Computer Vision and Image Understanding 239 (2024) 103903

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Sketch-based 3D shape retrieval via teacher–student learning

Shuang Liang [*], Weidong Dai, Yiyang Cai, Chi Xie

Tongji University, Shanghai 201804, China

ARTICLE INFO

Communicated by Kwang Moo Yi

Keywords:
Sketch
3D shape retrieval
Cross-modal difference
Teacher–student learning
Feature transfer

ABSTRACT

One of the main difficulties of sketch-based 3D shape retrieval is the significant cross-modal difference between 2D sketches and 3D shapes. Most previous works adopt one-stage methods to directly learn the aligned common embedding space of sketches and shapes by a shared classifier. However, the intra-class difference of the sketch is more significant than the shape, harming the feature learning of 3D shapes when the two modalities are considered under the shared classifier. This issue harms the discrimination of the learned common embedding space. This paper proposes a novel two-stage method to learn a common aligned embedding space via teacher–student learning to address the issue. Specifically, we first employ a classification network to learn the discriminative features of shapes. The learned shape features are considered a teacher to guide the feature learning of sketches. Moreover, we design a guidance loss to achieve the feature transfer with semantic alignment. The proposed method achieves an effective, aligned cross-modal embedding space. Experiments on three public benchmark datasets prove the superiority of the proposed method over state-of-the-art methods.

1. Introduction

With the development and maturity of 3D scanning and 3D modeling technology, it has become easier to obtain and generate 3D shapes. The growing number of 3D shapes leads to the pressing demand for the effective retrieval of 3D shapes. In the early years, keywords were first labeled for 3D shapes, and used as queries for retrieval. This traditional text-based 3D shape retrieval is difficult to describe the details of complex 3D shapes, and completing several text annotations is a challenging task. Therefore, researchers proposed content-based 3D shape retrieval frameworks, which mainly include example-based 3D shape retrieval and sketch-based 3D shape retrieval. Most of the existing works fall into the first group. They take a query 3D model as input and returns similar models (Su et al., 2015; He et al., 2018), for example, the view-based methods (Su et al., 2015; He et al., 2018), rotation-invariant methods and, non-rigid 3D shape retrieval methods. Point cloud networks and deep shape descriptors are proposed for example-based 3D shape retrieval (Qi et al., 2017; Guo et al., 2020). The example-based approaches are straightforward. They suffer from the difficulty to acquire the 3D query shape.

Recently, with the development of touchscreen technology, hand-drawing sketches have become a convenient way of human-machine interaction. Compared with text and 3D shapes, sketches are intuitive and convenient for users to manipulate. Therefore, sketch-based 3D shape retrieval has been receiving increased attention in the community of computer graphics and computer vision.

Despite its flexibility and convenience, sketch-based 3D shape retrieval is a quite challenging task for two reasons. First, sketches and 3D shapes have a large cross-modal difference, making it very difficult to retrieve 3D shapes from sketch queries directly. Second, the hand drawn sketches are highly abstract and subjective, resulting in large intra-class variations in sketches. Some examples from two public datasets illustrate this issues in Fig. 1.

Various methods have been proposed to address the two aforementioned issues. In recent years, deep learning-based methods have become mainstream and advanced the field's progress (Wang et al., 2015; Xie et al., 2017; Dai et al., 2017, 2018; Chen and Fang, 2018; Lei et al., 2019). The key issue of sketch-based 3D shape retrieval is to learn a common semantically aligned feature space. Most of the deep learning-based methods address this issue by adopting one-stage strategy, based on the Siamese network architecture (Wang et al., 2015). Specifically, they use two different Convolution Neural Networks (CNNs) to learn the features of sketches and 3D shapes. Then, the features from different modalities are directly mapped into a common feature space under the constraint of a shared classifier or loss function. However, we notice that the sketch data have relatively larger intra-class difference and relatively smaller inter-class difference than 3D shape data based on the public datasets. Therefore, in one-stage methods, the poor-quality features of sketches will impair the high-quality feature learning of 3D shapes under a shared classifier and loss function. This further impairs the discrimination of the learned common feature space. Thus, these methods are unable to reduce the cross-modal difference effectively because the cross-modal features cannot be aligned compactly. The issue will be discussed in detail in Section 5.2.

* Corresponding author.
E-mail addresses: shuangliang@tongji.edu.cn (S. Liang), 18wddai@tongji.edu.cn (W. Dai), 2131486@tongji.edu.cn (Y. Cai), chixie@tongji.edu.cn (C. Xie).

https://doi.org/10.1016/j.cviu.2023.103903
Received 3 September 2023; Received in revised form 1 November 2023; Accepted 9 December 2023
Available online 14 December 2023
1077-3142/© 2023 Elsevier Inc. All rights reserved.

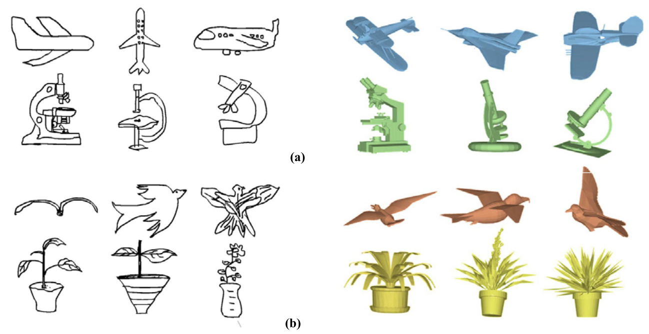

Fig. 1. Samples from two widely used benchmarks: (a) the SHREC 2013 dataset, (b) the SHREC 2014 dataset. Images on the left are sketches. Images on the right are 3D shapes. Samples in the same row belong to the same class.

In some works of sketch-based 3D shape retrieval and sketch-based image retrieval, generative adversarial network (GAN) is used to solve the cross-domain (or cross-modal) gap (Chen and Fang, 2018; Pandey et al., 2019, 2020). These methods avoid the issue of shared classifier and loss. However, they also have limitations. The method based on GAN (Chen and Fang, 2018) is unnecessarily complex. Some methods render the 3D shapes into multiple 2D view images (Su et al., 2015; Dai et al., 2017; Chen and Fang, 2018; Lei et al., 2019), but these 2D view images are different from the natural images. Natural images usually have rich backgrounds, while view images usually have simple backgrounds. Therefore, the GAN methods of sketch-based image retrieval (Pandey et al., 2019, 2020) are not directly applicable to sketch-based 3D shape retrieval.

Recently, the distillation network and its teacher–student strategy (Hinton et al., 2015; Zhou et al., 2018; Chen et al., 2018) have been applied to many research fields, such as model compression and knowledge transfer. The distillation strategy first adopts a complex and powerful teacher model to learn useful and powerful knowledge. Following this, the learned knowledge is seen as a teacher to supervise the learning of a simple and weak student model.

This paper is inspired by the distillation network and its variants. We propose a novel two stage method named Cross-Modal Feature Transfer via Teacher–Student Learning (CFTTSL), for sketch-based 3D shape retrieval. It separates the cross-modal retrieval task into two stages: classification of 3D shapes and feature transfer of 2D sketches. Fig. 2 shows the pipeline of the proposed model. In the first stage, we obtain the features of 3D shapes by training a classification network, which obtains a high-quality feature space of 3D shape without the interference from sketch data. In the second stage, we build a transfer network for the feature learning of sketches. The training of the transfer network is conducted under the guidance of the pre-learned high-quality features of 3D shapes and the supervision of our formulated guidance loss function. In this framework, the classification network of 3D shapes is regarded as a teacher model and the transfer network of sketches is regarded as a student model. The proposed cross-modal guidance learning achieves the cross-modal feature alignment by transferring the sketch features from the original feature space to the pre-learned 3D shape feature space. Consequently, the cross-modal difference between the two modalities is reduced, and the retrieval performance is improved.

This work is an extended version of the conference paper (Dai and Liang, 2020). Compared with the previous version, this work has made a lot of improvements concerning technology, the experiment and the writing:

(1) To supervise the feature transfer of sketches more effectively, the formulated guidance loss function is improved. It is more solid and achieves better performance.

(2) We conduct more experiments in a more comprehensive experimental setup, e.g., new benchmark, new hyper-parameter settings, various 3D shape rendering methods and network architecture to verify the robustness and explore the performance cap of our method.

(3) We further verify the effectiveness and superiority of the proposed method with providing a deep analysis to explain why two-stage method is more effective than one-stage ones.

(4) We provide more detailed method descriptions and convincing visualization. The whole text is improved to be more concise, accurate, and up to date for new experiment observations, changed details, refined notations, and new related works.

To summarize, the main contribution of this work is that we adopt distillation network, *i.e.* teacher–student strategy, to learn a common cross-modal feature space for sketch-based 3D shape retrieval. To our best knowledge, this work is the first to incorporate distillation network into sketch-based 3D shape retrieval. Unlike most existing deep methods, the proposed method is a two-stage framework, which takes advantages of the pre-learned feature space of 3D shapes to guide the feature learning of sketches. Compared with existing one-stage method, the proposed method can reduce cross-modal differences more efficiently. Moreover, we formulate a guidance loss function to supervise the feature transfer of sketches, which effectively achieves cross-modal feature alignment. Experiments on three large public benchmark datasets demonstrate that our method significantly outperforms state-of-the-art methods.

2. Related work

This section mainly introduces some related works of sketch-based image retrieval sketch-based 3D shape retrieval, and distillation networks.

2.1. Sketch-based image retrieval

The task of sketch-based image retrieval is similar to sketch-based 3D shape retrieval. The ideas in these two fields are often interlinked. Therefore, we introduce some related works of sketch-based image retrieval in this paper. Xu et al. (2018) propose a cross-paced partial curriculum learning (CPPCL) framework, which can effectively integrate the self-pacing philosophy with modality-specific partial curricula and investigate different self-paced regularizers. Pang et al. (2020) provide a novel mixed-modality jigsaw puzzle solver as an effective pre-training strategy for fine-grained sketch-based image retrieval. Sain et al. (2021) propose a novel style-agnostic model. Many GAN methods are proposed to solve the cross-domain gap between sketches and images. Pandey et al. (2019) propose a generative model based on joint-adversarial learning. They propose a Maximum Mean Discrepancy (MMD) loss to quantify the distance between the means of the two different class distributions. Pandey et al. (2020) propose a multi-stage GAN based generative model, which can transform the zero-shot sketch-based image retrieval (ZS-SBIR) problem into a conventional image-to-image retrieval problem. Bhunia et al. (2022) design a stroke subset selector to detect noisy strokes, which is helpful to the sketch-based imagre retrieval. Sain et al. (2023) introduces CLIP into the field the of sketch tasks, which greatly improves the performance of zero-shot retrieval.

2.2. Sketch-based 3D shape retrieval

The difference between 3d shapes and sketches is much greater than that between sketches and images. Therefore, sketch-based 3D shape retrieval is a much challenging task compared to sketch-based image retrieval.

Traditional methods. Early works propose various traditional methods based on the hand-draft features and the shallow learning-based features for sketch-based 3D shape retrieval (Eitz et al., 2012; Furuya and Ohbuchi, 2013; Li et al., 2013, 2014b,a). On one hand, various hand-crafted features are employed, such as Zernike moments,

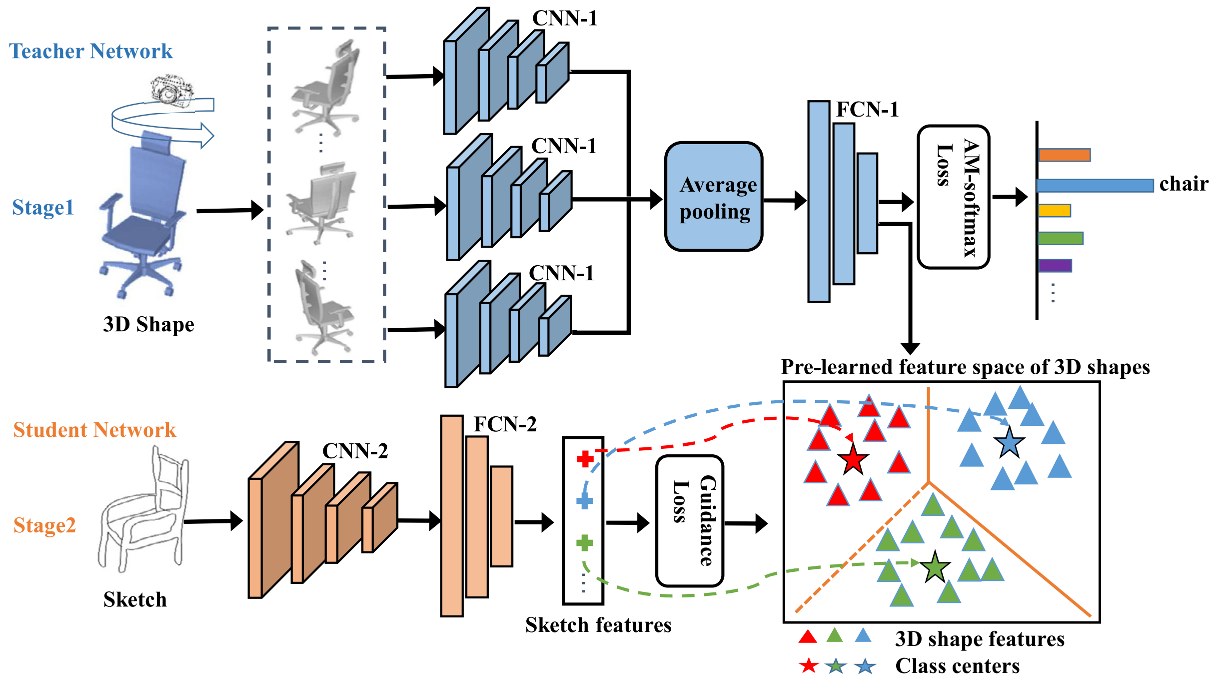

Fig. 2. The framework of the proposed method. It consists of a teacher/classification network, a student/transfer network and a pre-learned feature space of 3D shapes. The training process consists of two stages: (1) The pre-learned feature space of 3D shapes is learned by the teacher network. (2) The sketch features are learned by the student network under the guidance of pre-learned feature space.

coutour-based Fourier descriptors, eccentricity feature and circularity features (Li et al., 2014b), HoG-SIFT features (Yoon and Yoon, 2017). On the other hand, many shallow learning-based features are developed. Eitz et al. (2012) propose a Gabor local line-based feature with a bag-of-features framework for sketch-based 3D shape retrieval. Furuya and Ohbuchi (2013) propose the manifold ranking method to calculate the distance between sketches and the 2D projections of 3D shapes based on local feature extraction. Sahillioğlu and Sezgin (2017) propose a simple and efficient retrieval algorithm, which achieves articulated 3D shape retrieval from sketches for the first time.

Deep learning-based methods. In recent years, deep learning methods have been explored for sketch-based 3D shape retrieval. Wang et al. (2015) first use Siamese network architecture to address the cross-modal retrieval problem, which learns the discriminative features of sketches and 3D shapes by minimizing within-modality and cross-modality loss. Dai et al. (2017) propose the deep correlated metric learning method. This method combines discriminative loss and correlation loss to train two deep neural networks, reducing intra-class distances while increasing inter-class distances. Xie et al. (2017) additionally employ Wasserstein barycenters to aggregate multi-view deep features of rendered images from 3D shapes. He et al. (2018) propose the Triplet-Center loss (TCL), which combined the triplet loss and center loss. The feature learned by TCL is more discriminative than the traditional classification loss. Qi et al. (2021) the problem of fine-grained sketch-based 3D shape retrieval for the first time. Lei et al. (2019) propose a Deep Point-to-Subspace method with an improved center loss function, which projects a sketch into a feature vector and a 3D shape into a subspace spanned by a few selected basis feature vectors. However, most of these methods directly map the cross-modal features into a common feature space, failing to reduce the cross-modal difference effectively. Chen and Fang (2018) propose a feature transformation method, which reduce cross-modal divergence via semantics preserving adversarial learning (GAN-based method). However, this method is complicated and unnecessary. It also fails to effectively utilize the semantic information due to using the metric loss function, which makes the semantic alignment unsatisfactory. Liang et al. (2021) investigates the problem of noisy sketch data for 3D shape retrieval task and propose noise resistant sketch feature learning with uncertainty to reduce the negative impact from noisy data. Cai et al. (2023) improve the method in Liang et al. (2021) by designing an end-to-end uncertainty learning framework. Bai and Bai (2023) introduces a novel hierarchical domain-augmented adaptive learning method to exploit mutual information of the samples from the same class but different domains. Zhao et al. (2022) separate the feature extraction procedure to distinguish modality-general and modality-specific information so as to reduce the difference between sketch and shape descriptors. Bai et al. (2023) introduces a master-auxiliary learning framework to learn a better feature alignment.

Most of these works attempt to reduce the modality gap between sketches and 3D shapes directly. Thus, they improve the feature alignment ways within the traditional one-stage framework. However, this work focuses on the difference between 3D shapes and sketches. Inspired by knowledge distillation, we use discriminative 3D shape features to guide the hard sketch feature learning to learn the sketch embedding space with large inter-class variation and small intra-class variation.

Among these methods, DPSML (Lei et al., 2019) adopts a similar method compared to our method. They also formulate the task of sketch-based 3D shape retrieval as a classification problem and attempt to improve feature alignment through 3D shape branch. Specifically, they define a novel similarity measure method between the sketch and the 3D shape, which is the distance between the sketch feature and its closest point in the spanned subspace formed by a set of view features of 3D shapes. Their proposed similarity measure can reflect the closest viewpoint information captured by the query sketch. However, their method is still based on traditional one-stage framework. Although their method allows the sketch to match its closest view, it cannot address the issue caused by the complexity and abstraction of the sketch modality itself. This work proposes a novel two-stage framework based on knowledge distillation. Through using the relatively cleaner and simpler 3D model modality to guide sketch representation learning, our approach addresses the issues caused by the complexity and abstraction of the sketch and achieves better performance.

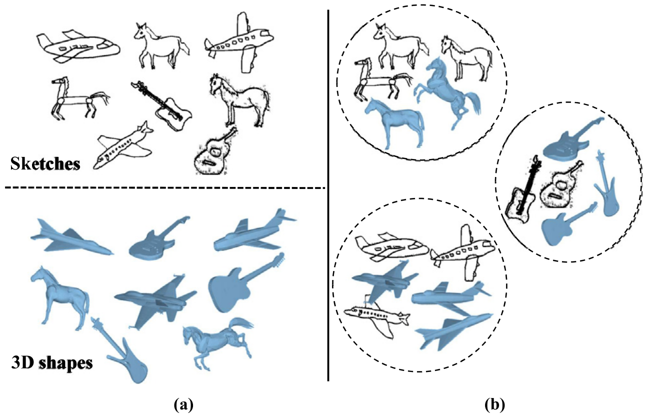

Fig. 3. An illustrated example of the key goal of sketch-based 3D shape retrieval. (a) The features of sketches and 3D shapes in their original feature spaces. (b) The desired common feature space where the cross-modal features are semantically aligned.

2.3. Distillation networks

Knowledge distillation with neural networks is pioneered by Hinton et al. (2015) and Cristian et al. (2006), a transfer learning method aiming to improve the training of a student network by relying on knowledge learned from a powerful teacher network. Zagoruyko and Komodakis (2016) employ the distillation method to deliver the attention information of the teacher network to the student network. An improved knowledge distillation method for metric learning is proposed by Chen et al. (2018). Zhou et al. (2018) propose to use the booster net to supervise the learning of the light network. In this paper, we incorporate the distillation method into sketch-based 3D shape retrieval to reduce the cross-modal difference effectively.

3. Proposed method

In this section, we will describe the motivation, the network architecture and the formulated guidance loss function in detail.

3.1. Motivation

The key goal of sketch-based 3D shape retrieval is to learn a common feature space where the cross-modal features are semantically aligned. Simultaneously, we need to make the intra-class difference small and inter-class difference large. Fig. 3 illustrates the key issue. Fig. 3(a) shows the original feature spaces of the sketches and 3D shapes, they are separated. Fig. 3(b) is the target, a common feature space for sketches and 3D shapes. Based on this goal, many previous works employ heterogeneous Siamese network architecture to learn the common feature space, failing to reduce the cross-modal difference and intra-class difference effectively. However, the methods of image classification are mature due to the development of deep learning. This first inspires us to use a classification network to learn a discriminative feature space for a single modality. Following this, the knowledge distillation inspires us to regard the results of the classification network as a teacher, guiding the feature transfer of another modality to the pre-learned feature space, to reduce the cross-modal difference between the two modalities. Moreover, compared with sketches with limited information and a high abstraction, 3D shapes have rich and geometrically realistic information for the classification task. Hence, it is more reasonable to choose the 3D shape modality for the classification network. The ablation study in Section 5 verifies this conclusion.

3.2. Network architecture

This paper proposes a cross-modal feature transfer method via teacher–student learning (CFTTSL) for sketch-based 3D shape retrieval, which uses the classification results of 3D shapes to guide the feature learning of sketches. Fig. 2 shows the network architecture. The framework consists of three parts: the teacher network, the student network and the pre-learned feature space of 3D shapes. The teacher network is the classification network of 3D shapes. The student network is the feature transfer network of sketches. Finally, the pre-learned feature space is the outputs of the teacher network based on the training data of 3D shapes. The training of the framework contains two stages, and the testing (retrieval) process is one-stage.

3.2.1. Two-stage training

Stage 1. We first train the teacher network to learn the discriminative features of 3D shapes. Specifically, following the previous methods (Su et al., 2015), the widely used multi-view representation for 3D shapes is adopted. That is, a 3D shape is rendered to $\mathbf{V}$ views by placing $\mathbf{V}$ virtual cameras around the 3D shape evenly ($\mathbf{V} = 12$ in this paper). Subsequently, the multiple views passed through CNN-1 separately. All branches of the CNN-1 share the same parameters. An average pooling layer is used to fuse the features that output from CNN-1. Following this, the output of the fused feature from the average pooling layer is fed into the metric network (FCN-1) to reduce the dimension of extracted features. Finally, a classification loss function is used to optimize the whole classification network.

The pre-learned feature space of the 3D shapes is expected to be quite discriminative. In other words, the small intra-class difference between the same class and the large inter-class difference between different classes is expected to be discriminative. The softmax loss is typically good at separating different classes but poor at compacting the same class features. Therefore, in this paper, we adopt the AM-softmax strategy as the loss function of the teacher (classification) network. AM-softmax is proposed for face recognition, which can enlarge the inter-class cosine distance and reduce the intra-class cosine distance. Specifically, we suppose that N is the number of classes. We denote the weights of the classifier layer as $\mathbf{W}^v = [\mathbf{w}^v_1, \mathbf{w}^v_2, \ldots, \mathbf{w}^v_N] \in \mathbb{R}^{l\times N}$. l is the dimension of the weights. v presents that it belongs to the modality of the 3D shape (We use s to represent the sketch modality).

The loss function of teacher (classification) network is

$$\mathbf{L}^v_{class} = \frac{1}{M}\sum_{i=1}^{M} \log \frac{e^{t(\bar{\mathbf{w}}^v_{y_i}\cdot\bar{\mathbf{f}}^v_i - m)}}{e^{t(\bar{\mathbf{w}}^v_{y_i}\cdot\bar{\mathbf{f}}^v_i - m)} + \sum_{j=1,j\neq y_i}^{c} e^{t(\bar{\mathbf{w}}^v_j\cdot\bar{\mathbf{f}}^v_i)}} \tag{1}$$

where the $\bar{\mathbf{w}}^v_{y_i} = \frac{\mathbf{w}^v_{y_i}}{\|\mathbf{w}^v_{y_i}\|_2}$, $\bar{\mathbf{f}}^v_i = \frac{\mathbf{f}^v_i}{\|\mathbf{f}^v_i\|_2}$, which are the normalization of $\mathbf{w}^v_{y_i}$ and $\mathbf{f}^v_i$. The M is the size of a mini-batch, the y_i denotes the label of ith sample, the $\mathbf{f}^v$ is the feature vectors of 3D shapes output from FCN-1 ($\mathbf{f}^v_i$ denotes the ith sample in the mini-batch). The t is the scale factor of the softmax function for regulating the loss convergence. The m is a cosine margin. The optimal values of them are: $t = 15$ and $m = 0.3$ in the task. The details of the two hyper-parameters will be discussed in Section 5.5.

After training the teacher network, the discriminative feature space of 3D shapes is obtained. The normalized weights of classifier $\bar{\mathbf{w}}^v_{y_i} = \frac{\mathbf{w}^v_{y_i}}{\|\mathbf{w}^v_{y_i}\|_2}$ can be treated as the typical center of class i. And $\bar{\mathbf{W}}^v = [\bar{\mathbf{w}}^v_1, \bar{\mathbf{w}}^v_2, \ldots, \bar{\mathbf{w}}^v_N] \in \mathbb{R}^{l\times N}$ is the set of centers of 3D shapes. The pre-learned feature space and class centers of 3D shape are the guidance of the training of the student network in the next stage.

Stage 2. In the second stage, we use another CNN (CNN-2) to learn the deep feature of the sketches. The goal is to transfer the sketch features from original feature space to the pre-learned feature space of 3D shapes, and achieve cross-modal semantic alignment. In the first stage, we obtained the pre-learned class centers of shapes, $\bar{\mathbf{W}}^v = [\bar{\mathbf{w}}^v_1, \bar{\mathbf{w}}^v_2, \ldots, \bar{\mathbf{w}}^v_N] \in \mathbb{R}^{l\times N}$. In this stage, we train the student network

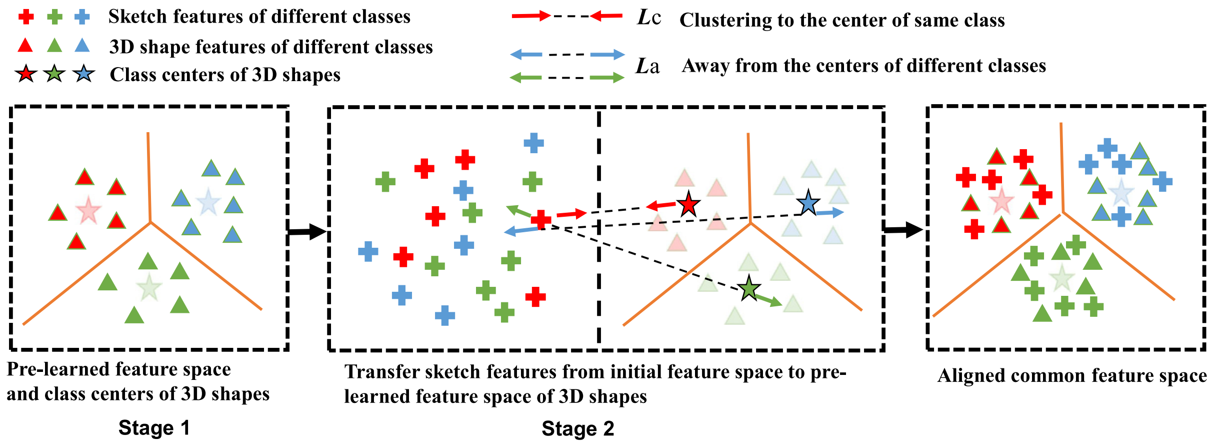

Fig. 4. A toy illustration of the sketch feature transfer under the guidance of pre-learned feature space of 3D shapes and the constraints of the guidance loss function. The different colors represent different classes.

(transfer network of sketches) under the guidance of the pre-learned class centers of 3D shapes, which transfer the feature of sketches to the pre-learned feature space under the supervision of our formulated guidance loss function. The loss function will be discussed in the next subsection with Section 3.3. Fig. 4 presents the process of feature transfer.

3.2.2. One-stage testing

In the testing (retrieval) phase, the features of query sketches and galley 3D shapes are extracted after CNN2-FCN2 and CNN1-FCN1 respectively. Following this, the similarity between the query sketches and the galley 3D shapes is calculated and ranked to obtain the rank list. Notably, the cosine distance is required for the similarity. Specifically, the feature vector of a sketch is denoted as $\mathbf{f}^s \in \mathbb{R}^l$, and the feature vector of a 3D shape is denoted as $\mathbf{f}^v \in \mathbb{R}^l$. l is the dimension of the feature vectors. Therefore, the similarity score between a sketch and a 3D shape could be expressed by

$$\mathrm{sim}(\mathbf{s}, \mathbf{v}) = \frac{\mathbf{f}^s \cdot \mathbf{f}^v}{\|\mathbf{f}^s\|_2 \|\mathbf{f}^v\|_2}. \tag{2}$$

3.3. Guidance loss function

To effectively constrain the feature transfer of sketches, we formulate a guidance loss to optimize the student network. The guidance loss consists of a classification loss and a transfer loss, defined as

$$\mathbf{L}_G = \mathbf{L}^s_{class} + \alpha \mathbf{L}_{trans} \tag{3}$$

where $\mathbf{L}_G$ is our guidance loss, $\mathbf{L}^s_{class}$ is a classification loss, and $\mathbf{L}_{trans}$ is a transfer loss. α is a trade-off hyper-parameter, and it is further analyzed in Section 5.5.

Transfer loss. On the one hand, we need to learn a common feature space for two modalities by transferring the features of sketches to the pre-learned feature space of 3D shapes. On the other hand, we also need to ensure that the sketch features are aligned with the 3D shape features with the same semantic information, that is, cross-modal feature alignment. Based on this goal, the transfer loss function is formulated as:

$$\mathbf{L}_{trans} = \mathbf{L}_c - \beta \mathbf{L}_a \tag{4}$$

where the $\mathbf{L}_c$ is the cosine distances between the features of sketches and the pre-learned class center of 3D shapes in the same class. Meanwhile, the $\mathbf{L}_a$ represents the mean of cosine distance between the features of sketches and other class centers of 3D shapes. β is a hyper-parameter, which controls the trade-off between the two terms. In the experiments, the optimal values of β is: $\beta = 0.1$. The details of this hyper-parameter will be discussed in Section 5.5.

Specifically, the $\mathbf{L}_c$ is formulated as

$$\mathbf{L_c} = \frac{1}{M} \sum_{i=1}^{M} \frac{\mathbf{f}^s_i \cdot \mathbf{w}^v_{y_i}}{\|\mathbf{f}^s_i\|_2 \|\mathbf{w}^v_{y_i}\|_2} \tag{5}$$

The $\mathbf{L}_a$ is defined as

$$\mathbf{L_a} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1, j \neq y^i}^{N} \frac{\mathbf{f}^s_i \cdot \mathbf{w}^v_j}{\|\mathbf{f}^s_i\|_2 \|\mathbf{w}^v_j\|_2} \tag{6}$$

In Eqs. (5) and (6), the M is the size of a mini-batch, N is the number of classes. $\mathbf{f}^s_i$ denotes the feature vector of ith sketch sample, and y^i denotes the label of the ith sample. Further, $\{\mathbf{w}^v_1, \mathbf{w}^v_2, \dots, \mathbf{w}^v_N\}$ are the pre-learned class centers of 3D shapes.

A toy illustration of $\mathbf{L}_c$ and $\mathbf{L}_a$ is shown in Fig. 4, The $\mathbf{L}_c$ loss function aims to cluster the features of sketches toward the class center of 3D shapes in the same class and meanwhile $\mathbf{L}_c$ makes the features away from the class centers of different classes. In short, it constrains the learning of sketches to transfer the sketch features to the pre-learned feature space of 3D shapes, achieving the semantic alignment of cross-modal features. Consequently, the cross-modal difference between 2D sketches and 3D shapes are reduced.

Joint supervision with classification loss. Our formulated transfer loss can be used independently from the classification loss. It can also be combined with the classification loss to achieve more discriminative embeddings and better retrieval performance. As shown in Eq. (3), we combine the transfer loss with AM-softmax loss. Notably, the AM-softmax $\mathbf{L}^s_{class}$ in student network is the same as the Eq. (1).

4. Experiment setups

This section introduces three widely used benchmark datasets for sketch-based 3D shape retrieval, the evaluation metrics and the implementation details.

4.1. Datasets

We perform experiments on three wildly used benchmark datasets, SHREC 2013 (Li et al., 2013), SHREC 2014 (Li et al., 2014b) and SHREC 2016 (Li et al., 2016).

SHREC 2013. SHREC 2013 is a large scale benchmark that evaluate algorithms for sketch-based 3D shape retrieval. The benchmark is constructed by collecting shared classes from both the Princeton Shape Benchmark (Shilane et al., 2004) and human-drawn sketches. The dataset has 7200 hand-drawn sketches and 1258 3D shapes. These sketches are divided into two subsets for each class: 50 samples for training and 30 samples for testing. For the 3D shapes, the numbers of

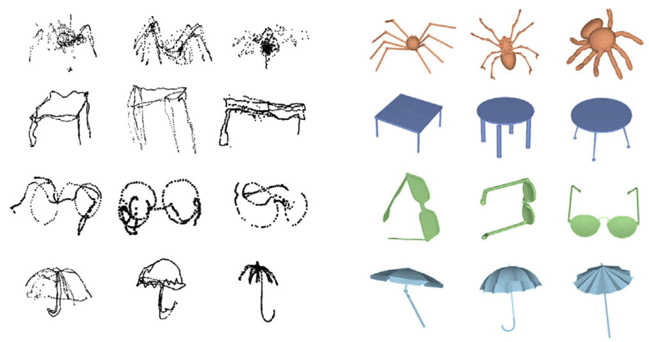

Fig. 5. Samples from SHREC the 2016 dataset. Images on the left are 3D sketches. Images on the right are 3D shapes. Samples in the same row belong to the same class.

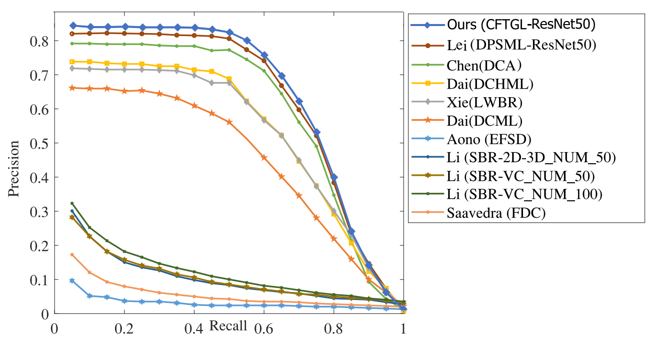

Fig. 6. Comparison of precision–recall curves on SHREC 2013.

distinct classes are different (14 on average). Fig. 1(a) shows the sketch examples and corresponding shapes in this dataset.

SHREC 2014. SHREC 2014 is a larger scale benchmark than SHREC 2013. SHREC 2014 has 171 classes, which contains 8987 3D shapes and 13,680 sketches. Each class of 3D shapes collection has an average of 53 samples. Each class of sketches has 80 images, including 50 training images and 30 test images. The SHREC 2014 dataset is more challenging since it has more categories and larger intra-class variations. Fig. 1(b) shows the sketch examples and corresponding shapes in SHREC 2014.

SHREC 2016. SHREC 2016 is a new benchmark and different from the SHREC 2013 and SHREC 2014 datasets. SHREC 2016 uses 3D sketches as the query sketches for 3D shape retrieval. The 3D sketches are collected by the Microsoft Kinect device. There are 300 3D sketches in the SHREC 2016 dataset, which is collected by Li et al. (2015a,b). These 3D sketches are divided into 30 classes. For each class, 7 sketches are used for training, 3 sketches are used for testing. The 3D shapes of SHREC 2016 come from SHREC 2013, which has 1258 3D shapes and 90 classes in total. Among the 30 classes of sketches, only 21 classes have the corresponding 3D shapes. The remaining 9 classes have no corresponding 3D shapes. In the model training phase, only 147 sketches from the 21 classes are used. The remaining 63 sketches are used as the test set. Fig. 5 shows the 3D sketch examples and corresponding shapes in SHREC 2016.

4.2. Evaluation metrics

In this work, we adopt the most widely used metrics for sketch-based 3D shape retrieval as follows: nearest neighbour (NN), first-tier (FT), second-tier (ST), E-measure (E), discounted cumulated gain (DCG) and the mean average precision (mAP) (Li et al., 2014a). We also report the precision–recall curve to demonstrate the retrieval performance visually.

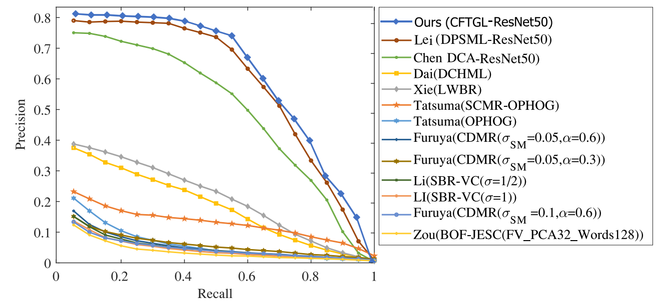

Fig. 7. Comparison of precision–recall curves on SHREC 2014.

4.3. Implementation details

This section introduces implementation details about data preprocessing, network backbones and parameter settings.

Data preprocessing. We resize the sketch images and the rendered views of 3D shapes to 224 × 224. For data augmentation, we use a random horizontal flip and random rotation with degree ranging from [−15°, 15°], for both sketches and views of 3D shapes. For the fair comparison with previous methods, we use the front view image of the 3D sketch as the query input on the SHREC 2016 dataset.

Backbone networks. We adopt five backbones for our framework, comprehensively evaluate our method, and fairly compare it with previous works. The backbones include AlexNet (Krizhevsky et al., 2012), VGG-16 (Simonyan and Zisserman, 2014), VGG-19 (Simonyan and Zisserman, 2014), ResNet-50 (He et al., 2016) and the latest vision transformer (ViT-B/16) (Dosovitskiy et al., 2020). The same backbones are used for the sketch network and view image network. All the backbones are pre-trained on ImageNet (Krizhevsky et al., 2012). Specifically, for the CNNs (CNN-1 and CNN-2), we adopt the layers of AlexNet before the "fc7" layer(inclusive), the layers of VGG1-16 before the "fc7" layer (inclusive) and the layers of ResNet-50 before the "pooling5" layer (inclusive). The FCNs (FCN-1 and FCN-2) consisted of fully connected layers. (4096-1024-256 for Alexnet, VGG-16 and VGG-19, 2048-1024-256 for ResNet-50, and 768-512-256 for ViT-B.)

Parameter setting. We set the initial learning rate as 0.001 and drop it by 0.9 every 10 epochs. The total number of training epochs is set as 90. We adopt the SGD as the optimizer, and its momentum and weight decay are set to 0.9 and 1e−4, respectively.

5. Results and analysis

5.1. Comparison with other methods

On the three benchmark datasets, we compare the proposed CFTTSL method with state-of-the-art methods, including CDMR (Furuya and Ohbuchi, 2013), SBR-VC (Li et al., 2013), DCML (Dai et al., 2017), Siamese (Wang et al., 2015), LWBR (Xie et al., 2017), DCHML (Dai et al., 2018), TCL (He et al., 2018), DCA (Chen and Fang, 2018), DPSML (Lei et al., 2019) and CGN (Dai and Liang, 2020), VTS (Dubey et al., 2022). Notably, CGN is the method in the conference vision (Dai and Liang, 2020). To evaluate the performance on vision transformer (Dosovitskiy et al., 2020), the state-of-the-art backbone, we introduce ViT-based image retrieval method, VTS (Dubey et al., 2022) into sketch-based shape retrieval and implement DPSML (Lei et al., 2019) and CGN (Dai and Liang, 2020) in ViT by ourselves for comparison.

Retrieval Performance on SHREC 2013. We draw the Precision-Recall (PR) curves to compare the proposed method and previous approaches to demonstrate the retrieval performance. Fig. 6 illustrates the comparison of PR curves on the SHREC 2013 dataset. From these curves, it can be observed that the proposed CFTTSL method significantly beat the compared models. Moreover, Table 1 reports the six

Table 1
The performance (%) on SHREC 2013 dataset.

Method	Backbone	NN	FT	ST	EE	DCG	mAP
CMDR	–	27.9	20.3	29.6	16.6	45.8	25.0
SBR-VC	–	16.4	9.7	14.9	8.5	34.8	11.6
Siamese	–	40.5	40.3	54.8	28.7	60.7	46.9
DCML	AlexNet	65.0	63.4	71.9	34.8	76.6	67.4
LWBR	AlexNet	71.2	72.5	78.5	36.9	81.4	75.2
DCHML	AlexNet	73.0	71.5	77.3	36.8	81.6	74.4
TCL	Alexnet	76.3	78.7	84.9	39.2	85.4	80.7
DCA	Resnet50	78.3	79.6	82.9	37.6	85.6	81.3
VTS	Vit-B	84.7	86.2	90.4	42.3	90.5	87.9
DPSML	AlexNet	74.1	76.1	82.1	38.5	83.6	78.5
	VGG19	80.1	81.6	85.2	39.8	87.0	83.1
	ResNet50	81.9	83.4	87.5	41.5	89.2	85.7
	ViT-B	85.4	87.1	90.7	42.4	91.0	88.3
CGN	AlexNet	77.0	80.1	84.7	39.9	86.1	81.7
	VGG19	81.9	83.0	87.5	41.8	88.6	84.9
	ResNet50	83.2	85.3	90.2	41.9	90.1	87.0
	ViT-B	86.3	87.8	91.5	42.7	91.8	89.1
Ours (CFTTSL)	AlexNet	**77.9**	**81.6**	**85.3**	**40.4**	**87.1**	**82.8**
	VGG19	**82.8**	**84.1**	**87.8**	**41.9**	**88.9**	**85.9**
	ResNet50	**83.9**	**86.0**	**90.6**	**42.0**	**90.7**	**87.8**
	ViT-B	**86.9**	**88.4**	**91.8**	**42.7**	**92.1**	**89.7**

metrics (NN, FT, ST, E, DCG and mAP) of our method and previous works. The table shows that the performance of CFTTSL method is superior to the state-of-the-art methods in all evaluation metrics based on the same CNN backbones. Specifically, our method outperforms the DPSML methods by a gain of 4.3% mAP based on the AlexNet backbone. Our method also outperforms the CGN method of conference vision by a gain of 1.1% mAP based on AlexNet backbone. Even compared with the ResNet-50 results achieved by DPSML (Lei et al., 2019), our VGG-19 results exceeded them in some evaluation metrics. For instance, our 85.9% mAP is higher than its 85.7% mAP. Further, our results based on the ResNet-50 and ViT backbone are better than the state-of-the-art results in all evaluation metrics.

Retrieval Performance on SHREC 2014. SHREC 2014 is a more difficult dataset due to the larger intra-class variations and more categories. On this dataset, we also evaluate the proposed method and compare the results with state-of-the-art results. Fig. 7 shows the PR curves of various methods on SHREC 2014 dataset, demonstrating that the proposed method outperforms the previous methods. Similar to the SHREC 2013 dataset, we also list the comprehensive evaluation metrics of our method and state-of-the-art methods in Table 2. The table demonstrates the superiority of our method to other methods in six evaluation metrics. For instance, Our CFTTSL method with ResNet-50 reaches 83.6% mAP, which is 2.3%, 3.3% and 0.6% higher than DPSML (Lei et al., 2019), DCA (Chen and Fang, 2018) and CGN (Dai and Liang, 2020), respectively. And our CFTTSL method with ViT reaches 84.3% mAP, which exceeds the previous best method by 1%. The performance on the SHREC 2014 dataset shows that our method is effective for difficult large-scale datasets.

Retrieval Performance on SHREC 2016. SHREC 2016 is a new benchmark for 3D shape retrieval. It is different from the SHREC 2013 and SHREC 2014 datasets since it uses the 3D sketches as query sketches. For the fair comparisons with the previous method (Lei et al., 2019), we use the front view image of the 3D sketch as the query input. And the view images selections of 3D shapes are the same as SHREC 2013 and SHREC 2014. Table 3 reports the retrieval performance of the proposed method in SHREC 2016. From this table, it can be observed that our method outperforms the previous works for all metrics. Specifically, the mAP of our method exceeds the DPSML method (Lei et al., 2019) with a gain of 3.1% on the same Alexnet backbone. As for the VGG-19 backbone, the proposed method also achieves state-of-the-art results for all metrics. The retrieval performance of the proposed method on the SHREC 2016 dataset further verifies the effectiveness and robustness of the proposed method.

Table 2
The performance (%) on SHREC 2014 dataset.

Method	Backbone	NN	FT	ST	EE	DCG	mAP
CMDR	–	10.9	5.7	8.9	4.1	32.8	5.4
SBR-VC	–	9.5	5.0	8.1	3.7	31.9	5.0
Siamese	–	23.9	21.2	31.6	14.0	49.6	22.8
DCML	AlexNet	27.2	27.5	34.5	17.1	49.8	28.6
LWBR	AlexNet	40.3	37.8	45.5	23.6	58.1	40.1
DCHML	AlexNet	40.3	32.9	39.4	20.1	54.4	33.6
TCL	Alexnet	58.5	45.5	53.9	27.5	66.6	47.7
VTS	Vit-B	81.6	84.1	87.6	42.7	89.9	85.3
DCA	AlexNet	49.8	46.4	51.3	29.4	62.7	50.2
	VGG16	68.2	69.8	72.3	37.5	78.3	71.1
	ResNet50	77.0	78.9	82.3	39.8	85.9	80.3
DPSML	AlexNet	67.7	73.2	79.5	37.9	83.0	75.1
	VGG19	74.8	78.5	83.9	40.6	86.6	80.0
	ResNet50	77.4	79.8	84.9	41.5	87.7	81.3
	ViT-B	82.3	85.2	88.4	43.5	90.3	86.1
CGN	AlexNet	73.4	74.3	80.8	39.1	84.4	77.8
	VGG16	76.9	78.9	83.1	41.2	86.9	80.8
	VGG19	77.5	79.2	84.4	41.3	87.1	81.5
	ResNet50	78.9	81.1	85.0	41.8	88.1	83.0
	ViT-B	83.3	86.2	89.8	43.6	91.1	87.2
Ours (CFTTSL)	AlexNet	**74.5**	**75.5**	**81.8**	**39.4**	**85.1**	**78.7**
	VGG16	**77.8**	**80.0**	**83.8**	**41.3**	**87.4**	**81.7**
	VGG19	**78.4**	**79.9**	**84.9**	**41.5**	**87.6**	**82.3**
	ResNet50	**79.5**	**82.0**	**85.5**	**41.9**	**88.4**	**83.6**
	ViT-B	**84.3**	**87.0**	**90.4**	**44.2**	**91.7**	**88.1**

Table 3
The performance (%) on SHREC 2016 dataset.

Method	Backbone	NN	FT	ST	EE	DCG	mAP
Siamese	–	0.0	3.1	10.8	4.8	29.3	7.2
CNN-SBR	–	22.2	25.1	32.0	28.6	47.1	31.4
DCHML	AlexNet	11.7	10.6	14.8	8.6	32.7	14.7
DPSML	AlexNet	42.9	47.8	56.3	27.9	60.9	49.9
	VGG19	47.6	51.0	57.2	29.0	64.0	53.3
Ours (CFTTSL)	AlexNet	**46.3**	**51.2**	**57.4**	**28.7**	**63.1**	**53.0**
	VGG19	**49.2**	**53.0**	**58.0**	**29.2**	**66.3**	**55.4**

Comparison under same setting We adopt AM-softmax as the loss of the proposed method, while previous SOTA methods use different loss functions respectively. TCL (He et al., 2018) method formulates the triplet-center loss as loss function. It combines the triplet loss (Florian et al., 2015) and center loss (Wen et al., 2016). DCA (He et al., 2018)

Table 4
Result comparison of our method and previous methods under the same loss functions on the SHREC 2014 dataset.

Method	Backbone	NN	FT	ST	EE	DCG	mAP	Params (M)
TCL	AlexNet	58.5	45.5	53.9	27.5	66.6	47.7	120.9
CFTTSL	AlexNet	**60.4**	**47.2**	**55.4**	**30.9**	**69.1**	**50.4**	123.6
DCA	ResNet-50	77.0	78.9	82.3	39.8	85.9	80.3	58.1
CFTTSL	ResNet-50	**78.8**	**80.5**	**83.2**	**41.5**	**86.6**	**82.1**	52.5
DPSML	ResNet-50	77.4	79.8	84.9	41.5	87.7	81.3	58.0
CFTTSL	ResNet-50	**79.5**	**82.5**	**85.8**	**41.6**	**88.9**	**82.7**	52.5

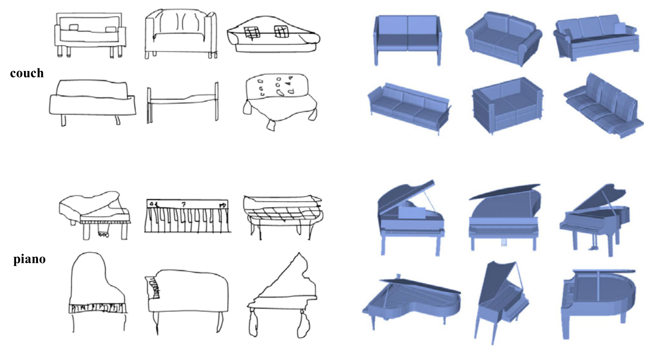

Fig. 8. Examples in SHREC 2014 dataset, showing that sketch data have relatively large intra-class difference than 3D shape data.

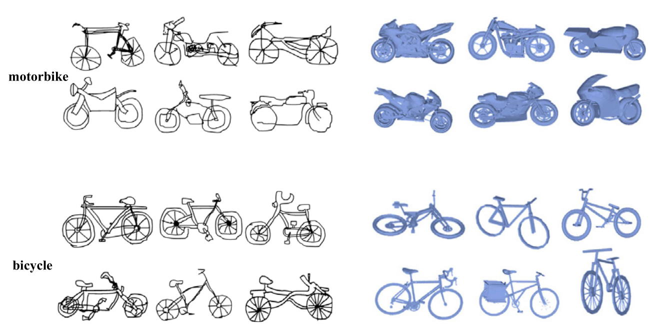

Fig. 9. Examples in SHREC 2014 dataset, showing that sketch data have relatively small inter-class difference than 3D shape data.

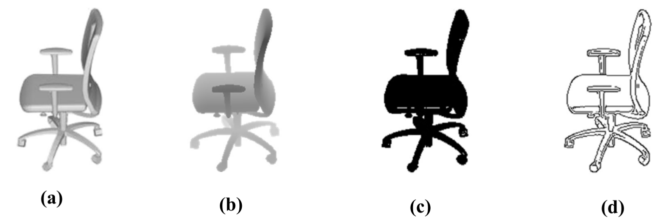

Fig. 10. Illustrations of 2D projections of 3D shapes. (a) view images, (b) depth-buffer images, (c) silhouette images, (d) contours of view images.

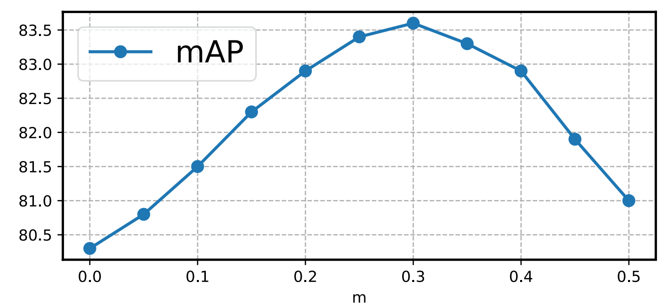

Fig. 11. The retrieval performance achieved by varying m and fixing the other hyper-parameters.

method adopts the variants of triplet loss as main loss function, which learns discriminative features for each modality. DPSML (Lei et al., 2019) method designs a classification-based loss by improving the center loss.

For fair comparison, we implement our two-stage method under the same loss functions of these methods. Since no open-source codes are provided, we reproduce these loss functions on our own, including TCL, DCA and DPSML. The results we reproduced are comparable with the original results in Table 2. Then, we incorporate these loss functions into our two-stage method based on the same backbone on the SHREC 2014 dataset. Specifically, we use their loss functions in stage 1 of the proposed framework to train the teacher network. In stage 2, we use transfer loss (Eq. (4)) to train the student network. Notably, for the metric loss function of DCA, we obtain the pre-learned class centers of 3D shapes by feeding the training data into 3D shape networks. Comparison results are given in Table 4. As shown, the proposed method outperforms previous methods under the same loss function. This further verifies the effectiveness of our approach. Also, it shows that the proposed framework is compatible with other loss functions.

5.2. Ablation study on knowledge distillation-based method

In this subsection, we conduct experiments to compare the proposed two stage feature alignment method based on knowledge distillation with one-stage feature alignment method.

One-stage method implementation. For comparison, based on the previous one-stage method (He et al., 2018; Wang et al., 2015; Lei et al., 2019), the one-stage method using the proposed network architecture and loss function is implemented. Specifically, we make the 3D shape network and the sketch network share fully connected layers and loss functions. That is, CNN1 and CNN2 share a FCN, and the FCN has the same structure as FCN1 and FCN2. The other parameter settings are the same as the settings in the proposed two-stage method. For comprehensive comparisons, we choose softmax loss and Am-softmax loss in the experiment.

Effectiveness of two-stage methods. We compare the results of our one-stage and two-stage methods in Table 5. It can be observed that our one-stage method achieves 82.1% mAP, outperforming the DPSML (Lei et al., 2019) approach by 81.3% mAP. Furthermore, the results of our two-stage method are better than the one-stage method by 1.5% mAP on softmax loss and 2.3% mAP on Am-softmax loss, which validate the effectiveness.

Effectiveness of 3D shape as the teacher. To validate that the 3D shape modality is more suitable for the teacher. We conduct ablation experiments that swap the student network and teacher network, *i.e.*, the network of sketches become the teacher and the network of 3D shapes become the student. The other settings are the same as before. Table 6 shows the comparison of the performance on the SHREC 2014 dataset. The results demonstrate clear superiority of 3D shapes as the teacher. And when sketches becomes the teacher, the two-stage method performs even worse than the one-stage method.

Table 5
Result comparison of our one-stage method and two-stage method on SHREC 2014 dataset.

Method	Loss	NN	FT	ST	EE	DCG	mAP	Params (M)
One-stage	AM-softmax	78.2	80.2	84.4	41.2	87.4	82.1	49.8
Two-stage	AM-softmax	**79.5**	**82.0**	**85.5**	**41.9**	**88.4**	**83.6**	52.5
One-stage	softmax	77.2	77.5	83.2	40.7	86.3	78.8	49.8
Two-stage	softmax	**77.9**	**79.4**	**83.9**	**40.9**	**87.2**	**81.1**	52.5

Table 6
Results of different teacher–student setting on SHREC 2014 dataset.

Tea-Stu	NN	FT	ST	EE	DCG	mAP
3D-sketch	**79.5**	**82.0**	**85.5**	**41.9**	**88.4**	**83.6**
sketch-3D	78.1	80.4	84.3	41.1	87.2	81.8

Table 7
Cosine similarity and mAP on SHREC 2014 dataset. cos (shape) and cos (sketch) represents the average cosine similarity between sketch/shape features and their corresponding class centers.

Method	cos (shape)	cos (sketch)	mAP
One-stage	0.89	0.83	82.1
Sketch-3D	0.88	0.80	81.8
3D-sketch	**0.93**	**0.88**	**83.6**

5.3. Analysis: How does the proposed two-stage method work?

In this section, we provide the qualitative and quantitative analysis about how the two-stage method works and why 3D shapes is more suitable for the teacher.

How the two-stage method works. Based on the observation of public datasets, hand-drawn sketches have strong subjectivity and abstractness, leading to large intra-class difference and small inter-class difference. Compared to sketches, 3D shapes have richer geometric information, making the distribution of 3D shape data more compact and easy to learn. Figs. 8 and 9 show the examples of above-these issues. In Fig. 8, the sketches have larger intra-class variation than 3D shapes from the same class. Fig. 9 shows that the sketches from the 'motorbike' class and the sketches from the 'bicycle' class are hard to distinguish. Meanwhile the 3D shapes from the two classes are easy to distinguish.

(1) *The proposed two-stage method improves the learning of 3D shape*. It makes the learned feature space of 3D shape more compact. While one-stage methods adopt a shared network to train both shapes and sketches, thus the learning process will consider the two modalities simultaneously. Consequently, the poor-quality features with large intra-class differences and small inter-class differences of sketches will impair the learning of high-quality features of 3D shapes. This further impairs the discrimination and compactness of the learned common feature space. In contrast, the two-stage method first learns the high-quality features of 3D shapes independently, avoiding the impairment from the poor-quality features of sketches. As shown in Table 7, 3D shape network is trained independently in '3D-Sketch', and its intra-class cosine similarity of 3D shape is significantly higher than 'One-stage' and 'sketch-3D' in which the training process of 3D shape is impaired by sketches. This result validates that training 3D shapes independently can learn a more compact 3D shape feature space.

(2) *The learned high-quality 3D feature space guides*. This can help improves the learned quality of sketch features. As shown in Table 7, the sketch cosine similarity of the sketches in '3D-sketch' is higher than other methods, which validates the sketch feature space under the guidance of the 3D shape is more compact. The reason is that the 3D shape guidance can help the learning of similar classes. Sketches from similar classes are often confusing and difficult to distinguish. These confusing sketches make large intra-class differences and small inter-class differences. And class centers pre-learned in 3D shapes are usually more discriminative. Guiding sketch features optimizing towards the 3D class centers can help separate sketch features of different classes that are close in the feature space and gather features of the same class, e.g., the average cosine similarity of sketches learned by the one-stage method from class 'bike' and 'motorbike' is 0.39 while the average cosine similarity of the two-stage method is 0.16. Therefore, the two-stage method can learn a more discriminative common feature space, improving the retrieval performance.

Why the 3D shape is more suitable for the teacher. The reason is that hand-drawn sketches are very abstract. Further, the pre-learned features of sketches have large intra-class distance and small inter-class distance, which is unable to effectively guide the feature transfer of 3D shapes to obtain a discriminative cross-modal feature space. The results in Table 7 validate this conclusion. 3D shapes as the teacher form more compact feature space and improve the retrieval performance.

5.4. Model complexity

Table 4 reports the number of model parameters as model complexity of our CFTTSL and previous SOTA methods. It shows that our method is also competitive in terms of efficiency. Table 5 shows that the proposed two-stage method based on knowledge distillation has only a small increase in model complexity compared to one-stage method, but has large performance increase. The results demonstrate the practicality of the proposed method.

5.5. Impact of hyper-parameter

This section reviews experiments conducted to study the impact of the two hyper-parameters in Eq. (1) (t, m) and Eq. 4 (α, β). All the experiments are conducted on the SHREC 2014 dataset, and the backbone is ResNet-50.

Setting t and m. We fix α and β as 0.1 and 0.1 to explore the influence of parameters t and m. Parameter t has already been discussed sufficiently in some face recognition works (Ranjan et al., 2017; Wang et al., 2017). It depends on the number of categories. In this paper, we directly fix it to 15 and will not discuss it in detail. We conduct experiments by varying m from 0 to 0.45 to explore the optimal value of hyper-parameter m in our task. We report the important and comprehensive metric mAP in Fig. 11. The curve shows that the optimal value of m in the task is $m = 0.3$. We note that we keep the hyper-parameter m in $\mathbf{L}^{v}_{class}$ and $\mathbf{L}^{s}_{class}$ the same.

Setting α and β. we fix the t and m to 15 and 0.3 to explore the influence of parameters in Eqs. (3) and (4). According to the analysis in Section 3.3, the transfer loss $\mathbf{L}_{trans}$ can be used to supervise the student network independently. Therefore, we first experiment by using a single transfer loss to explore the optimal value of β. We set the β as 0, 0.01, 0.1 and 1.0 respectively. Table 8 shows the experimental results. From the table, we can see that the best results come from the setting $\beta = 0.1$. Moreover, when the $\beta = 0.0$ (only the $\mathbf{L}_{c}$), our method can still perform well. It achieves 81.9% mAP and 77.6% NN, outperforming the DPSML method (Lei et al., 2019).

Our guidance loss consists of a transfer loss and a classification loss. we fix the β as 0.1 to explore the influence of hyper-parameter α. Then, we set α as 0, 0.01, 0.1 and 1.0, respectively. The experimental results are reported in Table 9. As shown, the proposed method achieves the best performance when $\alpha = 0.1$. We can see that our guidance loss with AM-softmax obtains better retrieval results than single transfer loss. Notably, when $\alpha = 0.0$ (guidance loss without transfer loss), the model

Table 8
Impact of hyper-parameter β in the guidance loss function.

β	NN	FT	ST	EE	DCG	mAP
0.0	77.6	79.6	84.6	41.5	87.5	81.9
0.01	78.3	80.4	85.0	41.7	87.8	82.4
0.1	**79.0**	**81.4**	**85.3**	**41.8**	**88.2**	**83.2**
1.0	77.5	79.3	84.5	41.4	87.3	81.7

Table 9
Impact of hyper-parameter α in the guidance loss function.

α	NN	FT	ST	EE	DCG	mAP
0.0	3.1	5.1	5.6	2.0	6.7	7.6
0.01	68.4	73.8	79.9	37.7	83.2	75.6
0.1	**79.5**	**82.0**	**85.5**	**41.9**	**88.4**	**83.6**
1.0	78.5	79.9	84.8	41.7	87.5	82.6

has very poor results because it fails to transfer the sketch feature to the feature space of 3D shapes. In summary, our method achieves the best results when $\alpha = 0.1$ and $\beta = 0.1$. Therefore, our guidance loss function is

$$\mathbf{L}_G = \mathbf{L}^s_{class} + 0.1 * \mathbf{L}_c - 0.01 * \mathbf{L}_a \tag{7}$$

5.6. Impact of feature mapping on different layers

In Fig. 2, we only map the output features of the two FCN models in the teacher and student networks. To investigate the results when feature mapping also occurs at the different layers of the two FCNs, we conduct some experiments to explore this issue. Specifically, in the training stage, we calculate the class centers of different layers in FCN-1. Then we use the transfer loss (Eq. (4)) to transfer the sketch features of different layers in FCN-2 into the corresponding layer in FCN-1. In the test stage, we only use the features of the output layer. Feature mapping occurs at different layers of the two FCNs are conducted, including at output layer (layer-3), at layer-1 and 3, at layer-2 and 3, and at layer-1, 2, and 3. The results are reported in Table 10. This table shows that when feature mapping occurs at some layers except the last layer, the retrieval performance will degrade. The reason is that the feature transfer of other layers will impair the feature learning of the output layer used for the test.

5.7. Comparison with different 2D projections and sketch styles

This subsection explores how it performs when we capture different 2D projections for 3D shapes, including different rendering styles and viewpoints. We also compare with the performance of different styles of hand-drawn sketches.

Comparison with different rendering styles of 3D shape. The rendering styles include depth-buffer (z-buffer) images, silhouette images and contours of view images, as shown in Fig. 10. Specifically, We replace the view images in the proposed framework with different 2D projections and the combination of them. Notably, this combination consists of different types of projections in equal proportion. *i.e.* 3 view images, 3 depth-buffer images, 3 silhouette images and 3 contours images of the 12 sketches for each 3D shape. Table 11 reports the results. The results show that the proposed method achieves the best retrieval performance when view images are used, while the silhouette images result in the worst performance. Moreover, the depth-buffer images have similar results to view images. The reason is that the view images and depth-buffer images have more information and detail than the silhouette images and contours of view images.

Comparison with different viewpoints of 2D projections. We set different numbers of rendered views numbers and view angles to explore effects on the retrieval performance. Specifically, the rendered view numbers are set to 2, 4, 6 and 12 by placing the cameras evenly

Fig. 12. Sketch examples in SBASR and SHREC 2014 dataset, showing that sketch data in SBASR dataset are more abstract and challenging.

Table 10
Results when feature mapping occur at different layers of two FCNs on SHREC 2014 dataset.

Layers	NN	FT	ST	EE	DCG	mAP
Layer-3	**79.5**	**82.0**	**85.5**	**41.9**	**88.4**	**83.6**
Layer-1, 3	77.3	79.2	83.2	39.8	86.0	80.7
Layer-2, 3	77.4	79.5	82.8	40.1	86.2	80.9
Layer-1, 2, 3	77.4	79.4	82.8	40.0	86.0	80.8

Table 11
Results on different 2D projections on SHREC 2014 dataset.

2D images	NN	FT	ST	EE	DCG	mAP
View	**79.5**	**82.0**	**85.5**	**41.9**	**88.4**	**83.6**
Depth-buffer	79.4	82.1	85.4	41.9	88.5	83.3
Contour	78.9	79.4	83.5	40.8	87.5	81.7
Silhouette	78.2	77.8	82.1	39.9	86.9	79.4
Combination	79.3	82.0	85.1	41.7	88.2	83.0

Table 12
Results on different viewpoints of 3D shapes on SHREC 2014 dataset.

Viewpoints	NN	FT	ST	EE	DCG	mAP
2 views	76.7	79.1	83.4	40.9	86.8	80.5
4 views	78.4	80.8	85.0	41.5	87.8	82.7
6 views	79.2	81.6	85.4	41.5	88.1	83.0
12 views	**79.5**	**82.0**	**85.5**	**41.9**	**88.4**	**83.6**
12 views - invert	79.4	82.0	85.2	41.8	88.2	83.3

around the shape every 180, 90, 60 and 30 degrees. We also put the shape upside down under the setting of 12 views to explore the effect of the shape pose. Table 12 reports the results that the increased ability to represent shapes due to the increased number of views improves retrieval performance. When the number of views is sufficient to comprehensively represent a 3D shape, The improvement caused by increasing the view numbers is relatively small. Meanwhile, as long as the 3D shape can be comprehensively represented, the pose of the shape is not important.

Comparison with different sketching styles. To validate the effectiveness of our methods on various sketching styles. We conduct experiments on the more challenging SBASR-McGill (Sahillioğlu and Sezgin, 2017) dataset. The dataset have 10 classes. Its sketches are collected from novice users, and shapes comes from McGill (Siddiqi et al., 2008) 3D shape benchmark. The sketches in this dataset are more challenging for retrieval as they lack details compared to the sketches in SHREC'13&'14, e.g., many sketches of 'human' and 'teddy' are in form of skeleton in Fig. 12, which is hard to distinguish. As the results shown in Table 13, our method still achieves promising performance. The results validate the robustness of our method to different sketching styles.

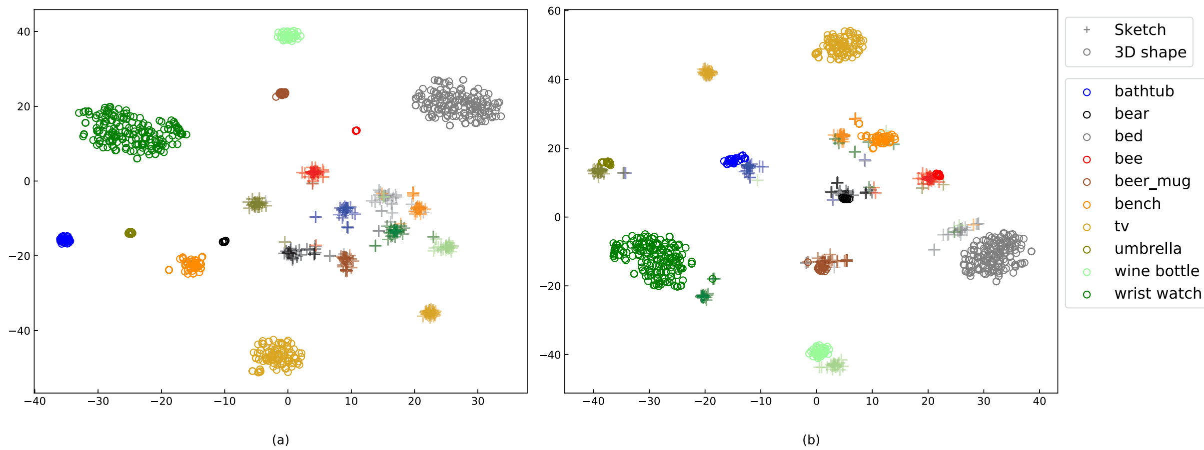

Fig. 13. Visualization of learned features of sketches and 3D shapes. (a) without sketch feature transfer, (b) with sketch feature transfer.

Table 13
Results on different styles of sketches on SBASR-McGill dataset.

Method	NN	FT	ST
Mahoney-Fromherz	56.7	52.3	59.9
SBASR	86.1	78.4	89.1
CFTTSL (Ours)	92.9	86.1	93.6

5.8. Visualization

This subsection shows a visualization of the learned deep features of the samples from the SHREC 2014 dataset. Then, we present some retrieval examples on the SHREC 2014 dataset.

Visualization of learned features. We adopt the t-distributed stochastic neighbor embedding (t-SNE) technique to reduce the dimension to visualize the feature distributions of 10 classes from the test set of SHREC 2014. We visualize the deep features of two modalities without sketch features transferring ($\alpha = 0.0$), and the deep features with sketch features transferring ($\alpha = 0.1$, $\beta = 0.1$). The learned feature distributions are plotted in Fig. 13. For the features without sketch features transferring in Fig. 13(a), the samples of all classes are well clustered within modalities, but they fail to cluster between modalities. For the features with sketch features transferring in Fig. 13(b), samples of most classes are well clustered within and between modalities. This comparison demonstrates that the CFTTSL framework and guidance loss achieve effective cross-modal feature semantic alignment.

Retrieval examples. Fig. 14 shows some retrieval examples of the SHREC 2014 dataset. The query sketches are listed on the left side, and their retrieved top ten 3D shapes are listed on the right side according to their ranking order. As shown, the proposed method achieves promising results for the classes of examples. However, some bad retrieved shapes with gray color are also shown. The main reason is that they have a similar appearance to the correct class.

6. Conclusion

In this paper, we propose a novel two-stage framework named Cross-Modal Feature Transfer via Teacher–Student Learning (CFTTSL) for sketch-based 3D shape retrieval, which adopts the idea of distillation networks to separate the cross-modal retrieval into a classification task and a feature transfer task. To our best knowledge, this work is the first one that incorporates distillation network into sketch-based 3D shape retrieval. In the first classification stage, we first learn the discriminative features of 3D shapes. Then in the feature transfer, our method uses the pre-learned features of 3D shapes to guide the feature learning of sketches, which avoids that the poor-quality sketch features impair the learning of common feature space. To effectively guide the feature transfer of sketches, we formulate a novel guidance loss function to supervise the training of sketch network. As a result, the features of sketches are transferred to the feature space of 3D shapes, resulting in the reduction of cross-modal difference. The results of experiments on three publicly available large-scale datasets (SHREC 2013, SHREC 2014, SHREC 2016) outperform the state-of-the-art approaches. Moreover, some ablation experiments also further demonstrate the effectiveness of the proposed method.

CRediT authorship contribution statement

Shuang Liang: Conceptualization, Supervision, Methodology, Project administration. **Weidong Dai:** Investigation, Software, Writing – original draft, Formal analysis. **Yiyang Cai:** Writing – review & editing, Software. **Chi Xie:** Writing – review & editing, Validation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Shuang Liang reports financial support was provided by National Natural Science Foundation of China. Shuang Liang reports financial support was provided by Natural Science Foundation of Shanghai. Shuang Liang reports financial support was provided by Shanghai Science and Technology Innovation Acton Project. Shuang Liang reports financial support was provided by Shanghai Municipal Science and Technology Major Project.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62076183, 61936014 and 61976159, in part by the Shanghai Science and Technology Innovation Action Project, China of under Grant 20511100700 and 22511105300, and in part by the Shanghai Municipal Science and Technology Major Project, China under Grant 2021SHZDZX0100. The authors would also like to thank the reviewers for their careful work and valuable suggestions.

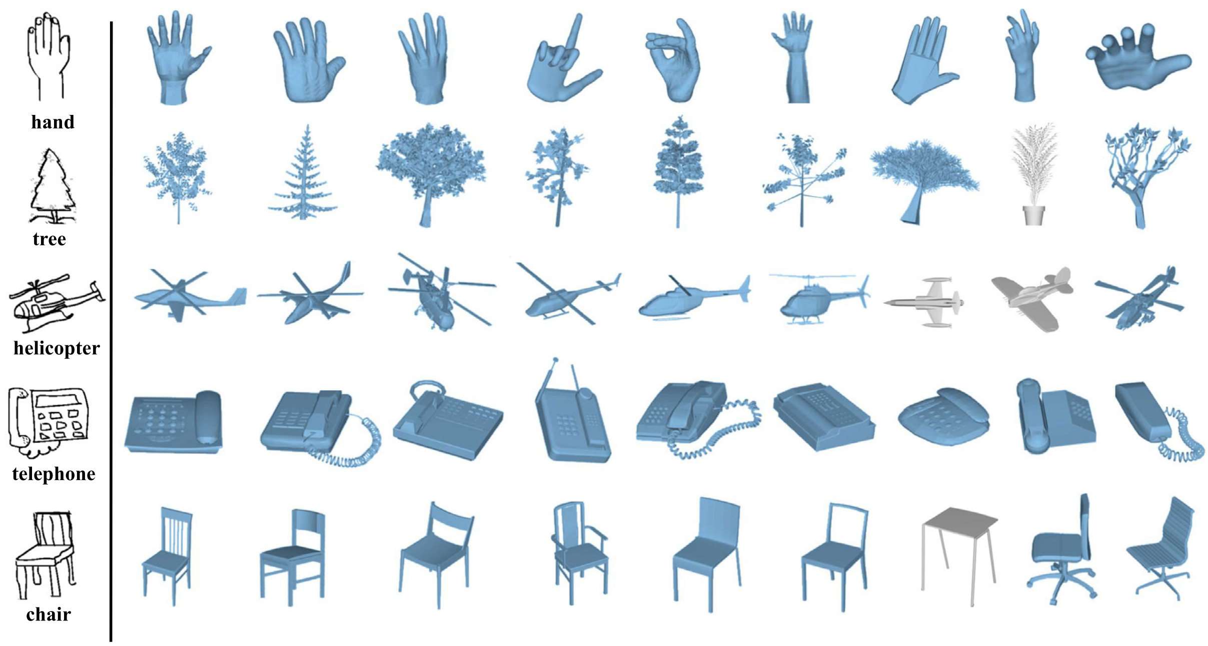

Fig. 14. Retrieval examples of SHREC 2014 dataset. Blue denotes the correct retrieval results while gray denotes the incorrect results.

References

Bai, S., Bai, J., 2023. HDA2L: Hierarchical domain-augmented adaptive learning for sketch-based 3D shape retrieval. Knowl.-Based Syst. 110302.

Bai, S., Bai, J., Xu, H., Tuo, J., Liu, M., 2023. PAGML: Precise alignment guided metric learning for sketch-based 3D shape retrieval. Image Vis. Comput. 104756.

Bhunia, A.K., Koley, S., Khilji, A.F.U.R., Sain, A., Chowdhury, P.N., Xiang, T., Song, Y.Z., 2022. Sketching without worrying: Noise-tolerant sketch-based image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (CVPR), pp. 999–1008.

Cai, Y., Lu, J., Wang, J., Liang, S., 2023. Uncertainty-aware cross-modal transfer network for sketch-based 3D shape retrieval. In: 2023 IEEE International Conference on Multimedia and Expo. (ICME), pp. 132–137. http://dx.doi.org/10.1109/ICME55011.2023.00031.

Chen, J., Fang, Y., 2018. Deep cross-modality adaptation via semantics preserving adversarial learning for sketch-based 3D shape retrieval. In: The European Conference on Computer Vision. pp. 605–620.

Chen, Y., Wang, N., Zhang, Z., 2018. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. In: AAAI Conference on Artificial Intelligence. pp. 4213–4217.

Cristian, B., Rich, C., Alexandru, N., 2006. Model compression. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 535–541.

Dai, W., Liang, S., 2020. Cross-modal guidance network for sketch-based 3d shape retrieval. In: IEEE International Conference on Multimedia and Expo. (ICME), pp. 156–162.

Dai, G., Xie, J., Fang, Y., 2018. Deep correlated holistic metric learning for sketch-based 3d shape retrieval. IEEE Trans. Image Process. 27 (7), 3374–3386.

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. pp. 4002–4008.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Dubey, S.R., Singh, S.K., Chu, W.T., 2022. Vision transformer hashing for image retrieval. In: 2022 IEEE International Conference on Multimedia and Expo. (ICME), pp. 1–6. http://dx.doi.org/10.1109/ICME52920.2022.9859900.

Eitz, M., Richter, R., Boubekeur, T., Hildebrand, K., Alexa, M., 2012. Sketch-based shape retrieval. ACM Trans. Graph. 31 (4), 13–15.

Florian, S., Dmitry, K., James, P., 2015. Facenet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 815–823.

Furuya, T., Ohbuchi, R., 2013. Ranking on cross-domain manifold for sketch-based 3D model retrieval. In: International Conference on Cyberworlds. pp. 274–281.

Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M., 2020. Deep learning for 3D point clouds: A survey. IEEE Trans. Pattern Anal. Mach. Intell..

He, K., Zhang, X., Reng, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

He, X., Zhou, Y., Zhou, Z., Bai, S., Bai, X., 2018. Triplet-center loss for multi-view 3D object retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1945–1954.

Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Krizhevsky, A., Sutskever, I., Hinton, G., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1097–1105.

Lei, Y., Zhou, Z., Zhang, P., Guo, P., Ma, Z., Liu, L., 2019. Deep point-to-subspace metric learning for sketch-based 3D shape retrieval. Pattern Recognit. 96 (7), 106–116.

Li, B., Lu, Y., Duan, F., Dong, S., Fan, Y., Qian, L., Laga, H., Li, H., Li, Y., Lui, P., et al., 2016. SHREC'16 track: 3D sketch-based 3D shape retrieval. In: Eurographics Workshop on 3D Object Retrieval. pp. 151–159.

Li, B., Lu, Y., Ghumman, A., Strylowski, B., Gutierrez, M., Sadiq, S., Forster, S., Feola, N., Bugerin, T., 2015a. 3D sketch-based 3D model retrieval. In: ICMR. pp. 555–558.

Li, B., Lu, Y., Ghumman, A., Strylowski, B., Gutierrez, M., Sadiq, S., Forster, S., et al., 2015b. KinectSBR: A kinect-assisted 3D sketch-based 3D model retrieval system. In: International Conference on Multimedia Retrieval. pp. 655–656.

Li, B., Lu, Y., Godil, A., Schreck, T., Aono, M., Johan, H., Saavedra, J., Tashiro, S., 2013. SHREC'13 track: large scale sketch-based 3D shape retrieval. In: Eurographics Workshop on 3D Object Retrieval. pp. 89–96.

Li, B., Lu, Y., Godil, A., Schreck, T., Bustos, B., Ferreira, A., Furuya, T., Fonseca, M.J., Johan, H., Matsuda, T., et al., 2014a. A comparison of methods for sketch-based 3D shape retrieval. Comput. Vis. Image Underst. 119 (6), 57–80.

Li, B., Lu, Y., Li, C., Godil, A., Schreck, T., Aono, M., Burtscher, M., Fu, H., Furuya, T., Johan, H., et al., 2014b. SHREC'14 track: Extended large scale sketch-based 3D shape retrieval. In: Eurographics Workshop on 3D Object Retrieval. pp. 121–130.

Liang, S., Dai, W., Wei, Y., 2021. Uncertainty learning for noise resistant sketch-based 3D shape retrieval. IEEE Trans. Image Process. 30, 8632–8643. http://dx.doi.org/10.1109/TIP.2021.3118979.

Pandey, A., Misha, A., Verma, V.K., Mittal, A., 2019. Adversarial joint-distribution learning for novel class sketch-based image retrieval. In: International Conference on Computer Vision. pp. 217–226.

Pandey, A., Mishra, A., Verma, V., Mittal, A., Murthy, H., 2020. Stacked adversarial network for zero-shot sketch based image retrieval. In: IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 765–779.

Pang, K., Yang, Y., Hospedales, T., Xiang, T., Song, Y., 2020. Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval. In: Conference on Computer Vision and Pattern Recognition. pp. 346–356.

Qi, A., Gryaditskaya, Y., Song, J., Yang, Y., Qi, Y., Hospedales, T.M., Xiang, T., Song, Y.Z., 2021. Toward fine-grained sketch-based 3D shape retrieval. IEEE Trans. Image Process. 30, 8595–8606. http://dx.doi.org/10.1109/TIP.2021.3118975.

Qi, C., Su, H., Mo, K., Guibas, K., 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 652–660.

Ranjan, R., Castillo, C.D., Chellappa, R., 2017. L2-constrained softmax loss for discriminative face verification. arXiv preprint arXiv:1703.09507.

Sahillioğlu, Y., Sezgin, M., 2017. Sketch-based articulated 3d shape retrieval. IEEE Comput. Graph. Appl. 37 (6), 88–101.

Sain, A., Bhunia, A.K., Chowdhury, P.N., Koley, S., Xiang, T., Song, Y.Z., 2023. CLIP for all things zero-shot sketch-based image retrieval, fine-grained or not. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (CVPR), pp. 2765–2775.

Sain, A., Bhunia, A.K., Yang, Y., Xiang, T., Song, Y.Z., 2021. StyleMeUp: Towards style-agnostic sketch-based image retrieval. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. (CVPR), pp. 8500–8509. http://dx.doi.org/10.1109/CVPR46437.2021.00840.

Shilane, P., Min, P., Kazhdan, M., Funkhouser, T., 2004. The princeton shape benchmark. In: Shape Modeling International. pp. 167–178.

Siddiqi, K., Zhang, J., Macrini, D., Shokoufandeh, A., Bouix, S., Dickinson, S., 2008. Retrieving articulated 3-D models using medial surfaces. Mach. Vis. Appl. 19, 261–275.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E., 2015. Multi-view convolutional neural networks for 3d shape recognition. In: IEEE International Conference on Computer Vision. pp. 945–953.

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. pp. 1875–1883.

Wang, F., Xiang, X., Cheng, J., Yuille, A., 2017. Normface: L2 hypersphere embedding for face verification. In: ACM MM. pp. 1041–1049.

Wen, Y., Zhang, K., Li, Z., Qiao, Y., 2016. A discriminative feature learning approach for deep face recognition. In: ECCV. pp. 499–515.

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. pp. 5068–5076.

Xu, D., Alameda-Pineda, X., Song, J., Ricci, E., Sebe, N., 2018. Cross-paced representation learning with partial curricula for sketch-based image retrieval. IEEE Trans. Image Process. 27 (9), 4410–4421.

Yoon, G., Yoon, S., 2017. Sketch-based 3D object recognition from locally optimized sparse features. Neurocomputing 267 (37), 556–563.

Zagoruyko, S., Komodakis, N., 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928.

Zhao, Y., Liang, Q., Ma, R., Nie, W., Su, Y., 2022. JFLN: Joint feature learning network for 2D sketch based 3D shape retrieval. J. Vis. Commun. Image Represent. 89, 103668.

Zhou, G., Fan, Y., Cui, R., Bian, W., Zhu, X., Gai, K., 2018. Rocket launching: A universal and efficient framework for training well-performing light net. In: AAAI Conference on Artificial Intelligence. pp. 3786–3790.