Semantic similarity metric learning for sketch-based 3D shape retrieval

Yu Xia, Shuangbu Wang(✉), Lihua You, and Jianjun Zhang

National Centre for Computer Animation, Bournemouth University, Poole, UK
swang1@bournemouth.ac.uk

Abstract. Since the development of the touch screen technology makes sketches simple to draw and obtain, sketch-based 3D shape retrieval has received increasing attention in the community of computer vision and graphics in recent years. The main challenge is the big domain discrepancy between 2D sketches and 3D shapes. Most existing works tried to simultaneously map sketches and 3D shapes into a joint feature embedding space, which has a low efficiency and high computational cost. In this paper, we propose a novel semantic similarity metric learning method based on a teacher-student strategy for sketch-based 3D shape retrieval. We first extract the pre-learned semantic features of 3D shapes from the teacher network and then use them to guide the feature learning of 2D sketches in the student network. The experiment results show that our method has a better retrieval performance.

Keywords: Sketch · 3D shape · Retrieval · Metric learning · Semantic feature

1 Introduction

The virtual 3D shape plays an increasingly important role in our daily lives due to the rapid development of digitalization techniques, such as visual effects, medical imaging and 3D printing. How to retrieve a desired 3D shape among a great number of 3D shapes is a popular research topic in many years [1,2,3,4]. Compared to using texts and 3D shapes as queries, sketches can easily describe the detailed information of complex 3D shapes, and are also more intuitive and convenient for humans to use. Therefore, sketch-based 3D shape retrieval has attracted considerable attention in the community of computer vision and graphics [5,6].

The main challenge for sketch-based 3D shape retrieval is the big domain discrepancies [7]. First, sketches are represented in a 2D space while 3D shapes are embodied in a 3D space, so their heterogenous data structures make it extremely difficult to directly retrieve 3D shapes from a query sketch. Second, sketches are abstract free-hand drawings, which usually consist of several simple lines and contain very limited information. Conversely, 3D shapes are realistic geometric objects and have many details of their shape characteristics. Third, sketches are presented with only one view of 3D shapes, and it is very hard to

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_5

find the best or most similar view of 3D shapes according to query sketches. Fig. 1 gives some examples of sketches and corresponding 3D shapes from the same class, and shows the large domain gap between them.

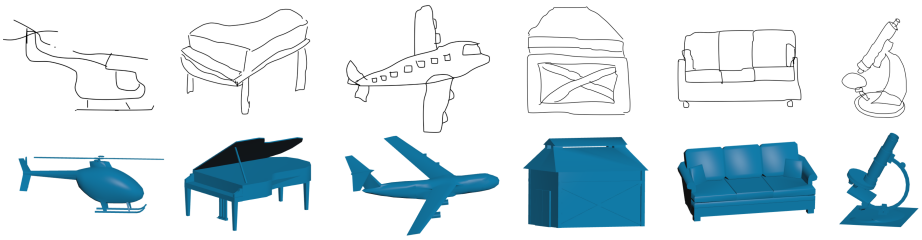

Fig. 1. Some examples of sketches and corresponding 3D shapes

In order to tackle the aforementioned challenge of sketch-based 3D shape retrieval, a variety of research efforts have been dedicated to this task, and their main purpose is to improve the retrieval accuracy. There are mainly two ways to achieve the accuracy improvement: 1) learning robust features representations for both sketches and 3D shapes [8,9,10], and 2) developing effective ranking or distance metrics between sketches and 3D shapes [7,11,12]. Due to the great success of deep convolutional neural networks (CNNs) applied in the image feature extraction in recent years, all state-of-the-art methods have used deep metric learning for sketch-based 3D shape retrieval and achieved a better retrieval accuracy compared with traditional methods [13]. However, these studies have two weaknesses. First, they address the domain discrepancy problem by mapping sketches and 3D shapes into a joint feature embedding space, where the similarity is measured using a shared loss function. It is difficult to effectively reduce the domain discrepancy because sketches and 3D shapes cannot be aligned perfectly within the same embedding space. Second, they have two different network structures to extract features of sketches and 3D shapes, respectively, and the parameters of the two networks are unshared and updated simultaneously during the training process, which leads to a high computational cost.

In this paper, we propose a novel semantic similarity metric learning to overcome the above-mentioned disadvantages of recent studies. Note that the aim of sketch-based 3D shape retrieval is to find 3D shapes belonging to the class labels of query sketches, so their label spaces are shared and can be used as a semantic embedding space. In such a semantic space, sketches and 3D shapes are aligned perfectly [7]. Inspired by the knowledge distillation technique, which uses a large teacher network to guide a small student network [14], we adopt a teacher-student strategy to obtain efficient networks for learning semantic similarity between sketches and 3D shapes. It can not only reduce the computational burden but also make the semantic features alignment easier. In our method, our proposed metric learning network consists of a teacher network and a student

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_5

network, as shown in Fig. 2. The teacher network is a pre-trained classification network based on MVCNN [15] to extract the semantic features of 3D shapes and the student network is a transfer network based on ResNet-50 [16] to learn the semantic features of sketches. We train the transfer network by the guide of a new similarity loss for optimizing the semantic feature distance between sketches and 3D shapes. The main contributions of our work are listed as follows:

- A metric learning network using the teacher-student strategy is proposed to conduct sketch-based 3D shape retrieval in a joint semantic embedding space.
- A similarity loss function is developed to optimize the semantic feature distance between sketches and 3D shapes.
- Several experiments are carried out on a large benchmark dataset of sketch-based 3D shape retrieval and show that our method outperforms other state-of-the-art methods.

The remaining parts of this paper are organized as follows. The related works on sketch-based 3D shape retrieval and the teacher-student strategy in metric learning are briefly reviewed in Sec. 2. Our proposed method is described in Sec. 3 and the experimental results and analysis are presented in Sec. 4, and finally the conclusion is drawn in Sec. 5.

2 Related works

Our proposed method is related to sketch-based 3D shape retrieval and the teacher-student strategy in metric learning. In this section, we briefly review the most related work in the two fields.

2.1 Sketch-based 3D shape retrieval

In the early stage, most sketch-based 3D shape retrieval methods relied on the handcrafted features for describing sketches and 3D shapes [4,5]. With the rapid growth of CNNs, learning-based methods have developed in recent years. Wang et al. [11] used two projection views to characterize 3D shapes and applied a Siamese network to learn a joint embedding space for sketches and 3D shapes. Zhu et al. [17] developed pyramid cross-domain neural networks to reduce the cross-domain discrepancies between sketches and 3D shapes. To address the same problem, Chen et al. [8] proposed a cross-modality adaptation model using an importance-aware metric learning method. Unlike these projection-based methods, Dai et al. [12] presented a deep correlated metric learning method to mitigate the discrepancy by directly extracting the feature of 3D shapes, and Qi et al. [7] used the PointNet network to extract 3D shape features and developed a deep cross-domain semantic embedding model. Chen et al. [13] developed a deep sketch-shape hashing framework for sketch-based 3D shape retrieval with a stochastic sampling strategy for 3D shapes and a binary coding strategy for learning discriminative binary codes. However, most of these retrieval methods

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_5

have two operative networks which cause a high computational cost. Besides, since they directly mapped features into a joint embedding space, it is difficult to effectively reduce the domain discrepancy.

2.2 Teacher-student strategy in metric learning

Since Hinton et al. [14] showed that a complex and powerful teacher model can guide the training of a small student network which can decrease the inference time and improve its generalization ability, this teacher-student strategy has received attention in the field of metric learning. Chen et al. [18] proposed cross sample similarities for knowledge transfer in deep metric learning, and modified the classical list-wise rank loss to bridge teacher networks and student networks. Yu et al. [19] presented a network distillation to compute image embeddings with small networks and developed two loss functions to communicate teacher and student networks. For the sketch-based 3D shape retrieval, Dai and Liang [20] proposed a cross-modal guidance network by using teacher-student strategy and used pre-learned features of 3D shapes to guide the feature learning of 2D sketches. However, their method cannot effectively minimize between-class similarity as well as maximize within-class similarity.

3 Method

3.1 Network Architecture

The network architecture of our proposed sketch-based 3D shape retrieval method is described in Fig. 2, which consists of a teacher network and a student network. Since sketches are abstract simple lines with limited information and 3D shapes are realistic geometric objects with more details, we select 3D shapes as the input of the teacher network and extract the semantic features from them to guide the training of the student network that takes sketches as input. By using the similarity loss to measure the cosine distance between sketches and 3D shapes, the features of sketches are optimized and gradually close to the pre-learned semantic features of 3D shapes during the training process of the student network.

In the teacher network, we apply the MVCNN [15] architecture, including CNN_1 and CNN_2, which are connected by a view-pooling layer, to represent multi-views of 3D shapes and extract the semantic features. First, we render a 3D shape from 12 different views by placing 12 virtual cameras around it every 30 degrees. Since there is still a big discrepancy between rendered images and sketches, we adopt the classic canny edge detector [21] to extract the edges of rendered images which are similar to the sketch lines. After that, the edge images are passed through CNN_1 separately to obtain view based features. Note that all branches of CNN_1 share the same parameters. In order to synthesize the information from all views into a single, we use element-wise maximum operation across the views in the view-pooling layer. Finally, these pooled feature maps are

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_5

passed through CNN_2 to obtain the shape descriptor. After finishing training the teacher network, we make all data of 3D shapes pass through the teacher network and obtain the pre-learned semantic features of 3D shapes.

In the student network, we adopt a transfer network CNN_3 to learn the semantic features of sketches. The input sketches are directly passed through CNN_3 to obtain the features. The student network is trained according to the optimization objective function, i. e., the similarity loss, which is guided by the pre-learned semantic features of 3D shapes.

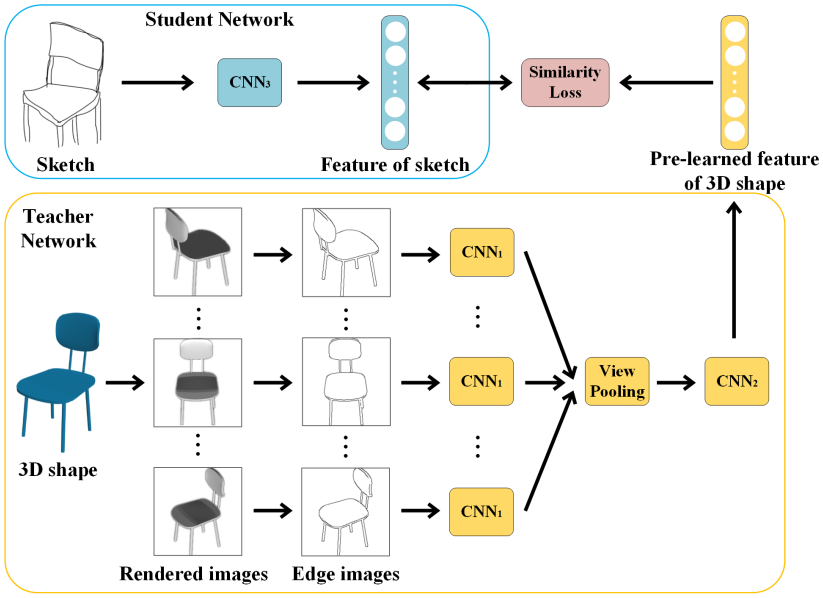

Fig. 2. The network architecture of our method.

3.2 Similarity loss

In order to find the desired 3D shape, we always want that the extracted feature of the sketch is more similar to the same-class 3D shape and more dissimilar to the different-class 3D shape, i. e., maximizing the within-class similarity and minimizing the between-class similarity. However, a query sketch usually has tens or hundreds related 3D shapes with the same class label, and it is difficult to tell which 3D shape is more similar or dissimilar to the query sketch. Note that our aim is to find 3D shapes belonging to the class labels of query sketches rather than find the most similar 3D shapes. Therefore, we focus on extracting the class features rather than the individual features of 3D shapes. The class feature is the mean value of the pre-learned features of the 3D shapes in the

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_5

same class. We use cosine similarity to measure the distance between a sketch and a 3D shape, which is defined as:

$$s = \frac{f_s \cdot f_c}{\|f_s\|_2 \, \|f_c\|_2} \tag{1}$$

where f_s is the sketch feature and f_c is the class feature of the 3D shape.

In a mini-batch with size N, we have N sketches and N corresponding 3D shapes. For each sketch i, we calculate its cosine similarity with all 3D shapes in the mini-batch. We denote the cosine similarity between the sketch and the same-class 3D shape by s_p^i, i. e., the positive pair, and the cosine similarity between the sketch and the rest 3D shapes by $s_n^i = \{s_1, s_2, \ldots, s_{N-1}\}$, i. e., the negative pairs. In order to maximize the similarity score of the positive pair and minimize the similarity score of the negative pair, the similarity loss function is defined as:

$$L = \frac{1}{N} \sum_{i=1}^{N} [\max(s_n^i) - s_p^i + m]_+ \tag{2}$$

where $[]_+$ is a ramp function and m is a margin for a better similarity separation between positive and negative pairs.

The reason why we choose the maximum similarity score from the group of s_n^i to represent the negative pair in Eq. 2 is that it can ensure the scores of all negative pairs are smaller than the positive pair and also increase the difficulty of learning as the same effect of m. Since it is difficult to optimize the Eq. 2, we adopt a smooth approximation by using a LogSumExp function to replace $\max(s_n^i)$ and a softplus function to replace $[]_+$, and then obtain the smooth similarity loss function:

$$L_{smooth} = \frac{1}{N} \sum_{i=1}^{N} \log \left\{ 1 + \exp \left[\log \left(\sum\nolimits_{n=1, n \neq p}^{N} \exp(r s_n^i) \right) - s_p^i + m \right] \right\} \tag{3}$$

where r is a scale factor. By training the student network with L_{smooth}, the sketch feature f_s is gradually close to the pre-learned class feature f_c of the same-class 3D shapes and keeps away from the different-class 3D shapes simultaneously.

4 Experiments

4.1 Datasets

We evaluate our proposed method on a frequently-used benchmark dataset, i. e., SHREC'13 [5], for sketch-based 3D shape retrieval. Some examples of sketches and corresponding 3D shapes in the dataset are shown in Fig. 1. The dataset is built by collecting large-scale hand-drawn sketches from TU-Berlin sketch dataset [22] and 3D shapes from Princeton Shape Benchmark [23], which consists of 90 classes including 7,200 sketches and 1,258 shapes. In each class, there are a total of 80 sketches, and 50 of which are for the training and the rest are for

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_5

the test. The number of 3D shapes varies in different classes. For example, the largest class is 'airplane', which has 184 3D shapes, and there are 12 classes containing only 4 3D shapes.

4.2 Implementation details

Our method is implemented on Pytorch with two NVIDIA GeForce GTX 2080 Ti GPUs.

Network structure The structure is illustrated in Fig. 2. The teacher network adopts the MVCNN [15] architecture and the CNN_1 and CNN_2 use the VGG-11 network [24]. In the student network, CNN_3 utilizes the ResNet-50 network [16].

Prepossessing The prepossessing includes the network pre-training and data processing. The teacher network is pre-trained on ImageNet [25] with 1k categories, and then fine-tuned on all edge images of the 3D shapes. The student network is first pre-trained for the classification task based on a part of QuickDraw dataset [26] with 3.45 million sketches in 345 categories, and then fine-tuned on the training dataset of sketches according to minimize Eq. 3. For the data processing, we uniformly resize the sketch images and the edge images of 3D shapes into a resolution of $224 \times 224 \times 1$.

Parameter settings In the teacher network, the learning rate and batch size are 5×10^{-5} and 8, respectively, and the number of training epochs is set to 20. In the student network, the learning rate and batch size are 1×10^{-4} and 48, respectively, and the number of training epochs is 10. Moreover, the margin m and the scale factor r are set to be 0.15 and 64, respectively. The Adam is employed as an optimizer for both networks and the weight decay is set to 0.

4.3 Experimental results

We show some retrieval results on the SHREC'13 dataset in Fig. 3. The query sketches are listed on the left including the class of chair, bicycle, piano, table, palm tree and sea turtle, and their retrieved top 8 3D shapes are listed on the right according to the ranking of similarity scores. As shown in Fig. 3, our method is effective in retrieving the corresponding 3D shapes of the query sketches. The reasons for generating incorrect results are the limited number of 3D shapes (e. g., the classes of bicycle and sea turtle only contain 7 and 6 3D shapes in the dataset, respectively) and the high similarity score of similar shapes from different classes (e. g., the couch and bench shapes get high similarity scores according to the query sketch of piano).

In order to demonstrate the effectiveness of our proposed method, we compare our method with several state-of-the-art methods, including SBR-VC [5], Siamese [11], Shape2Vec [10], DCML [12], LWBR [9], DCA [8], SEM [7] and DSSH [13]. In addition, we adopt the widely-used evaluation metrics for the sketch-based 3D shape retrieval, including the nearest neighbor (NN), first tier

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_5

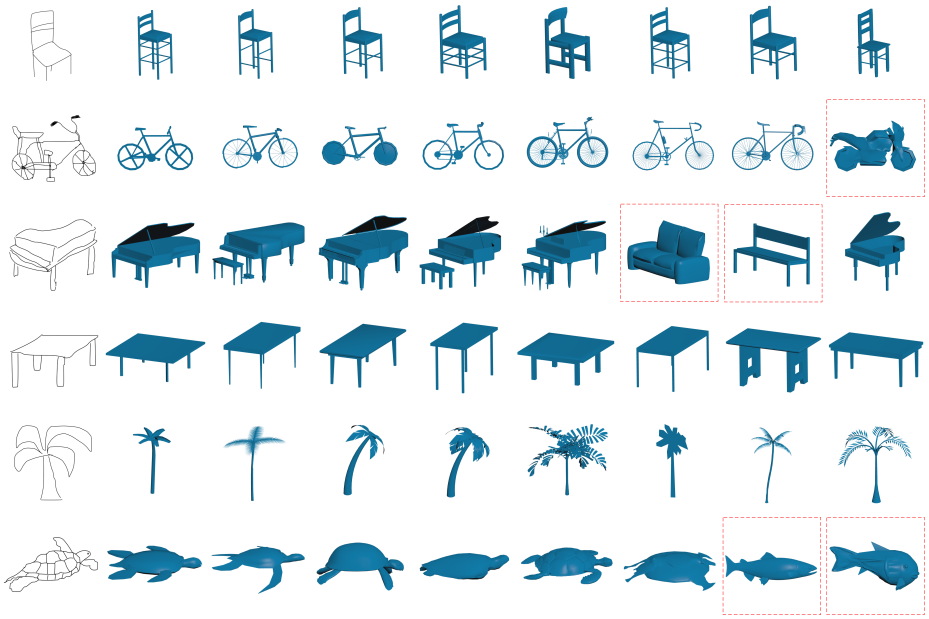

Fig. 3. Some examples of retrieval results. The left column is the query sketches and the right columns are the top 8 retrieved 3D shapes, and the wrong results are highlighted by red dashed squares.

(FT), second tier (ST), E-measure (E), discounted cumulated gain (DCG) and mean average precision (mAP) [4]. Table 1 shows the quantitative comparison of our method with the state-of-the-art methods on the SHREC'13 dataset. Except for the DSSH, it is clear to see that our method achieves the best performance than the state-of-the-art methods for all the evaluation metrics. Compared to the latest method DSSH, our method performs better or equally in the NN, E and DCG metrics.

We also visually compared our method with DSSH to show our advantages. As shown in Fig. 4, for the hand and horse sketch examples, our retrieved 3D shapes are more accurate than DSSH. First, the retrieved 3D shapes with mismatched details have a low-ranking in our method. For example, an unextended hand is ranked last in our method but ranked second in DSSH, and a horse with a lifted leg is ranked fourth in our method but ranked second in DSSH. Second, our wrong results are similar to the right results. For example, the wrong shapes of DSSH are guitars which are extraordinarily different to the horse, whereas our retrieved dogs are similar to the horse. Therefore, compared with DSSH, our method is more suitable for measuring feature distance between sketches and 3D shapes.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_5

Table 1. The comparison of our method and the state-of-the-art methods on the SHREC'13 dataset.

Method	NN	FT	ST	E	DCG	mAP
Siamese [11]	0.405	0.403	0.548	0.287	0.607	0.469
Shape2Vec [10]	0.620	0.628	0.684	0.354	0.741	0.650
DCML [12]	0.650	0.634	0.719	0.348	0.766	0.674
LWBR [9]	0.712	0.725	0.785	0.369	0.814	0.752
DCA [8]	0.783	0.796	0.829	0.376	0.856	0.813
SEM [7]	0.823	0.828	0.860	0.403	0.884	0.843
DSSH [13]	0.831	**0.844**	**0.886**	0.411	0.893	**0.858**
Ours	**0.836**	0.833	0.883	**0.411**	**0.896**	0.853

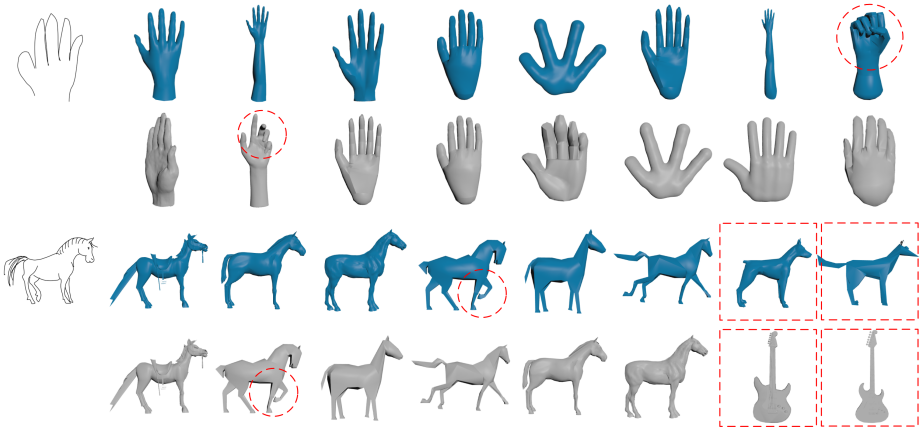

Fig. 4. The comparison of our method and DSSH [13] in two retrieval examples. The blue and gray colors denote the retrieval results of our method and DSSH, respectively, and the wrong results and mismatched details are highlighted by red dashed squares and circles, respectively.

5 Conclusion

In this paper, we propose a novel semantic similarity metric learning method for sketch-based 3D shape retrieval, and use a teacher-student strategy to obtain efficient networks for learning semantic similarity between sketches and 3D shapes. We first adopt the pre-trained classification network as the teacher network to extract the semantic features of 3D shapes, and then train the student network by using the pre-learned features of 3D shapes with a similarity loss function and finally learn the semantic features of sketches. As a result, our method effectively maximizes the within-class similarity and minimizes the between-class similarity. The experiments show that our method performs better than the state-of-the-art methods.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_5

Acknowledgements

This research is supported by the PDE-GIR project which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 778035.

References

1. Chen, D.Y., Tian, X.P., Shen, Y.T. and Ouhyoung, M.: On visual similarity based 3D model retrieval. Computer Graphics Forum **22**(3), 223-232 (2003)
2. Shih, J.L., Lee, C.H. and Wang, J.T.: A new 3D model retrieval approach based on the elevation descriptor. Pattern Recognition **40**(1), 283-295 (2007)
3. Shao, T., Xu, W., Yin, K., Wang, J., Zhou, K. and Guo, B.: Discriminative sketch-based 3d model retrieval via robust shape matching. Computer Graphics Forum **30**(7), 2011-2020 (2011)
4. Li, B., Lu, Y., Godil, A., Schreck, T., Bustos, B., Ferreira, A., Furuya, T., Fonseca, M.J., Johan, H., Matsuda, T. and Ohbuchi, R.: A comparison of methods for sketch-based 3D shape retrieval. Computer Vision and Image Understanding **119**, 57-80 (2014)
5. B. Li, Y. Lu, Afzal Godil, Tobias Schreck, Masaki Aono, Henry Johan, Jose M. Saavedra, S. Tashiro, In: Biasotti, S., Pratikakis, I., Castellani, U., Schreck, T., Godil, A. and Veltkamp R. (eds.) SHREC'13 Track: Large Scale Sketch-Based 3D Shape Retrieval, Eurographics Workshop on 3D Object Retrieval 2013, pp. 89-96 (2013).
6. Li, B., Lu, Y., Li, C., Godil, A., Tobias S., Aono, M., Burtscher, M., Fu, H., Furuya, T., Johan, H., Liu, J., Ohbuchi, R., Tatsuma, A. and Zou, C: Shrec'14 track: Extended large scale sketch-based 3d shape retrieval. In Eurographics workshop on 3D object retrieval 2014, pp. 121–130 (2014)
7. Qi, A., Song, Y. and Xiang, T.: Semantic embedding for sketch-based 3d shape retrieval. In British Machine Vision Conference (2018)
8. Chen, J. and Fang, Y.: Deep cross-modality adaptation via semantics preserving adversarial learning for sketch-based 3d shape retrieval. In Proceedings of the European Conference on Computer Vision, pp. 605-620 (2018).
9. Xie, J., Dai, G., Zhu, F. and Fang, Y.: Learning barycentric representations of 3d shapes for sketch-based 3d shape retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5068-5076 (2017)
10. Tasse, F.P. and Dodgson, N.: Shape2vec: semantic-based descriptors for 3d shapes, sketches and images. ACM Transactions on Graphics (TOG) **35**(6), 1-12 (2016)
11. Wang, F., Kang, L. and Li, Y.: Sketch-based 3d shape retrieval using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1875-1883 (2015)
12. Dai, G., Xie, J., Zhu, F. and Fang, Y.: Deep correlated metric learning for sketch-based 3d shape retrieval. In Proceedings of the AAAI Conference on Artificial Intelligence, **31**(1) (2017)
13. Chen, J., Qin, J., Liu, L., Zhu, F., Shen, F., Xie, J. and Shao, L.: Deep sketch-shape hashing with segmented 3d stochastic viewing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 791-800 (2019)
14. Hinton, G., Vinyals, O. and Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_5

15. Su, H., Maji, S., Kalogerakis, E. and Learned-Miller, E.: Multi-view convolutional neural networks for 3d shape recognition. In Proceedings of the IEEE international conference on computer vision, pp. 945-953 (2015)
16. He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778 (2016)
17. Zhu, F., Xie, J. and Fang, Y.: Learning cross-domain neural networks for sketch-based 3d shape retrieval. In Proceedings of the AAAI conference on artificial intelligence, **30**(1) (2016)
18. Chen, Y., Wang, N. and Zhang, Z.: Darkrank: Accelerating deep metric learning via cross sample similarities transfer. In Proceedings of the AAAI Conference on Artificial Intelligence, **32**(1) (2018)
19. Yu, L., Yazici, V.O., Liu, X., Weijer, J.V.D., Cheng, Y. and Ramisa, A.: Learning metrics from teachers: Compact networks for image embedding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2907-2916 (2019)
20. Dai, W. and Liang, S.: Cross-Modal Guidance Network For Sketch-Based 3d Shape Retrieval. In 2020 IEEE International Conference on Multimedia and Expo, pp. 1-6 (2020)
21. Canny, J.: A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence **8**(6), 679-698 (1986)
22. Eitz, M., Hays, J. and Alexa, M.: How do humans sketch objects?. ACM Transactions on graphics (TOG) **31**(4), 1-10 (2012)
23. Shilane, P., Min, P., Kazhdan, M. and Funkhouser, T.: The princeton shape benchmark. In Proceedings Shape Modeling Applications, pp. 167-178 (2004).
24. Simonyan, K. and Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
25. Krizhevsky, A., Sutskever, I. and Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84-90 (2017)
26. Ha, D. and Eck, D.: A neural representation of sketch drawings. arXiv preprint arXiv:1704.03477 (2017)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_5