch, we extracted histogram of oriented distance (HOD); for 3D shape, we extract 3D-SIFT feature, which is extended from the well-known 2D SIFT. The extracted 3D-SIFT is further encoded by locality-constrained linear coding (LLC) to get a global shape descriptor. Then we learn two deep neural networks to transform the raw features of both domains into a new feature space, mitigating the domain discrepancy as well as maintaining the discrimination. The loss of the proposed network includes two parts, discriminative term which is constructed with pairwise distance within each domain and correlation term which is constructed with pairwise distance across different domains. The former one minimizes the variations of the deep learned features from the same class and maximizes the variations of the deep learned features from different classes within each domain; the latter one aims to alleviate the domain discrepancy, making the distributions of features from both domains as consistent as possible. Apart from adding the proposed loss at the output layer, similar loss is also imposed at the hidden layer to guide features in hidden layer also with desired properties. And it could further increase the robustness of deep learned features at the output layer.

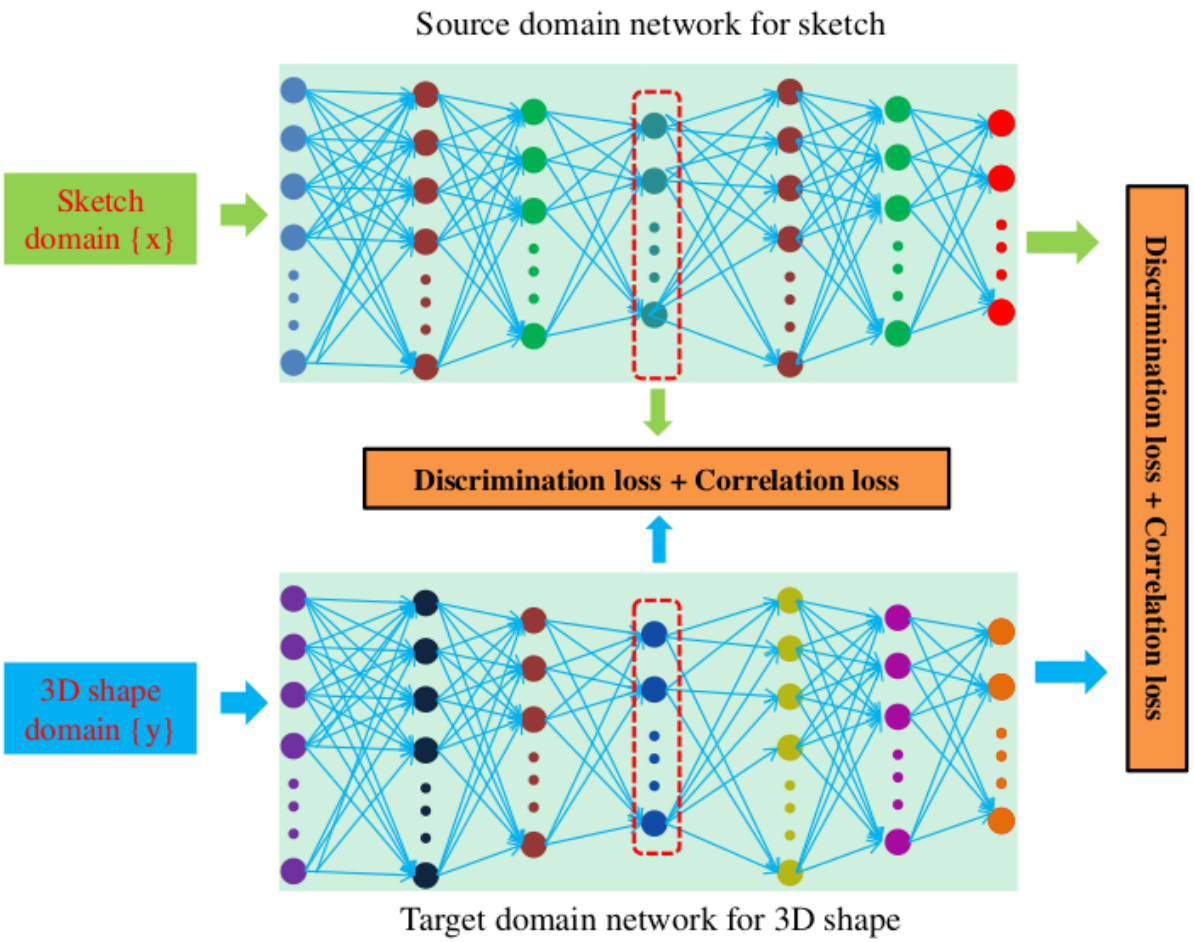

Figure 17: The detailed framework of our proposed deep correlated holistic metric learning network. The whole network structure mainly includes two components, source domain network and target domain network. The proposed loss function is imposed at both output layer and hidden layer.

Results

Our proposed method is evaluated on three well-known benchmarks, SHREC 2013, SHREC 2014 and SHREC 2016. Besides, we also compared our proposed method with the state-of-the-art methods using several common metrics. Precision-recall curve is provided to visualize the performance of our proposed method. Overall, the experimental results demonstrate that our proposed method could outperform the state-of-the-art methods.

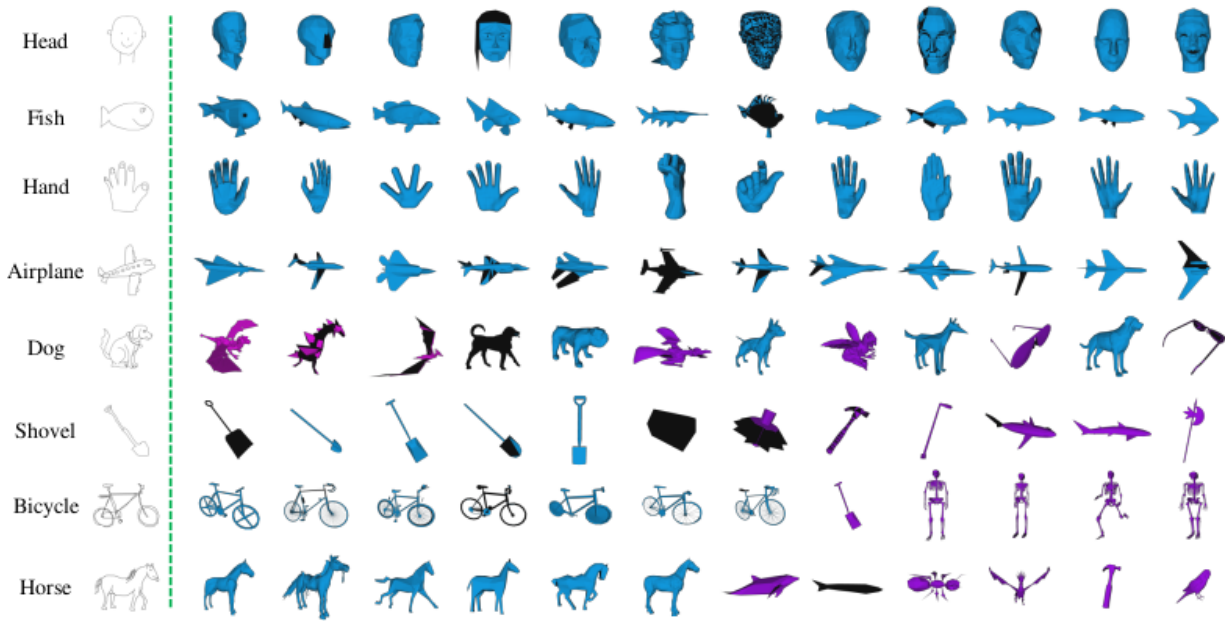

Figure 18: Illustration of retrieved examples on SHREC 2013 dataset. The query sketch is listed on the left first column, and the top 12 retrieved 3D models are listed on the right side, according to their ranking orders. The correct retrieved examples are marked with blue color, while the incorrect retrieved examples are marked with purple color.

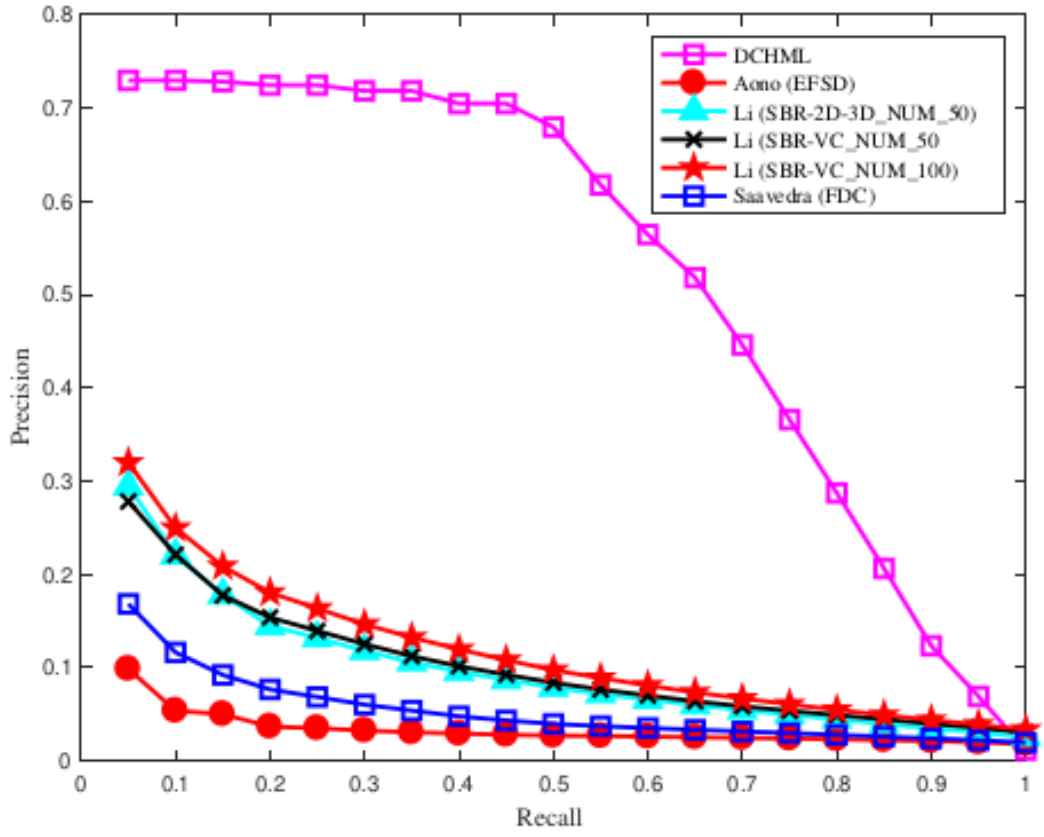

Figure 19: Performance comparison of precision-recall curve on SHREC 2013 dataset.

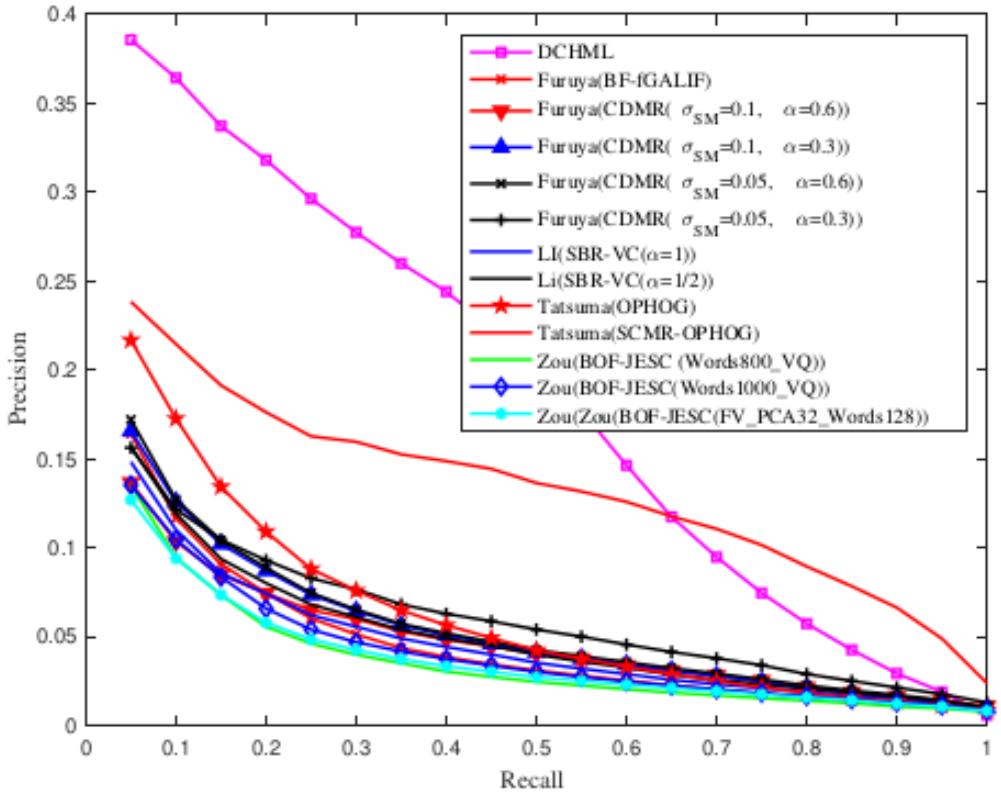

Figure 20: Performance comparison of precision-recall curve on SHREC 2014 dataset.

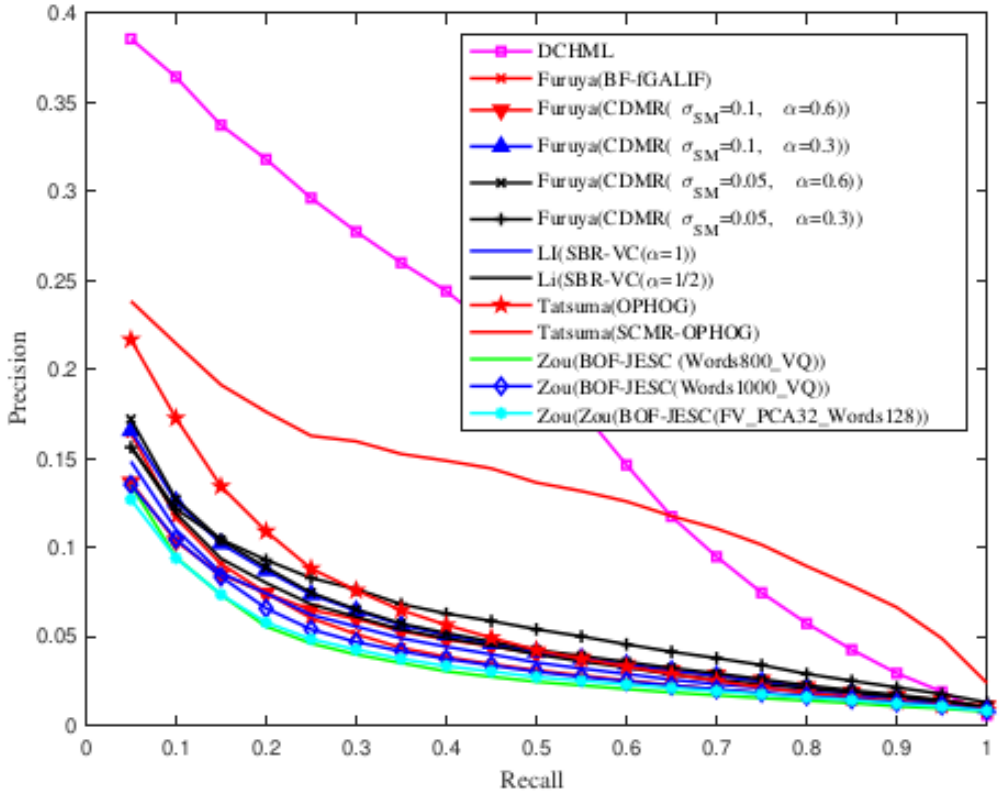

Figure 21: Examples of 3D sketches and shapes from SHREC 2016.