Guidance Cleaning Network for Sketch-Based 3D Shape Retrieval

Qi Liu\textsuperscript{1,a}, Shengjie Zhao\textsuperscript{1,b}\textsuperscript{*}

\textsuperscript{1}School of Electronics and Information Engineering, Tongji University, Shanghai, 201804 - China
\textsuperscript{a}email: liuqiskan@tongji.edu.cn
\textsuperscript{b}\textsuperscript{*}Corresponding author's email: shengjiezha@tongji.edu.cn

Abstract. The sketch-based query for 3D shapes has drawn growing attention from the academic community due to its flexibility and accessibility. However, most recent works focus on reducing the cross-modality discrepancy between 2D sketches and 3D shapes, which neglect the influence caused by noise information in some low-quality free-hand sketches. To address this issue, we proposed a novel network to decrease the impact in two ways: 1) an attention weighting module to detect the noisy samples by a self-attention mechanism; 2) a data cleaning module to clear up low-quality sketches according to a ranking regularization. Experiments on two widely-used datasets demonstrate the effectiveness of our method.

1. Introduction

With the rapid growth of computing power of microchips, 3D shapes are used more and more widely in various industries. How to retrieve 3D shapes quickly and accurately has become a hot topic in the computer vision community. Compared with the use of texts or 3D shapes as the retrieval method, sketch-based 3D shape retrieval is more intuitive, clear and convenient because of the popularity of touch-based devices.

Fig.1. Some examples in the datasets, where samples in same column belong to same category.

However, the cross-modality discrepancy between 2D images and 3D shapes as well as large variations among sketches make it a very challenging task. In recent years, traditional methods based on hand-craft features have gradually been surpassed by methods using deep neural networks. Specifically, most deep learning approaches follow the pipeline in MVCNN [2] to project 3D shapes into several 2D views. Then, the features of sketches and views are extracted by two CNNS respectively,
and finally the features are converted into a common space to achieve feature alignment. Nevertheless, those methods focus on reducing the cross-modality discrepancy while neglect the noisy information in sketches. As Fig.1 shows, since the painting skills vary among individuals, there are relatively larger intra-class variations and smaller inter-class variations among hand-drawing sketches compared to 3D shapes. Those low-quality sketches contain incorrect semantic information, and over-fitting to them will be harmful to the retrieval precision.

Lately, deep learning methods have been applied to detect noisy data. Specifically, a deep convolutional neural network is used to learn a weighting coefficient representing the level of importance, which increases from low-quality samples to high-quality samples. The learning of noisy data is suppressed to prevent the impact caused by low-quality images.

![Diagram](image)

**Fig.2.** The block diagram of the proposed method. Our network consists of 3 modules: (1) knowledge distillation module to transfer features of sketches to pre-learned feature space of 3D shapes, (2) attention weighting module to detect the noisy sketches, (2) data cleaning module to clear up low-quality sketches.

To address the problems mentioned above, we proposed a novel framework named Guidance Cleaning Network, as shown in Fig.2. For the cross-modality discrepancy, we follow the pipeline proposed in CGN [7], which utilized knowledge distillation to transfer features of sketches to pre-learned feature space of 3D shapes; For the problem of high variations among sketches, we apply the attention mechanism which learns the weight of each sketch to measure its importance to the loss function. Then, the ranking regularization in Self-Cure network [13] is adopted to split the mini-batch into high importance group and low importance group. Finally, the data cleaning module is adopted to clear up low-quality sketches. In order to combine the two aspects above tightly, we introduced a novel loss function that is mainly composed of regression loss term ($L_r$) and classification loss term ($L_c$). Specifically, $L_r$ is used to constrain the process of transfer learning, which transfers the features of sketches to the feature space of 3D shapes, and $L_c$ can accelerate the transfer learning process as well as assist the attention module to learn the importance weight.

Overall, the contributions of our paper can be summarized as follows:

1. We innovatively proposed a Guidance Cleaning Network for sketch-based 3D shape retrieval. A data cleaning module is adopted to clear up sketches according to importance weights gained by a self-attention mechanism, which can reduce the impact of low-quality data.
2. We introduced a novel loss function to combine two aspects that influence retrieval precision immensely, namely large cross-modality discrepancy and high variations among sketches.

3. The experimental results of our method exceed the state-of-art approaches on two widely-used datasets.

2. Related work
In this section, we review some related works of sketch-based 3D shape retrieval and noise detection.

2.1. sketch-based 3D shape retrieval
The main contributions of existing methods are to extract discriminative and modality-invariant features, which could be roughly classified into traditional methods and deep learning methods. For traditional methods, features are extracted manually [1][4]. However, the discriminative abilities of hand-craft features are limited, which have been surpassed by deep learning methods gradually. For deep learning methods, Su H, et al. [2] proposed MVCNN that extracts a compact shape descriptor by combining information from multiple 3D shape projections. Based on MVCNN method, Y Lei, et al. [5] proposed the "point-to-subspace" metric learning for feature alignment. [7] adopted distillation network to reduce the cross-modal differences. Specifically, they used the features extracted through a pre-trained teacher network to guide the training of the student network. However, those methods neglect the noise information in sketches, and overfitting them will be harmful to the retrieval results.

2.2. noise detection
Noise detection for reliability assessment and risk analysis has been studied for a long time [9]. Some approaches exploit a validation set to reduce the impact of noisy data [11]. Other methods leverage extra neural networks to learn sample weights of dataset [14-15]. For instance, Self-Cure network [13] relabeled noisy samples according to the weights learned by a self-attention mechanism. Inspired by [13], we incorporate self-attention into the student network to decrease the impact of noisy samples.

3. Proposed Method

3.1 Motivation
Due to the subjectivity of free-hand paintings, there are big intra-class variations and small inter-class variations among sketches. But most methods focus on reducing cross-modal differences while ignore the variations of sketches. In recent years, noise detection based on deep learning has been applied in the field of facial expression recognition, which inspired us to find and clear up noisy samples by deep neural networks. Different from relabeling of those samples in [13], the noise information in sketches mainly comes from the subjectivity of paintings, and it is more appropriate to simply remove these low-quality samples. Therefore, we use the data cleaning module to clear up those samples, namely the samples that have low importance weight and are incorrectly predicted during the training process.

3.2 Network Architecture
To reduce the cross-modality discrepancy as well as intra-class variations, we proposed a Guidance Cleaning Network, which consists of 3 modules: i) knowledge distillation module, ii) attention weighting module, iii) data cleaning module.

(i) knowledge distillation module
Inspired by CGN [7], we use the pre-learned 3D shape feature space to guide the learning process of sketch features. Specifically, we follow the MVCNN method to train a classification network and extract class centers of 3D shapes. Let \( \{c_1, c_2, \ldots, c_N\} \in \mathbb{R}^{D \times N} \) denotes the class center features of \( N \) kinds of 3D shapes. We assume that the 3D shapes and sketches of the same type have the same semantic information, that is, the same class centers. So, we can use the class centers of 3D shapes to guide the learning of the sketch network. In transfer learning stage, the guidance loss in CGN [7] is adopted to transfer the features of sketches effectively, which is formulated as,
\[ L_r = \frac{1}{M} \sum_{i=1}^{M} \frac{f_i \cdot c_{y_i}}{\|f_i\|_2 \cdot \|c_{y_i}\|_2} \tag{1} \]

where \( M \) is the size of mini-batch, \( f_i \) is the extracted feature of the \( i_{th} \) sketch and \( y_i \) denotes its corresponding label.

(ii) attention weighting module

In sketch network, we introduce attention weighting module to detect the noisy samples, where poor-painted sketches are expected to get lower weights than high quality sketches. Specifically, attention module consists of several fully-connected layers following by a sigmoid function. Let \( \alpha_i \) denotes the weight of \( i_{th} \) sketch. To constrain the optimizing of \( \alpha \), we choose Logit-Weighted Cross-Entropy loss in Self-Cure Network [13]:

\[ L_c = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{\alpha_i w_{y_i}^T f_i}}{\sum_{j=1}^{C} e^{\alpha_i w_{y_j}^T f_i}} \tag{2} \]

where \( W_j \) is the \( j_{th} \) classifier.

(iii) data cleaning module

To find out low-importance sketches, the attention weights are ranked in descending. Then, each mini-batch is divided into two groups with a ratio \( r \). Formally, we introduce the rank regularization loss in [13] to make sure those two groups are separated by a margin \( m \) as follows:

\[ L_{RR} = \max \{0, m - \frac{1}{r \cdot N} \sum_{i=0}^{rN} \alpha_i - \frac{1}{(1-r) \cdot N} \sum_{i=0}^{(1-r)N} \alpha_i \} \tag{3} \]

A sample in low-weights group is removed only if the predictive label is inconsistent with original one as well as the corresponding probability is bigger than the one of given label with a threshold \( t \), which is defined as follows:

\[ d = \begin{cases} 1 & \text{if } P_p - P_g > t, \\ 0 & \text{otherwise,} \end{cases} \tag{4} \]

where \( d = 1 \) means this sample will be deleted, \( P_p \) denotes corresponding probability of predictive label, \( P_g \) is the predictive probability of given label and \( r \) is the threshold.

To sum up, the loss function of student network can be formulated as follows:

\[ L = L_r + L_c + L_{RR} \tag{5} \]

Our loss function aims to draw the class centers of sketches and 3D shapes closer after clear away noisy samples.

4. Experiments

In this section, we first introduce 2 public datasets. Then, we demonstrate the experiment setups and performance of our approach.

4.1. Datasets

**SHREC 2013** is a large-scale benchmark [4] to evaluate approaches for sketch-based 3D shape retrieval. 7,200 sketches and 1,258 3D shapes are divided into 90 classes. There are 80 images in each class (50 for training, 30 for testing) for sketches, and about 14 samples on average per class for 3D shapes. **SHREC 2014** is a benchmark larger than SHREC 2013, containing 13,680 sketches and 8987 3D shapes from 171 classes. Similar to SHREC 2013, there are 80 sketches in each class (50 for training, 30 for testing). The number of 3D shapes per class is about 53 on average. Compared to SHREC 2013, SHREC 2014 is more challenging because of larger variations within class and more categories.

4.2. Implementation Details

The proposed method is implemented on Pytorch 1.4 and python 3.6 platform with 2 Nvidia RTX 2080Ti GPU. Several pre-trained backbones are adopted for the CNN networks of sketches and 3D
shapes, such as AlexNet [18] or ResNet-50 [19]. Specifically, we use all backbones by removing the fully-connected classifier. The MLPs consists of 3 fully connected layers (i.e., 4096-1024-256 for AlexNet, 2048-1024-256 for ResNet-50).

Followed by CGN [7], the training process is divided into two stages. In the first stage, the teacher network is trained to extract features of 3D shapes. In transfer learning stage, the mini-batch is split into high importance group and low importance group with ratio $r = 0.7$. The margin $m$ to separate those 2 groups is set to 0.15. The learning rate is initialized to be 0.001 which decays by 0.9 after every 10 epochs. The training finishes after 100 epochs. The data cleaning module works at the $i_{th}$ epoch ($i = 10$ in this paper) to remove noisy sketches, and the threshold $t$ is set to 0.2.

4.3. Evaluation of the Proposed Method

For performance evaluation, the widely adopted metrics are used as follows: precision-recall curve, nearest neighbor (NN), first tier (FT), second tier (ST), E-measure(E), discounted cumulated gain (DCG) and mean average precision (mAP) [1]. In order to demonstrate the effectiveness of the proposed method, we report our results on the above two datasets by comparing with several state-of-the-art approaches, including Siamese[17], DCML[16], LWBR[12], TCL[10], DCA[3], DPSML[5], DSSH[8] and HEAR[6].

Table 1. Quantitative comparison on SHREC 2013 and SHREC 2014. The best results are in bold font.

| Method  | Backbone | SHREC 2013 | SHREC 2014 |
|---------|----------|------------|------------|
|         |          | NN | FT | ST | EE | DCG | mAP | NN | FT | ST | EE | DCG | mAP |
| Siamese | -        | 40.5 | 40.3 | 54.8 | 28.7 | 60.7 | 46.9 | 23.9 | 21.2 | 31.6 | 14 | 49.6 | 22.8 |
| DCML    | AlexNet  | 65  | 63.4 | 71.9 | 34.8 | 76.6 | 67.4 | 27.2 | 27.5 | 34.5 | 17.1 | 49.8 | 28.6 |
| LWBR    | AlexNet  | 71.2 | 72.5 | 78.5 | 36.9 | 81.4 | 75.2 | 40.3 | 37.8 | 45.5 | 23.6 | 58.1 | 40.1 |
| TCL     | AlexNet  | 76.3 | 78.7 | 84.9 | 39.2 | 85.4 | 80.7 | 58.5 | 45.5 | 53.9 | 27.5 | 66.6 | 47.7 |
| DCA     | ResNet-50| 78.3 | 79.6 | 82.9 | 37.6 | 85.6 | 81.3 | 77  | 78.9 | 82.3 | 39.8 | 85.9 | 80.3 |
| DPMSL   | AlexNet  | 74.1 | 76.1 | 82.1 | 38.5 | 83.6 | 78.5 | 67.7 | 73.2 | 79.5 | 37.9 | 83  | 75.1 |
|         | ResNet-50| 81.9 | 83.4 | 87.5 | 41.5 | 89.2 | 85.7 | 77.4 | 79.8 | 84.9 | 41.5 | 87.7 | 81.3 |
| DSSH    | ResNet-50| 79.9 | 81.4 | 86  | 40.4 | 87.3 | 83.1 | 77.5 | 78.8 | 83.1 | 40.4 | 87  | 80.6 |
| HEAR    | ResNet-50| 82.1 | 83.7 | 87.8 | 40.9 | 88.8 | 85.4 | 79.2 | 80.7 | 84.6 | 40.9 | 87.7 | 82.2 |
| Ours    | AlexNet  | 77.2 | 80.5 | 85.5 | 40.5 | 86.4 | 81.9 | 73.9 | 76.8 | 81.3 | 39.4 | 84.7 | 78.3 |
|         | ResNet-50| 83.6 | 85.5 | 90.2 | 42.5 | 90.5 | 87.1 | 79.9 | 82.1 | 86.2 | 41.9 | 88.6 | 83.5 |

Evaluation on SHREC 2013 and SHREC 2014. The PR curves of the proposed method and state-of-the-arts on SHREC 2013 and SHREC 2014 are shown in Fig.3 and Fig.4, respectively. It is obvious that the precision rate of our method exceeds the compared methods when the recall rate changes from 0 to 1. As demonstrated in Table 1, our method also achieves best performance on SHREC 2013 and SHREC 2014 using standard evaluation metrics. For instance, the proposed method improves the mAP of DCA, DPSML, DSSH, HEAR by 3.98%, 2.71%, 3.60%, 1.58% on SHREC 2014 with ResNet-50 as backbone. Due to SHREC 2014 being more challenging than SHREC 2013, the experimental results have proven the effectiveness of our method.
5. Conclusion
This paper presents a novel Guidance Cleaning Network for sketch-based 3D shape retrieval, which consists of 3 modules including knowledge distillation module, attention weighting module and data cleaning module. The training process can be divided into two stages, namely teacher network training stage and transfer learning stage. The experimental results demonstrate the superiority of our method.

References
[1] Li B, Lu Y, Godil A, et al. (2014) A comparison of methods for sketch-based 3D shape retrieval. Computer Vision and Image Understanding, 119: 57-80.
[2] Su, H., Maji, S., et al. (2015) Multi-view convolutional neural networks for 3d shape recognition. In: IEEE international conference on computer vision. Santiago. pp. 945-953.
[3] Chen, J., & Fang, Y. (2018) Deep cross-modality adaptation via semantics preserving adversarial learning for sketch-based 3d shape retrieval. In: European Conference on Computer Vision. Munich. pp. 605-620.
[4] Godil A, Li B, Lu Y, et al. (2013) SHREC’13 Track: Large Scale Sketch-Based 3D Shape Retrieval. In: Eurographics Workshop on 3D Object Retrieval. Girona. pp. 89-96.
[5] Lei Y, Zhou Z, Zhang P, et al. (2019) Deep point-to-subspace metric learning for sketch-based 3D shape retrieval. Pattern Recognition, 96: 106-116.
[6] J. Chen, J. Qin*, Y. Shen, L. Liu, F. Zhu and L. Shao. (2020) Learning Attentive and Hierarchical Representations for 3D Shape Recognition. In: European Conference on Computer Vision. Online. pp. 106-123
[7] Dai, W., & Liang, S. (2020) Cross-Modal Guidance Network for Sketch-Based 3D Shape Retrieval. In: International Conference on Multimedia and Expo. London. pp. 1-6.
[8] Chen, J., et al. (2019) Deep sketch-shape hashing with segmented 3d stochastic viewing. In: IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA. pp. 791-800.
[9] M Elisabeth Paté-Cornell. (1996) Uncertainties in risk analysis: Six levels of treatment. Reliability Engineering & System Safety, 54(2-3):95-111.
[10] He, X., Zhou, Y., Zhou, et al. (2018) Triplet-center loss for multi-view 3d object retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition. Munich. pp. 1945-1954.
[11] Jenni, S., & Favaro, P. (2018) Deep bilevel learning. In: European Conference on Computer Vision. Munich. pp. 618-633.
[12] Xie, J., Dai, G., Zhu, F., & Fang, Y. (2017) Learning barycentric representations of 3d shapes for sketch-based 3d shape retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition. Hawaii. pp 5068-5076.
[13] Wang, K., Peng, X., Yang, J., Lu, S., & Qiao, Y. (2020). Suppressing uncertainties for large-scale facial expression recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. online. pp. 6897-6906.
[14] Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., ... & Sugiyama, M. (2018) Co-teaching: Robust training of deep neural networks with extremely noisy labels. In: Conference on Neural Information Processing Systems. Montréal. pp. 8527–8537.
[15] Jiang, L., Zhou, Z., Leung, T., Li, L. J., & Fei-Fei, L. (2018) Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In: International Conference on Machine Learning. Beijing. pp. 2304-2313.

[16] Dai, G., Xie, J., Zhu, F., & Fang, Y. (2017) Deep correlated metric learning for sketch-based 3d shape retrieval. In: AAAI Conference on Artificial Intelligence. San Francisco. 31(1): 4002-4008.

[17] Wang, F., Kang, L., & Li, Y. (2015) Sketch-based 3d shape retrieval using convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition. Boston. pp. 1875-1883.

[18] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012) Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 25: 1097–1105.

[19] He, K., Zhang, X., et al. (2016) Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition. Las Vegas. pp. 770-778).