Sketch-Based Shape Retrieval via Multi-view Attention and Generalized Similarity

Yongzhe Xu *, Jiangchuang Hu†, Kun Zeng‡
School of Data and Computer Science, Sun Yat-Sen University
*xuyzh6@mail2.sysu.edu.cn, †hujch3@mail2.sysu.edu.cn, ‡zengkun2@mail.sysu.edu.cn

Abstract—Sketch based shape retrieval has received increasing attention in computer vision and computer graphics. It suffers from the challenge gap between sketch and 3D shape. In this paper, we propose a generalized similarity matching framework based on a multi-view attention network(MVAN), where users can retrieval 3D shapes that share the same semantics of input query sketch. We first project 3D shape from multiple view points and use a convolutional neural network to extract low level feature maps of these 2D projections. Then a multi-view attention network is adopted to fuse the feature maps and forms a more accurate 3D shape representation. We use another CNN to extract the feature of sketch. Finally we measure the similarity of sketch and 3D shape via a generalized similarity model, which fuses some traditional similarity model into a general form and optimizes its parameters using data-driven method. We combine the MVAN and generalized similarity model into a unified network (Fig. 1) and train the whole model in an end-to-end manner. We evaluate our method on SHREC'13 and SHREC'14 sketch track benchmark datasets. The experimental results demonstrate that the proposed method can outperform state-of-the-art methods.

Index Terms—Sketch, Shape, Attention, Generalized Similarity

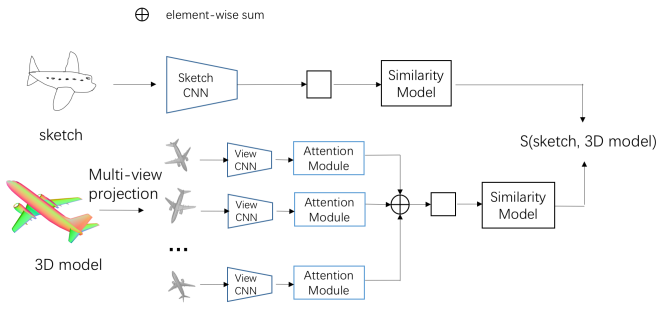

Fig. 1. Our MVAN and generalized similarity model for sketch-based shape retrieval. We use a convolutional neural network to extract the features of projections of 3D shape and then fuse the features via attention model to get prominent representation for 3D shape. We then use another CNN to extract the feature of query sketch. Finally a generalized similarity model is applied to calculate the distance between sketch and shape. The smaller the distance, the more similar the sketch and shape are.

I. Introduction

With the development of virtual reality, a large number of 3D shapes have been created, which causes a growing demand for retrieval of 3D shapes. The research for retrieving 3D shapes mainly focus on three ways: retrieval by keyword, retrieval by 3D shape and retrieval by sketch. For keyword-based shape retrieval, users are supposed to provide very accurate keywords to describe the feature of 3D shape and meanwhile the labels of 3D shape should be rich and accurate enough. It not only increases the burden for labeling but is inefficient for users to query a 3D shape. The 3D-shape-based method seems more direct as a user can input a similar 3D shape and retrieval what he needs. However, obtaining a 3D shape itself is not an easy thing, so it's not friendly in this situation. Compared to previous two methods, sketch-based shape retrieval is more convenient since users don't need to be skillful in painting and 3D modeling and can draw arbitrary lines to express their intents. Therefore sketch-based shape retrieval has received considerable attention in recent years.

Sketch-based shape retrieval needs to overcome the cross-domain similarity matching between sketches and 3D shapes. To resolve such problem, many existing methods project 3D shapes into 2D representations and then measure the similarity between query sketch and projections. It converts the difficult matching problem between image and 3D shape into a more simple one between image and image, where the latter has been heavily studied in computer vision. However, due to the large variation of sketch, there still exists some visual difference between sketch and 2D projection. To overcome such problem, recently many researchers tried to use different algorithms to extract features from different domains and applied some similarity model to project the feature vectors into the same feature space and measure the distance of feature vectors [1]–[3]. Besides the challenge of domain gap, another problem comes from the unknown pose of 3D shapes, which makes it difficult to find the best view point for projection. We point out that different projections characterize differently to 3D shape. It is straightforward to recognize 3D shape from some of its projections. But there also exists some unreasonable view points from where even human are unlikely to distinguish the original 3D shape (See Fig. 2).

According to this observation, in this paper we propose a multi-view attention network(MVAN) to fuse different projections into a more compact representation, which has stronger ability to characterize 3D shape. We uniformly sample several view points on a sphere with the center

of the shape as its center. We render the shape using Phong reflection model [4] by placing virtual cameras on each view point which point towards the centroid of the shape.Then the MVAN and another CNN are used to extract features of 3D shape and sketch separately. Finally a similarity model is applied to calculate the distance between sketch and 3D shape and we can retrieval the most similar shape for a query sketch based on knn algorithm. In terms of similarity model, many existing works prefer to Euclidean distance [3], [5], Mahalanobis distance [6] or Cosine similarity [7]. However, we mention that similarity model is highly related to the property of feature vectors. It's hard to define the most suitable similarity model to match what the CNN outputs. Therefore, we explore a generalized similarity model proposed by Lin et al. [8] to calculate the similarity between different domains. The generalized similarity model is proved to be a general form of Mahalanobis distance and Cosine similarity. Its parameters are optimized via data-driven method and thus can be embedded into the feature learning framework and has better performance to calculate domain similarity.

In summary, our contribution includes:

- We fuse the features from different projections into a prominent representation of 3D shape based on attention mechanism.
- We further explore a generalized similarity model to learn the similarity parameters between different domains.
- We perform our experiments on SHREC'13 and SHREC'14 benchmark. The result illustrates that our method can achieve state-of-the-art.

The rest of the paper is organized as follows. Section 2 reviews related work. In section 3 we introduce our multi-view attention network and generalized similarity model. Section 4 presents the experimental results. We conclude the paper in section 5.

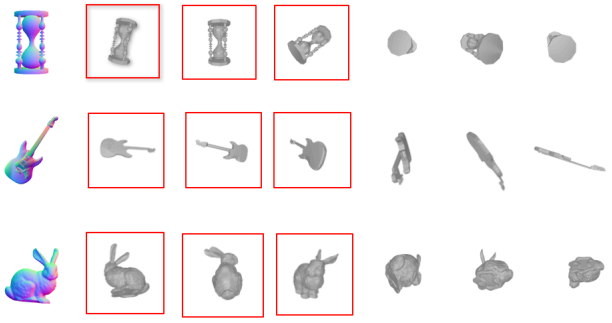

Fig. 2. The projections of different 3D shapes. From top to bottom are hourglass, guitar, rabbit and their projections separately. The projections labeled with red rectangle retain the most significant feature of shapes, while the left are not easy to recognize the original shapes even for human eyes.

II. Related Work

Currently, there are mainly three methods for shape retrieval: query-by-keyword, query-by-shape and query-by-sketch. In this paper, we focus on sketch-based shape retrieval.

Sketch-based shape retrieval has received considerable attention in recent years since sketch is human nature and it is very convenient for users to describe their intents. At present, most of the methods project 3D shapes into 2D representations and measure the similarity between sketch and projections. Eitz et al. [7] render 3D shapes by occluding contours and suggestive contours from numerous view points. They extract the Gabor local line-based features (GALIFs) from projections and construct a dictionary by performing bag-of-feature method. Then the histograms of sketch and projections are extracted from the dictionary and the Cosine similarity is used for retrieval. Saavedra et al. [9] extract the external contours of projections and sketches and then apply three kinds of descriptors, HELO, HOG and fourier descriptors to describe the contours. Finally the Manhattan distance is used for similarity matching. In [10], a ZFEC descriptor is used to encode the sihouette and external contours of projections and sketches, which aims to find the similar candidate projections quickly. Then the relative shape context matching method is adopted to more accurately measure the similarity between sketch and candidate projections.

Since the success of deep learning in computer vision, researchers have started to apply neural network, especially deep convolutional neural network to extract the deep features of sketches and projections. Wang et al. [1] randomly sample two view points to render 3D shapes and then two CNNs of different parameters are applied to extract the features of sketches and projections respectively. A objective function based on within-domain similarity and cross-domain similarity is defined to optimize the parameters. As the first attempt of deep learning on this subject, their performance outperforms other state-of-the-art methods based on hand crafted features. Xie et al. [3] propose to learn barycentric representation of 3D shape based on Wasserstein distance. They apply another CNN to extract the feature of sketch and train the model according to contrastive loss. Li et al. [2] adopt two VGGs [11] of different parameters to extract features of sketches and projections respectively. They then propose to learn the multi-view pairwise relationship between sketch and projections, which can facilitate to handle the semantic similarity between sketch and shape and thus improve the performance.

In this paper, we propose to fuse the projections of 3D shape based on attention mechanism to learn a compact shape representation. Attention mechanism is motivated by human visual system. It means to pay attention to key parts of image and ignore the less important areas. In

recent years attention model has been applied successfully in image description [12], [13], fine grained image classification [14], [15] and natural language processing [16], [17]. It can be divided into soft attention and hard attention. In this paper we adopt the soft attention model. We extract the feature maps of projections from CNN and then learn the attention masks from each projection. The attention masks record the weight of each projection, which can be used to do weighted summation of feature maps and form a prominent feature map for 3D shape. Note that different from the barycentric method proposed in [3], we fuse the low level feature maps which can preserve more about the structural information of projections, while in [3] they learn barycentric representation from fully connected layer of CNN thus their method focus more on the probability distribution of semantic features. In addition, most of the aforementioned approaches adopt Euclidean distance or Cosine similarity as their similarity models. We point out that it is hard to specify the most suitable similarity model for deep features of different domain. Thus, in this paper we adopt a generalized similarity model proposed by Lin et al. [8], which can learn its parameters through a data-driven manner.

III. Method

In this section, we introduce our sketch-based shape retrieval framework in detail. The multi-view attention network will be covered in section 3.1. Then in section 3.2 we demonstrate how to apply the generalized similarity model to measure the similarity of sketch and 3D shape. Finally in section 3.3 we introduce the embedding of feature learning and similarity model, enabling an end-to-end training.

A. Multi-View Attention Network

In sketch-based shape retrieval, most existing works tend to render 3D shape from multiple view points in order to preserve the characteristic of shape as much as possible. However, as we have mentioned before, due to the unknown pose of shape, actually not all the projections can represent the original shape, some of which even act as noise. In Fig. 3 we display four projections of two shapes from different view points. One can easily identify the original shapes from their side views. However, it will be more difficult from top views. In addition, as their projections from top views are very similar to each others, it further affects the ability of neural network to distinguish 3D shapes. Besides, as we have claimed in section 1, some of the projections can characterize 3D shape well, while the others have difficulties highlighting the feature of 3D shape. Therefore, different projections are supposed to own varied weights in order to represent shape more accurately.

According to above observations, we propose a multi-view fusion network based on attention mechanism. Firstly, We evenly sample k view points on a sphere centered at the center of shape and then render projections by Phong reflection model. These k projections are further input into the network. The base architecture of the network is adopted as AlexNet [18], however, we add an attention module after the first pooling layer. In our experiment, the attention module is another network consisting of one convolutional layer with kernel size 1 and the number of filters is the same as the channel number of input feature map. The feature map extracted from the first pooling layer is fed into this attention module and an attention mask with the same size is then obtained. As we input k projections into the network, k feature maps and their attention masks are obtained from attention module. Each attention mask will weight their corresponding feature map by element-wise production. We further merge these weighted feature map by element-wise summation and fed it into the subsequent layers. The multi-view attention network is shown in Fig. 4. To be more specifically, we denote the feature map as $f \in R^{H\times W\times C}$ where H, W, C represent the height, width and channel of the feature map respectively. Then the attention mask is calculated as follow:

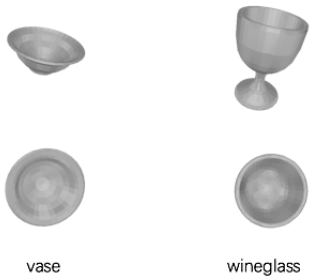

Fig. 3. Projections of 3D shapes from different view points. From left to right are projections of vase and wineglass. From top to bottom are side views and top views separately.

$$\alpha_{i,j} = F^{att}(f_{i,j}, W_{att})$$

where F^{att} is the attention layer with parameter W_{att} and $f_{i,j}$ is the feature vector of feature map at location (i, j). We further normalize the attention mask with softmax function:

$$\alpha^v_{i,j} = \frac{e^{\alpha_{i,j}{}^v}}{\sum_{l=1}^{k} e^{\alpha_{i,j}{}^l}}$$

The filtered feature map is then obtained by element-wise production of the attention mask:

$$f^{att}_{i,j} = f_{i,j} \otimes \alpha_{i,j}$$

After these k filtered feature maps are generated, we sum them to form a more compact representation and continue to feed it into the subsequent layers.

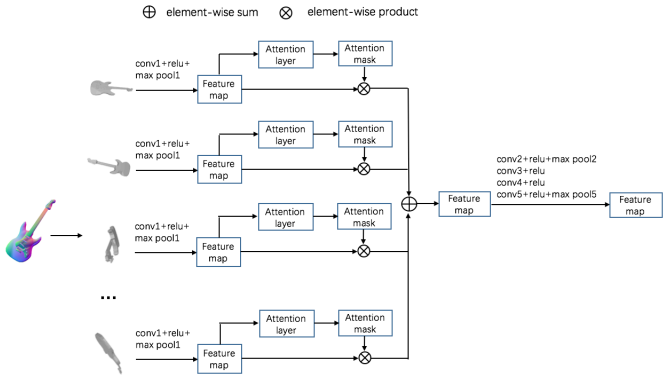

Fig. 4. Our multi-view fusion network.

B. Cross-Domain Similarity Model

In most existing works, researchers focus more on feature learning and care less about the effect of similarity model. However, as similarity model is feature related, it is nearly impossible to specify the best similarity measure by hand. Therefore, we explore the cross domain similarity model [8], of which the parameters are learned in data-driven way and apply it into the similarity measure between sketch and shape.

In [8], most similarity models are combined into a general form by affine transformation:

$$S(\mathbf{x},\mathbf{y}) = \begin{bmatrix}\mathbf{x}^T & \mathbf{y}^T & 1\end{bmatrix}\begin{bmatrix}\mathbf{A} & \mathbf{C} & \mathbf{d}\\ \mathbf{C}^T & \mathbf{B} & \mathbf{e}\\ \mathbf{d}^T & \mathbf{e}^T & f\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\\ 1\end{bmatrix} \quad (1)$$

where $\mathbf{A}$,$\mathbf{B}$ are restricted as simi-difinite matrix and $\mathbf{C}$ is the correlation matrix between domain $\mathbf{x}$ and $\mathbf{y}$. Thus they can be represented as:

$$\begin{aligned}\mathbf{A} &= {\mathbf{L_A}}^T\mathbf{L_A}\\ \mathbf{B} &= {\mathbf{L_B}}^T\mathbf{L_B}\\ \mathbf{C} &= -{\mathbf{L_C^x}}^T\mathbf{L_C^y}\end{aligned} \quad (2)$$

Substitute (2) into (1) we obtain the general similarity model:

$$\begin{aligned}\tilde{S}(\mathbf{x},\mathbf{y}) =& S(\mathbf{f_1}(\mathbf{x}),\mathbf{f_2}(\mathbf{y}))\\ =&||\mathbf{L_A}\mathbf{f_1}(\mathbf{x})||^2 + ||\mathbf{L_B}\mathbf{f_2}(\mathbf{y})||^2 + 2\mathbf{d}^T\mathbf{f_1}(\mathbf{x})\\ &- 2(\mathbf{L_C^x}\mathbf{f_1}(\mathbf{x}))^T(\mathbf{L_C^y}\mathbf{f_2}(\mathbf{y})) + 2\mathbf{e}^T\mathbf{f_2}(\mathbf{y}) + f\end{aligned} \quad (3)$$

Note that this model is antisymmetric, meaning $\tilde{S}(\mathbf{x},\mathbf{y}) \neq \tilde{S}(\mathbf{y},\mathbf{x})$, which is reasonable as $\mathbf{x}$ and $\mathbf{y}$ belong to different domains and own different parameters. Since f has no influence on training, we set it to be -1.9 throughout this paper.

We use triplet ranking loss to train our similarity model. Given a triplet sample in the form of (x, y^+, y^-), where x and y are sketch and shape respectively and x is more similar to y^+ than to y^-, triplet ranking loss is defined as follow:

$$\tilde{l}_{triplet}(\mathbf{x},\mathbf{y}^+,\mathbf{y}^-) = max(f(x,y^+) - f(x,y^-), C)$$

where C represents the margin between similarity distance. We replace f with equation(3) and obtain the following objective function:

$$\tilde{l}_{triplet}(\mathbf{x},\mathbf{y}^+,\mathbf{y}^-) = max(\tilde{S}(\mathbf{x},\mathbf{y}^+) - \tilde{S}(\mathbf{x},\mathbf{y}^-), C)$$

In the experiment, we set C to be -50.

C. End-to-End Training

We now describe how to train feature learning and similarity model in an end-to-end way. The feature extractor(MVAN) for shape has already been demonstrated in section 3.1. We use AlexNet to extract the feature of sketch. In order to combine with similarity learning, we adopt a similar architecture as in [8], where we prune the FC layer of AlexNet and MVAN and then connect them using a shared sub-network. The output feature of the sub-network will be fed into the similarity network, which will output similarity parameters for sketch and shape respectively. Note that due to large bias of different domain, the parameters of batch normalization in shared sub-network and of similarity network are different. The whole network architecture is shown in Fig. 5.

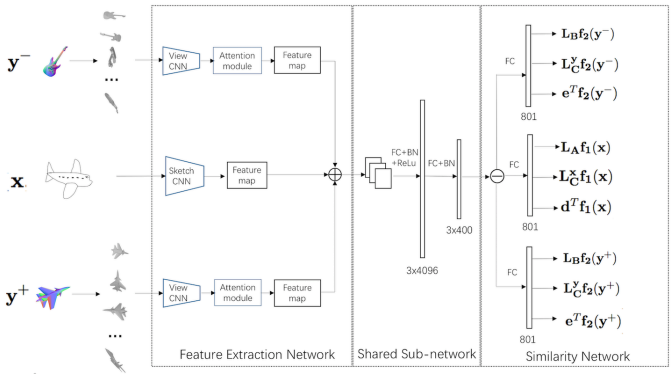

Fig. 5. The combination of feature learning and similarity model. The View CNN(MVAN) and Sketch CNN can extract feature of shape and sketch separately. The feature map of both two domains will be fed into shared sub-network, which shares parameters except BN layer. The final similarity network can calculate the similarity parameters. Different domain owns different parameters of their similarity network.

We denote the output of network as:

$$\begin{aligned}\tilde{x} &\triangleq \begin{bmatrix}\mathbf{L_A}\mathbf{f_1}(\mathbf{x}) & \mathbf{L_C^x}\mathbf{f_1}(\mathbf{x}) & \mathbf{d}^T\mathbf{f_1}(\mathbf{x})\end{bmatrix}^T\\ \tilde{y}^+ &\triangleq \begin{bmatrix}\mathbf{L_B}\mathbf{f_2}(\mathbf{y}^+) & \mathbf{L_C^y}\mathbf{f_2}(\mathbf{y}^+) & \mathbf{e}^T\mathbf{f_2}(\mathbf{y}^+)\end{bmatrix}^T\\ \tilde{y}^- &\triangleq \begin{bmatrix}\mathbf{L_B}\mathbf{f_2}(\mathbf{y}^-) & \mathbf{L_C^y}\mathbf{f_2}(\mathbf{y}^-) & \mathbf{e}^T\mathbf{f_2}(\mathbf{y}^-)\end{bmatrix}^T\end{aligned}$$

As in [8], we introduce three auxiliary matrix:

$$\begin{aligned}\mathbf{P_1} &= \begin{bmatrix}\mathbf{I}^{r\times r} & \mathbf{0}^{r\times(r+1)}\end{bmatrix}\\ \mathbf{P_2} &= \begin{bmatrix}\mathbf{0}^{r\times r} & \mathbf{I}^{r\times r} & \mathbf{0}^{r\times 1}\end{bmatrix}\\ \mathbf{P_3} &= \begin{bmatrix}\mathbf{0}^{1\times 2r} & 1^{1\times 1}\end{bmatrix}^T\end{aligned}$$

where r represents the feature dimension of shared network, $\mathbf{I}$ is identity matrix ans $\mathbf{0}$ is zero matrix. Then the similarity model can be written as:

$$\begin{aligned}\tilde{S}(x,y^+) =&(\mathbf{P_1}\tilde{x})^T\mathbf{P_1}\tilde{x} + (\mathbf{P_1}\tilde{y}^+)^T\mathbf{P_1}\tilde{y}^+ \\ &- 2(\mathbf{P_2}\tilde{x})^T\mathbf{P_2}\tilde{y}^+ + 2\mathbf{P_3}^T\tilde{x} + 2\mathbf{P_3}^T\tilde{y}^+ + f\end{aligned} \tag{4}$$

$$\begin{aligned}\tilde{S}(x,y^-) =&(\mathbf{P_1}\tilde{x})^T\mathbf{P_1}\tilde{x} + (\mathbf{P_1}\tilde{y}^-)^T\mathbf{P_1}\tilde{y}^- \\ &- 2(\mathbf{P_2}\tilde{x})^T\mathbf{P_2}\tilde{y}^- + 2\mathbf{P_3}^T\tilde{x} + 2\mathbf{P_3}^T\tilde{y}^- + f\end{aligned}$$

The objective function can be represented as:

$$\tilde{l}_{triplet}(\mathbf{x},\mathbf{y}^+,\mathbf{y}^-) = max(L(\mathbf{x},\mathbf{y}^+,\mathbf{y}^-), C) \tag{5}$$

where

$$\begin{aligned}L(\mathbf{x},\mathbf{y}^+,\mathbf{y}^-) =&(\mathbf{P_1}\tilde{y}^+)^T\mathbf{P_1}\tilde{y}^+ - 2(\mathbf{P_2}\tilde{x})^T\mathbf{P_2}\tilde{y}^+ + 2\mathbf{P_3}^T\tilde{y}^+ \\ &- (\mathbf{P_1}\tilde{y}^-)^T\mathbf{P_1}\tilde{y}^- + 2(\mathbf{P_2}\tilde{x})^T\mathbf{P_2}\tilde{y}^- - 2\mathbf{P_3}^T\tilde{y}^-\end{aligned}$$

The gradients with respect to $\tilde{x}$, $\tilde{y}^+$, $\tilde{y}^-$ are

$$\frac{\partial\tilde{l}}{\partial\tilde{x}} = (2\mathbf{P_2}^T\mathbf{P_2}\tilde{y}^- - 2\mathbf{P_2}^T\mathbf{P_2}\tilde{y}^+) \times I_{\tilde{S}(x,y^+)-\tilde{S}(x,y^-)>C}$$

$$\begin{aligned}\frac{\partial\tilde{l}}{\partial\tilde{y}^+} =&[2\mathbf{P_1}^T\mathbf{P_1}\tilde{y}^+ - 2(\mathbf{P_2}\tilde{x})^T\mathbf{P_2} + 2\mathbf{P_3}^T] \\ &\times I_{\tilde{S}(x,y^+)-\tilde{S}(x,y^-)>C}\end{aligned}$$

$$\begin{aligned}\frac{\partial\tilde{l}}{\partial\tilde{y}^-} =&[2(\mathbf{P_2}\tilde{x})^T\mathbf{P_2} - 2\mathbf{P_1}^T\mathbf{P_1}\tilde{y}^- - 2\mathbf{P_3}^T] \\ &\times I_{\tilde{S}(x,y^+)-\tilde{S}(x,y^-)>C},\end{aligned}$$

where $I_{condition} = 1$ if *condition* is true, otherwise $I_{condition} = 0$. Hence the objective function in (5) can be trained using back propagation.

Once the model is optimized, one can calculate the similarity parameters between sketch and shape using forward propagation and compute the similarity according to (4). After ranking the similarity distance between sketch and all the shapes, the most similar shape can be obtained by the minimal distance. Since the calculation of shapes can be computed offline, the performance of retrieval can be very fast.

IV. Experiments

In this section, we evaluate our method on SHREC'13 [19] and SHREC'14 [20] and compare to the state-of-the-art methods. We further conduct an experiment to explore the performance of different numbers of projections and visualize the effect of our proposed attention module.

A. Datasets

We test our method on SHREC'13 and SHREC'14 sketch track benchmark datasets. SHREC'13 is constructed based on a human sketch dataset [21] and Princeton shape benchmark [22]. It contains 7200 sketches and 1258 shapes, totally with 90 class. Each class consists 80 sketches, where 50 samples for training and 30 samples for testing. The construction of SHREC'14 is similar to SHREC'13. Instead, it contains 171 class, totally with 13680 sketches ans 8987 shapes, which is more challenge than SHREC'13. We show some samples of SHREC'13 in Fig. 6.

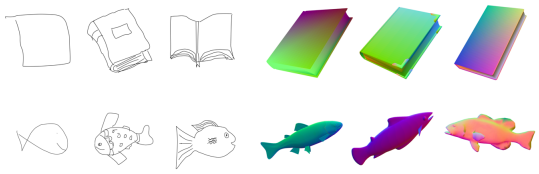

Fig. 6. The first row are sketches and shapes of book and the second row are fish. Both samples are drawn from SHREC'13 dataset.

In our experiment, we initialize the convolution layers of AlexNet and MVAN(except for attention module) by the model pre-trained on ImageNet images from 1K parameters. Other parameters of the network are randomly initialized. The output feature size of shared network is set to 400, resulting in the feature size of similarity network to be 801. The initial learning rate is set to 0.001 and we decrease it by 0.8 ratio after every 10000 iterations. We employ the mini-batch gradient descent to train our method and the batch size is set to 20. For every training sketch sample, we randomly choose 10 shapes from the same class as positive samples and choose 10 shapes from different class as negative samples. Since the number of training sketches is limited, we adopt the sketch deformation technique in [23] to augment the training samples so as to reduce the overfitting of the network. We report our results using the following evaluation methods: precision-recall curve(PR curve), nearest neighbor(NN), first tier(FT), second tier(ST), E-measure(E), discounted cumulated gain(DCG) and mean average precision(mAP).

B. Comparative Results

1) SHREC'13 benchmark dataset: We compare our method with CDMR [24], HOG-SIL [25], SBR-2D-3D [10], SBR-VC [25], Siamese Network [1] and LWBR [3] on SHREC'13 benchmark. The comparison result of NN, FT, ST, E, DCG, mAP is shown in table 1. It is notable that methods based on deep learning (Siamese, LWBR, Ours) are superior to those based on hand crafted features. While compared with other methods utilizing neural network, our method also outperforms Siamese CNN by 30.9% and LWBR by 2.6% in mAP respectively. The precision-recall curves in Fig. 7 demonstrates that our method can

significantly outperform others and keep high precision even when recall reaches 1.

TABLE I
Retrieval results on SHREC'13 dataset.

Method	NN	FT	ST	E	DCG	mAP
CDMR	0.279	0.203	0.296	0.166	0.458	0.250
HOG-SIL	0.110	0.069	0.107	0.061	0.307	0.086
SBR-2D-3D	0.132	0.077	0.124	0.074	0.327	0.095
SBR-VC	0.164	0.097	0.149	0.085	0.348	0.116
Siamese	0.405	0.403	0.548	0.287	0.607	0.469
LWBR	0.712	0.725	0.785	0.369	0.814	0.752
Ours	0.744	0.737	0.848	0.403	0.846	0.778

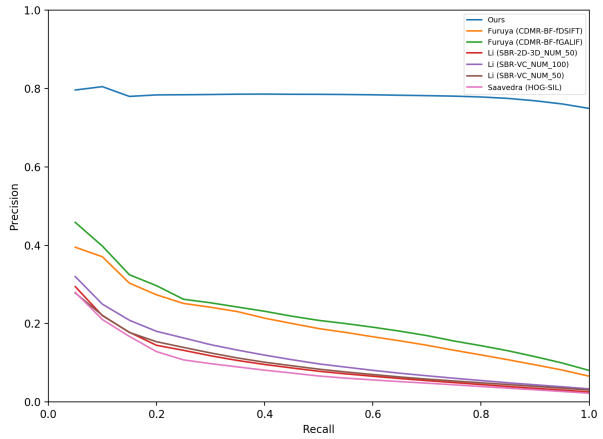

Fig. 7. The precision-recall curve for the CDMR, HOG-SIL, SBR-2D-3D, SBR-VC and our method on SHREC'13 benchmark dataset.

2) SRREC'14 benchmark dataset: We also evaluate our method on SHREC'14 dataset and compare with several methods: CDMR [24], SBR-VC [25], SCMR-OPHOG [26], BOF-JESC [27], Siamese [1], LWBR [3] and MVPR [2]. The result of NN, FT, ST, E, DCG, mAP is shown in table 2. Due to the complexity of SHREC'14 dataset, most methods including deep learning can't achieve high performance. However, our method still outperforms MVPR by 2.0% in mAP, which achieves best accuracy among the state-of-the-art methods. In Fig. 8, the PR curve further validates our proposed method is more stable than others since it keeps nearly 60% at precision when recall reaches 50% and decreases slowly as recall increases until the latter reaches 80%.

Furthermore, we perform another experiment to evaluate the effect of view number on our method. The retrieval results of different view numbers for our method are reported in table 3. It can be observed that the mAP of our method improves from 50.8% of 2 view to 59.1% of 20 views. This is reasonable as more projections provide richer information about 3D shape.

TABLE II
Retrieval results on SHREC'14 dataset.

Methods	NN	FT	ST	E	DCG	mAP
CDMR	0.109	0.057	0.089	0.041	0.328	0.054
SBR-VC	0.095	0.050	0.081	0.037	0.319	0.050
SCMR-OPHOG	0.160	0.115	0.170	0.079	0.376	0.131
BOF-JESC	0.086	0.043	0.068	0.030	0.310	0.131
Siamese	0.239	0.212	0.316	0.140	0.496	0.228
LWBR	0.403	0.378	0.455	0.236	0.581	0.401
MVPR	0.546	0.506	0.642	0.301	0.715	0.543
Ours	0.614	0.555	0.687	0.310	0.775	0.563

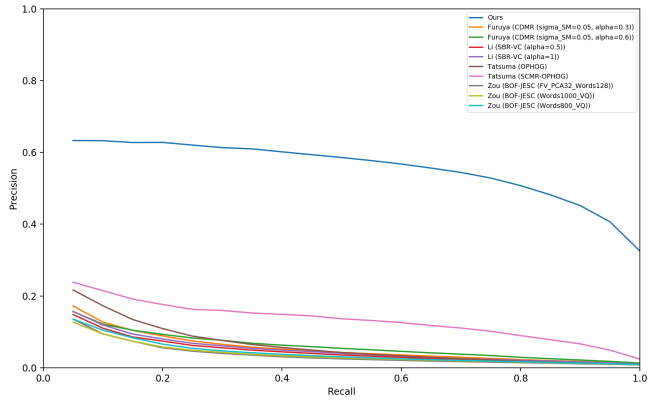

Fig. 8. The precision-recall curve for the CDMR SBR-VC SCMR-OPHOG BOF-JESC and our method on SHREC'14 benchmark dataset.

C. Visualization of Attention Mask

To explore the effect of our multi-view attention network, we visualize the attention mask using method in [12]. The visualization of attention mask for three shapes are shown in Fig. 9. One can note that the mask focus more (with deeper red) on the prominent parts of projection and ignores the less relevant parts. In the knife example, our model pays more attention on point and handle, which are distinctive for knife. For shark example, it highlights fishtail and head. And for spider shape our model learns that the feet can distinguish spider from other shapes easily. Another interesting example happens in the fifth projection of knife shape and the forth and sixth projections of shark shape. As these projections are

TABLE III
The effect of different view numbers on our method

view number	NN	FT	ST	E-Measure	DCG	mAP
2	0.570	0.496	0.631	0.291	0.741	0.508
4	0.586	0.519	0.655	0.297	0.756	0.528
8	0.598	0.535	0.671	0.305	0.765	0.547
12	0.614	0.555	0.687	0.310	0.775	0.563
20	0.629	0.570	0.693	0.330	0.781	0.591

difficult to be recognized even by human eyes, our model applies little weight on them, meaning that our method has the ability to filter 'noisy' projections.

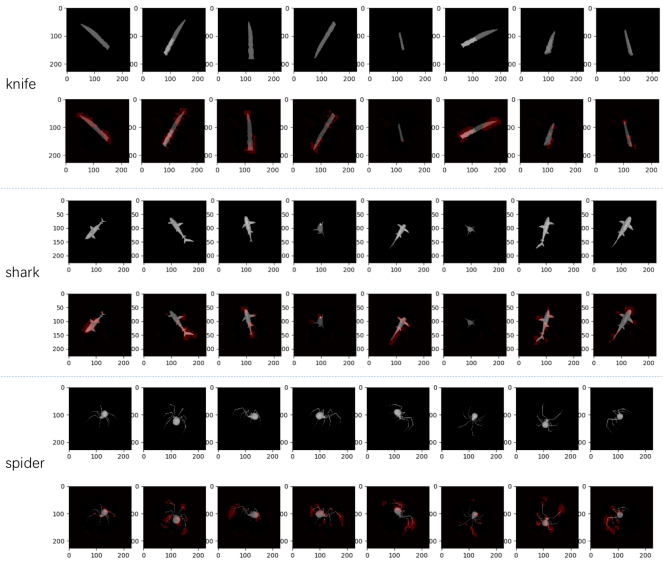

Fig. 9. Attention mask on different projections. The first row in each shape displays the projections from 8 different view points and the attention mask of each projection is shown on the second row. The deeper of the red color, the heavier of the weight.

V. Conclusion

In this paper, we propose a multi-view attention network to obtain a compact representation of shapes. And a generalize similarity model is applied to calculate the similarity distance between features of sketch and shape. We adopt triplet ranking loss to train feature learning network and similarity model in an end-to-end manner. Extensive experiments on SHREC'13 and SGREC'14 benchmark demonstrate that our method can achieve promising results.

References

[1] F. Wang, L. Kang, and Y. Li, "Sketch-based 3D Shape Retrieval using Convolutional Neural Networks," ArXiv e-prints, Apr. 2015.

[2] H. Li, H. Wu, X. He, S. Lin, R. Wang, and X. Luo, "Multi-view pairwise relationship learning for sketch based 3d shape retrieval," in 2017 IEEE International Conference on Multimedia and Expo (ICME), July 2017, pp. 1434–1439.

[3] J. Xie, G. Dai, F. Zhu, and Y. Fang, "Learning barycentric representations of 3d shapes for sketch-based 3d shape retrieval," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, pp. 3615–3623.

[4] B. T. Phong, "Illumination for computer generated pictures," Commun. ACM, vol. 18, no. 6, pp. 311–317, Jun. 1975. [Online]. Available: http://doi.acm.org/10.1145/360825.360839

[5] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," ArXiv e-prints, Mar. 2015.

[6] H. Su, S. Maji, E. Kalogerakis, and E. G. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in Proc. ICCV, 2015.

[7] M. Eitz, R. Richter, T. Boubekeur, K. Hildebrand, and M. Alexa, "Sketch-based shape retrieval," ACM Trans. Graph. (Proc. SIGGRAPH), vol. 31, no. 4, pp. 31:1–31:10, 2012.

[8] L. Lin, G. Wang, W. Zuo, X. Feng, and L. Zhang, "Cross-domain visual matching via generalized similarity measure and feature learning," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1089–1102, June 2017.

[9] J. M. Saavedra, B. Bustos, T. Schreck, S. M. Yoon, and M. Scherer, "Sketch-based 3D Model Retrieval using Keyshapes for Global and Local Representation," in Eurographics Workshop on 3D Object Retrieval, M. Spagnuolo, M. Bronstein, A. Bronstein, and A. Ferreira, Eds. The Eurographics Association, 2012.

[10] B. Li and H. Johan, "Sketch-based 3d model retrieval by incorporating 2d-3d alignment," Multimedia Tools and Applications, vol. 65, no. 3, pp. 363–385, Aug 2013. [Online]. Available: https://doi.org/10.1007/s11042-012-1009-0

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," CoRR, vol. abs/1409.1556, 2014.

[12] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ser. ICML'15. JMLR.org, 2015, pp. 2048–2057. [Online]. Available: http://dl.acm.org/citation.cfm?id=3045118.3045336

[13] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, pp. 3242–3250.

[14] Y. Peng, X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," IEEE Transactions on Image Processing, vol. 27, no. 3, pp. 1487–1500, March 2018.

[15] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015, pp. 842–850.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," ArXiv e-prints, Jun. 2017.

[17] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," CoRR, vol. abs/1409.0473, 2014. [Online]. Available: http://arxiv.org/abs/1409.0473

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 1097–1105. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999134.2999257

[19] B. Li, Y. Lu, A. Godil, T. Schreck, M. Aono, H. Johan, J. M. Saavedra, and S. Tashiro, "SHREC'13 Track: Large Scale Sketch-Based 3D Shape Retrieval," in Eurographics Workshop on 3D Object Retrieval, U. Castellani, T. Schreck, S. Biasotti, I. Pratikakis, A. Godil, and R. Veltkamp, Eds. The Eurographics Association, 2013.

[20] B. Li, Y. Lu, C. Li, A. Godil, T. Schreck, M. Aono, M. Burtscher, H. Fu, T. Furuya, H. Johan, J. Liu, R. Ohbuchi, A. Tatsuma, and C. Zou, "Extended large scale sketch-based 3d shape retrieval," in Proceedings of the 7th Eurographics Workshop on 3D Object Retrieval, ser. 3DOR '15. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2014, pp. 121–130. [Online]. Available: http://dx.doi.org/10.2312/3dor.20141058

[21] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" ACM Trans. Graph. (Proc. SIGGRAPH), vol. 31, no. 4, pp. 44:1–44:10, 2012.

[22] "The princeton shape benchmark," in Proceedings of the Shape Modeling International 2004, ser. SMI '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 167–178. [Online]. Available: https://doi.org/10.1109/SMI.2004.63

[23] Q. Yu, Y. Yang, F. Liu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Sketch-a-net: A deep neural network that beats

humans," International Journal of Computer Vision, vol. 122, no. 3, pp. 411–425, May 2017. [Online]. Available: https://doi.org/10.1007/s11263-016-0932-3

[24] T. Furuya and R. Ohbuchi, "Ranking on cross-domain manifold for sketch-based 3d model retrieval," in 2013 International Conference on Cyberworlds, Oct 2013, pp. 274–281.

[25] B. Li, Y. Lu, A. Godil, T. Schreck, B. Bustos, A. Ferreira, T. Furuya, M. J. Fonseca, H. Johan, T. Matsuda, R. Ohbuchi, P. B. Pascoal, and J. M. Saavedra, "A comparison of methods for sketch-based 3d shape retrieval," Comput. Vis. Image Underst., vol. 119, pp. 57–80, Feb. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2013.11.008

[26] B. Li, Y. Lu, C. Li, A. Godil, T. Schreck, M. Aono, M. Burtscher, Q. Chen, N. K. Chowdhury, B. Fang, H. Fu, T. Furuya, H. Li, J. Liu, H. Johan, R. Kosaka, H. Koyanagi, R. Ohbuchi, A. Tatsuma, Y. Wan, C. Zhang, and C. Zou, "A comparison of 3d shape retrieval methods based on a large-scale benchmark supporting multimodal queries," Computer Vision and Image Understanding, vol. 131, pp. 1 – 27, 2015, special section: Large Scale Data-Driven Evaluation in Computer Vision. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314214002100

[27] C. Zou, C. Wang, Y. Wen, L. Zhang, and J. Liu, "Viewpoint-aware representation for sketch-based 3d model retrieval," IEEE Signal Processing Letters, vol. 21, no. 8, pp. 966–970, Aug 2014.