Toward Fine-Grained Sketch-Based 3D Shape Retrieval

Anran Qi, Yulia Gryaditskaya, Jifei Song, Yongxin Yang, Yonggang Qi, *Member, IEEE*, Timothy M. Hospedales, *Member, IEEE*, Tao Xiang, and Yi-Zhe Song, *Senior Member, IEEE*

***Abstract*— In this paper we study, for the first time, the problem of fine-grained sketch-based 3D shape retrieval. We advocate the use of sketches as a fine-grained input modality to retrieve 3D shapes at instance-level – e.g., given a sketch of a chair, we set out to retrieve a specific chair from a gallery of all chairs. Fine-grained sketch-based 3D shape retrieval (FG-SBSR) has not been possible till now due to a lack of datasets that exhibit one-to-one sketch-3D correspondences. The first key contribution of this paper is two new datasets, consisting a total of 4,680 sketch-3D pairings from two object categories. Even with the datasets, FG-SBSR is still highly challenging because (i) the inherent domain gap between 2D sketch and 3D shape is large, and (ii) retrieval needs to be conducted at the instance level instead of the coarse category level matching as in traditional SBSR. Thus, the second contribution of the paper is the first cross-modal deep embedding model for FG-SBSR, which specifically tackles the unique challenges presented by this new problem. Core to the deep embedding model is a novel cross-modal view attention module which automatically computes the optimal combination of 2D projections of a 3D shape given a query sketch.**

***Index Terms*— Sketch, 3D shape, FG-SBSR, dataset, cross-modal, view-attention.**

I. Introduction

THE ability to retrieve a specific 3D shape from a large collection of 3D shape models underpins many important applications in AR/VR, 3D printing, architectural modelling and film animation. This task is relatively straightforward if retrieval is conducted at category-level, where text queries are often sufficient – typing chair into a retrieval engine will return all chairs. In this paper, we ask a more challenging question – can we conduct retrieval on a fine-grained level, so instead of returning any 3D chair, the system will yield just the one particular chair that is desired?

To answer this question, we first need to consider an appropriate input modality, which would be convenient to use and which naturally encodes a sufficient level of detail to drive retrieval. We follow the findings of the fine-grained sketch-based *image retrieval* community [1]–[6] , who argue that sketch is a favored input modality over traditionally used text for fine-grained (instance-level) retrieval. This is intuitive since, compared to text, a sketch can easily convey appearance and structure details that would be otherwise cumbersome to describe in text.

Sketch-based 3D shape retrieval (SBSR) has been considered before. However, instead of the instance-level or fine-grained SBSR (FG-SBSR) problem tackled in this work, all existing works [7]–[14] focus on category-level SBSR. They typically learn a joint embedding space between 2D sketches and 3D shapes. To mitigate the domain gap between 2D and 3D, state-of-the-art methods mostly employ a MVCNN [15] architecture to represent 3D shapes in 2D.

This paper, for the first time, tackles the fine-grained SBSR problem. There are two reasons why this has never been attempted before. First, instance-level cross-domain matching between sketches and 3D shape models is hard due to the intrinsic domain gaps between the two modalities. The domain gaps of FG-SBSR can be broadly factorized into (i) the dimensionality gap: sketches are represented in 2D, whereas 3D shapes have a third dimension, (ii) the abstraction gap: sketches are highly abstract, yet 3D shapes are geometrically realistic, and (iii) the view gap: sketches are drawn from specific view points, while 3D shape models are entirely view-independent. Although category-level SBSR is also faced with the same domain gaps, bridging them becomes more important when the subtle inter-instance differences need to be modeled.

The second reason is the lack of suitable datasets. To bridge the domain gap, large quantities of sketch-3D shape pairs need to be collected and annotated. However, no such dataset exists. This prohibits any serious attempt to solving the problem. In particular, all existing SBSR datasets, such as the SHREC series of datasets [9], [16], provide only category-level pairings between sketches and 3D shapes. More importantly, they were obtained cheaply by merging existing 3D shape datasets with off-the-shelf sketch datasets that share the same categories.

Manuscript received September 21, 2020; revised April 20, 2021, May 26, 2021, and August 10, 2021; accepted September 24, 2021. Date of publication October 14, 2021; date of current version October 21, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Raymond Fu. *(Corresponding author: Anran Qi.)*

Anran Qi, Yulia Gryaditskaya, Jifei Song, Yongxin Yang, Tao Xiang, and Yi-Zhe Song are with the SketchX Lab, Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford GU2 7XH, U.K. (e-mail: a.qi@surrey.ac.uk).

Yonggang Qi is with the SketchX Lab, Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford GU2 7XH, U.K., and also with the Beijing University of Posts and Telecommunications, Beijing 100876, China.

Timothy M. Hospedales is with the SketchX Lab, Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford GU2 7XH, U.K., and also with The University of Edinburgh, Edinburgh EH8 9YL, U.K.

Digital Object Identifier 10.1109/TIP.2021.3118975

That is, none of the sketches are specifically sourced for a given 3D shape, hence does not exhibit *one-to-one mapping* that we seek. We are faced with a much harder problem of collecting instance-level sketch and 3D pairings. The common practice to collect such data in the sketch-based image retrieval community [1], [17] is to let participants draw by recollection. It is worth noting that free-hand sketches are distinctly different from synthetic edges automatically rendered from 3D shapes, making a synthetic sketch and 3D shape dataset in-viable for practical FG-SBSR (verified by an experiment (in Sec. V-C)).[1]

As the first contribution of this paper, we present two FG-SBSR datasets, consisting of a total of 4,680 sketch-3D *one-to-one pairings* across two categories (chair and lamp). The dataset is built via crowd-sourcing by asking users to finger sketch on a touchscreen device *by recollection*, after observing a model for a fixed amount of time. A key problem that needs to be addressed is collecting a dataset that contains views representative of those that people are most likely and are most comfortable to draw, covering the space of such views well. We address this problem by (i) conducting a pilot study to determine salient views that ordinary users are accustomed to draw, and (ii) collecting more than one sketched view per 3D model from each participant. As a result, we source three corresponding sketches for each 3D model, each of which drawn from a specific view angle. We hope that, by making these two datasets publicly available, we will stimulate research interest in this new computer vision problem.

As the second contribution, we propose to learn a deep joint embedding space that simultaneously addresses all the aforementioned domain gaps. In such a space, the two modalities are aligned, and sketch and 3D shape instances can be compared by simply computing their distance. To overcome the dimensionality gap, we follow the common practice [15], [18] in 3D shape recognition by projecting a 3D shape into multiple 2D views. The key problem left now is to match a sketch from a certain view to the projection of the 3D shape along one or more views (the view gap). To this end, we introduce a novel cross-modal view attention module which automatically selects the best combination of 3D shape views for further deep alignment. The joint embedding model is trained with a triplet ranking loss, which is the most popular choice to tackle the abstraction gap for fine-grained sketch-based image retrieval [1], [17]. With the proposed cross-modal view attention module, a novel triplet sampling strategy is devised which allows for large amount of triplets in the batch, leading to better cross-modal alignment.

We conduct extensive experiments on the two new datasets. The results show that the proposed model significantly outperforms alternatives extended from existing category-level SBSR models and instance-level sketch-based image retrieval models. Importantly, we show that the proposed cross-modal view attention module together with the tailor-made triplet sampling strategy is the key for the superior performance.

[1]This is akin to photo edge maps being a poor substitute for free-hand sketches in sketch-based image retrieval [1], [17].

II. Related Work

A. Shape Recognition

Recent deep recognition methods for 3D shapes can be broadly categorized into four categories, according to how 3D shapes are represented. References [19]–[22] represent shapes as point clouds, a natural representation for a scanned data. Volumetric-based methods [23]–[27] apply 3D convolutional neural network on 3D voxels. Spherical function-based methods [28], [29] encode 3D shape as spherical signals and extend convolutional neural networks to have built-in spherical invariance in order to cope with 3D orientations. View-based methods [15], [18], [30], [31] encode 3D models using a collection of their 2D projections. Notable works include [15], which projects 3D objects into multiple views. Each view is passed through a network which learns discriminative view descriptors, which are combined by view-pooling. Recently, [18] proposed to use bilinear pooling to effectively aggregate convolutional features of different views. In this paper, we adapt a view-based approach to encode 3D shapes, but for the first time study a cross-modal retrieval problem with view attention.

B. Sketch-Based 3D Shape Retrieval

Existing sketch-based 3D shape retrieval (SBSR) methods all focus on category-level retrieval, *i.e.*, given a query sketch, the retrieved 3D shape is considered to be correct as long as it belongs to the same category. The earlier hand-crafted feature based methods [7]–[9], [32] have been followed by the more recent deep learning based models [10]–[14], [33]. All the existing deep category-level SBSR models aim to learn a joint embedding space for the 3D shape and 2D sketch modalities. Most of them follow the multi-view CNN (MVCNN) [15] approach originally designed for 3D shape recognition to project 3D shapes into 2D images of evenly distributed views, with the exception of [33] which models 3D shapes as point clouds and employs PointNet [19], [20] for feature extraction.

Our approach differs significantly from the existing ones in that we for the first time tackle the instance-level FG-SBSR problem, which is made possible by the two new datasets contributed in this paper. Though the problem of focus is different, the proposed FG-SBSR model is related to the deep joint embedding based SBSR models [11]–[14] in the use of 2D projections of 3D shape and triplet ranking loss for embedding space learning. However, our model enables instance-level retrieval and can be trained on a single category. It is uniquely able to select the optimal projection views for 3D shape feature extraction, further reducing domain gaps, and has an effective triplet sampling strategy tailor-made for our view attention module.

C. Instance-Level Sketch-Based Image Retrieval

Another closely related problem is fine-grained (instance-level) sketch-based image retrieval (FG-SBIR), which has received increasing interest recently [1]–[6], [34]. Comparing FG-SBIR with FG-SBSR, the latter is more challenging in that (i) sketch and photo are both in 2D, yet there is a

dimensionality mismatch between sketch and 3D shape, (ii) all existing FG-SBIR datasets assume a common pose between sketch-photo pairs [1], [17], whereas view correspondence has to be separately established in FG-SBSR. As a result, although the models in [1], [3], [5], [34] are also cross-modal joint embedding models, the cross-modal view attention module introduced in this paper is critical to cope with the dimensionality mismatch and view selection problems, as validated in our experiments (see Sec. V-C). Note that FG-SBIR and FG-SBSR share the same difficulties in data collection due to the costly sketch-drawing process. Existing FG-SBIR datasets [1], [17] thus have moderate sizes with hundreds of sketches per object category – similar to those of our FG-SBSR datasets.

D. Attention Mechanism

Attention modules have been introduced to deep models for addressing a variety of different tasks, including but not limited to visual question answering (VQA) [35]–[37], image captioning [38]–[40], object retrieval [5] and detection [41], [42]. Reference [43] first proposed to use soft attention with a recurrent neural network for a machine translation task. Later, [44] proposed the Transformer, an encoder-decoder architecture with self-attention modules, which achieved state-of-the-art results in machine translation. Recently, [45]–[48] adopted the Transformer architecture to diverse vision tasks. Reference [45] used the self-attention mechanism for image generation problem. Reference [47] proposed a Vision Transformer, which divides an image into 16×16 patches before feeding these patches into the standard transformer. For a detailed overview of Transformer-based architectures please refer to the recent survey [49]. Different from most existing attention modules, our cross-modal view attention module is (a) cross-modal and (b) designed for 2D projection view selection/reweighting rather than spatial feature reweighting. Cross-modal attention has been exploited in text-visual multi-modal modelling tasks such as VQA [35], [50], which again serves a different purpose (image spatial-sentence word co-attention vs. view attention).

III. Fine-Grained Shape-Sketches Datasets

The first contribution of our paper lies in introducing the two fine-grained datasets for chair and lamp categories.[2] The two datasets contain a total of 1,560 quadruplets, comprising 3D shapes paired with 3 distinctive image and sketch views each (Fig. 1), forming 4,680 sketch-3D shape pairs. The chair dataset has 1,005 3D shape-sketches quadruplets and the lamp dataset has 555 3D shape-sketches quadruplets.

In this section, we first discuss the data collection task design (III-A). We then provide the rationales behind the choice of the two shape categories: lamps and chairs (III-B.1) and of particular shape instances (III-B.2).

A. Sketch Collection

A good dataset of sketches, on one hand, should reflect well real-world free-hand sketching from one's mind eye. On the other hand, to enable training for fine-grained tasks each sketch in our dataset should depict a particular shape instance. These two criteria are contradictory and form the main challenge in collecting fine-grained datasets of sketches paired with given shape or image instances. The common solution to this problem is to let participants study a reference image for a certain period of time and then let them draw from memory. This approach was used by Eitz *et al.* [51] to collect 43 scene sketches for retrieval performance evaluation, by Antol *et al.* [52] to collect clipart illustrations of people and their interactions, and by Sangkloy *et al.* [17] to collect a fine-grained dataset of sketch-image pairs on 125 categories.

Unlike previous works, which deal with images, we want to have an association between a 3D shape and its sketch representation. This problem raises a new question of which views should the dataset include. The design of our data-collection task aims at the dataset that (1) contains commonly selected views, and (2) covers the space of such views well. To achieve this, we first render each shape reference images from three distinctive views, those number and viewpoints are selected according to sketch literature and our pilot study (Sec. III-A.1). Then, following previous work, we show one reference image to a volunteer for 15 seconds, then display a blank canvas and let the volunteer sketch the object that he/she just saw from memory using fingers on a tablet/phone. 30 volunteers are recruited to sketch the reference images. Volunteers did not have any art training and thus represent the general population who might use the fine-grained SBSR system. First, each reference image is sketched by two different volunteers. After finishing collecting all sketches, for quality control purposes, three additional volunteers vote to select the best sketch out of the two, for each reference image.

As can be observed in Fig. 1, sketches collected this way convey well the reference shape, but are not pixel-aligned, exhibiting typical to free-hand sketches distortions. Moreover, viewpoints do not match the reference precisely, as expected from amateur drawers, and contributing to the diversity of viewpoints in our dataset. Below we detail the selection of the reference viewpoints.

1) Views Selection: Our data collection task design requires providing the participants with a reference viewpoint. In design literature [53] and sketching systems [54]–[58], to represent a 3D shape, it is common to use frontal, side and an "informative" 3/4 bird's-eye perspective view, which reduces shape ambiguity. To cover all these settings in our dataset we ask each participant to sketch from 3 viewpoints. We ensure the bird's-eye perspective view by setting the camera zenith angle to 20°. Since in our sketch collection task participants are sketching from memory, the sketched viewpoints will slightly vary around the reference viewpoint. Thus, the reference viewpoint should just reflect well the mean preferred viewpoints. To select the most preferred azimuth camera positions for each category, we conduct a pilot study. 20 participants are each presented with 200 3D models (100 chairs and 100 lamps), that they can manually rotate from azimuth 0° to 90° at 15° intervals.[3] While rotating, each is asked to choose 3 views per model that they are mostly

[2]The datasets link: http://sketchx.ai/downloads/: AmateurSketch-3DChair.

[3]We empirically found that ordinary people are unable to reliably produce sketches for finer view differences.

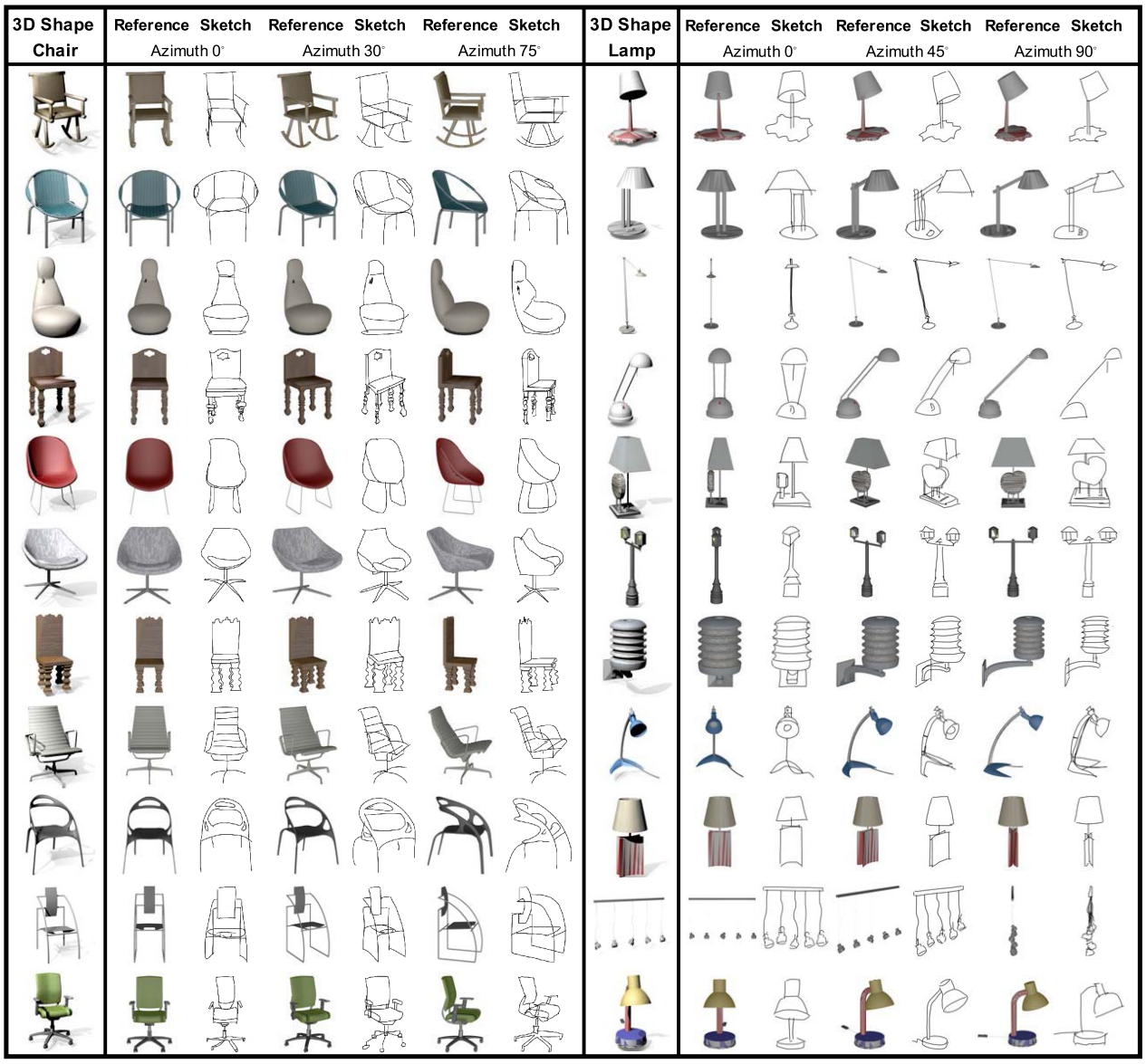


Fig. 1. Example quadruplets of 3D shapes and 3 reference viewpoints with the corresponding sketched views from the proposed chair and lamp datasets. This figure demonstrates the diversity of shapes and highlights the appearance difference/domain gap between 3D shape renderings and realistic free-hand sketches. See Sec. III for the details on data collection and views selection.

likely to sketch. We then aggregate this view selection data ($6{,}000 = 20 \times 100 \times 3$ data points per category), and choose the top 3 most selected views as the ones which we collect sketches for. They are 0°, 30°, 75° for chairs, and 0°, 45°, 90° for lamps.

The selected viewpoints are in agreement with the previous work on sketching and design literature, mentioned above. Our study highlights that the most preferred viewpoints differ slightly between the two set of shapes. The 45° and 90° views represent the lamps shape well. In case of chairs these views can cause clutter and accidental occlusions. This explains a choice of 30° and 75° viewpoints for this category, as more "informative" views, instead.

B. Category and Shape Instance Selection

1) 3D Shape Category Selection: The 3D shapes used in our datasets are selected from the largest 3D shape dataset ShapeNet [59]. Among the 270 object categories, chair and lamp categories are chosen for the following reasons: (1) They are among only a handful of categories that provide over 1000 instances per category. (2) Objects in these two categories have a lesser degree of symmetry; as a result, when viewed from different angles, the appearance varies (see Fig. 1). In contrast, categories such as wine bottle are much less sensitive to view angle. This view-sensitive nature of 3D shapes makes these two categories more challenging for fine-grained sketch-based 3D shape retrieval (FG-SBSR).

2) 3D Shape Instance Selection: For each category, we manually select 3D shape instances to be used in our datasets. Inspired by [60] and [61], the following criteria are used for instance selection:

1) **Representativeness**: There are many subcategories of chairs and lamps in ShapeNet (*e.g.*, armchair, lounge chair, Windsor chair for chairs, and floor lamp, table

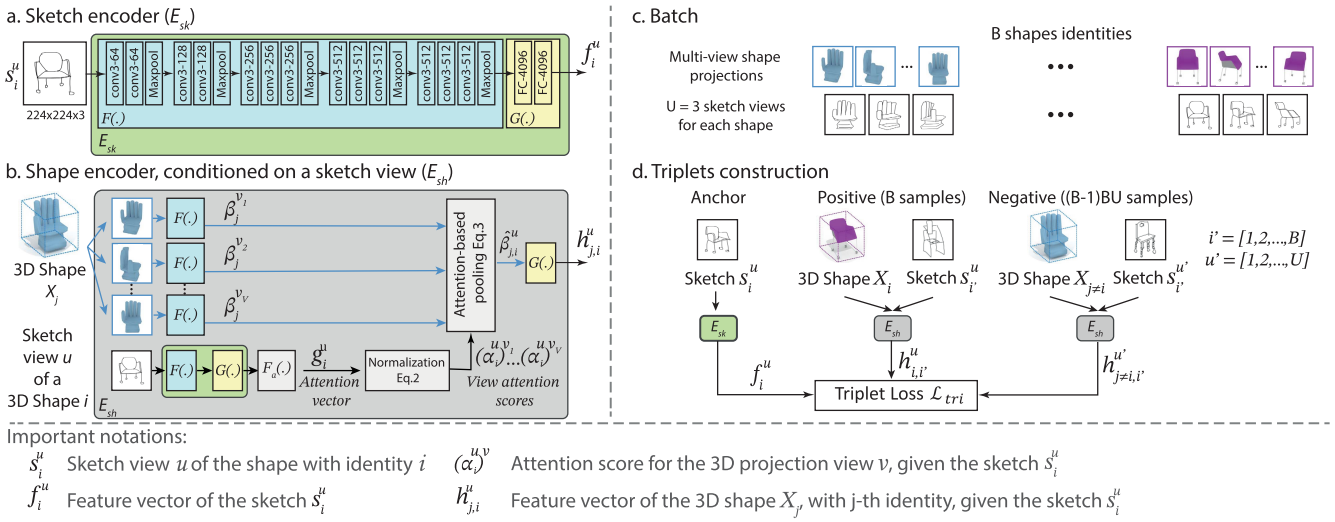


Fig. 2. An illustration of the proposed fine-grained SBSR with cross-modal view attention. The details of the sketch (a.) and shape encoders (b.) are described in Sec. IV-B. The attention module $F_a(\cdot)$, shown in (b.) is detailed in Sec. IV-C. The details of triplets construction are provided in Sec. IV-D.

lamp for lamps). We select the representative instances from each subcategory.

2) **Distinctness**: The selected instances in each subcategory need to be visually distinct so that their differences can be visually depicted by sketches.
3) **View-sensitivity**: As mentioned earlier, the two categories are chosen because they are in general view-sensitive. However, there are still some instances which will produce identical images when projected to different views. These instances are not chosen.
4) **Sketchability**: The 3D shape should be easy to sketch. The free-hand sketches are drawn by people with diverse drawing skills to represent real-world application scenarios. We, therefore, avoid 3D shapes that contain complicated texture that poses a distraction for the sketch drawers.

Following these four criteria, 1,005 and 555 3D shapes are selected for chair and lamp categories, respectively.

IV. Sketch-Based Shape Retrieval Method

In this section we provide details of our sketch-based shape retrieval method.

A. Problem Definition and Model Overview

Our goal is to retrieve a 3D shape given one free-hand sketch view from an unknown viewpoint. To reduce the domain gap, we represent a 3D shape via a discrete set of its 2D projections. To perform retrieval we need to define a strategy for computing relevance scores between a given sketch and a 3D shape. We compute the scores in a feature space as the Euclidean distance between sketch and shape feature vectors. Our key idea is to introduce a cross-modal view attention module (Sec. IV-C) that automatically computes the relevance weights of 3D shape projections to an input sketch view. These weights are used to fuse the feature vectors of each of the individual shape projections to obtain a 3D shape feature vector, conditioned on a given sketch view. To obtain joint embedding of sketch and shape views we exploit a Siamese network, namely the parameters of sketch and shape views encoders are shared. We, further, train the model with a triplet ranking loss formulation with a specifically designed triplet sampling strategy. Fig. 2 shows a schematic illustration of the model, and summarizes the notations used in this section.

B. Sketch and 3D Shape Projections Embedding

Siamese networks have been shown to be effective in instance-level sketch-based image retrieval [1]. We adopt a Siamese architecture in our model to learn a joint embedding space for sketch and 3D shape modalities. For the backbone network that extracts feature vectors from a sketch view and 3D shape projections, we employ VGG-16 (config. D) [62], and remove the final class label prediction layer (Fig. 2a.). The output of the final layer (without ReLU function) is used as the deep feature. The feature extraction network consist of two networks. The first part of the network $F(\cdot)$, consisting of 13 convolutional layers, converts an input sketch view or projected view to a $7 \times 7 \times 512$ feature map. The second part of the network $G(\cdot)$, consisting of 2 fully connected layers, extracts the final feature vector of a dimensionality 1×4096.

To obtain a feature vector of a 3D shape, each projected view is first passed separately to $F(\cdot)$. Views features are then aggregated to one feature map via cross-modal view attention, described in the next section. Finally, the aggregated 3D shape feature map is passed through the network $G(\cdot)$ to obtain the final 3D shape feature vector (Fig.2b.).

C. Cross-Modal View Attention

Certain views of the 3D shape can be drastically different from that of the sketch, even if a sketch and 3D shape projections depict the same shape. It is thus important to dynamically, given an input sketch view, determine the relevance of each 3D shape projection view, prior to fusing

individual 3D shape projections feature vectors into the 3D shape feature vector. To this end, we propose a cross-modal view attention module, which generates a view selection vector used for the fusion of 3D shape projections.

Let's denote a sketch view as s_i^u, where u indicates the viewpoint and i is a depicted 3D shape identity. Then, $f_i^u = G(F(s_i^u)) \in \mathbb{R}^{4096}$ denotes the sketch feature vector.

We obtain the *attention vector* $\mathbf{g}_i^u \in \mathbb{R}^V$, where V is the number of 3D shape views, by feeding the sketch feature vector into the attention module $F_a(\cdot)$:

$$\mathbf{g}_i^u = F_a(f_i^u; W_a), \tag{1}$$

where W_a are the weights/parameters of that module. In our model, the view attention module is a network consisting of one fully connected layer ($4096 \rightarrow V$). The v-th element in the attention vector $\mathbf{g}_i^u$, denoted as $(g_i^u)^v$, represents the attention score for the 3D shape projection view v, given the sketch view u of the shape with an identity i.

We then design a specific normalization scheme to refine the attention vector. The final view *attention score* $(\alpha_i^u)^v$ is calculated as follows:

$$\begin{aligned}(\alpha_i^u)^v &= \ell_2\_\text{softmax}((g_i^u)^v; \tau) \\ &= \frac{\exp(\tau^{-2}(g_i^u)^v / \|\mathbf{g}_i^u\|_2)}{\sum_{v=1}^{V} \exp(\tau^{-2}(g_i^u)^v / \|\mathbf{g}_i^u\|_2)}.\end{aligned} \tag{2}$$

We use softmax to obtain valid probability vectors. We observe that some elements of $\mathbf{g}_i^u$ are very large, while others are small. This leads to numerical instability if these values are directly passed to the exponential function. We, therefore, normalize $\mathbf{g}_i^u$ by its ℓ_2 norm. We further employ scaling by a temperature parameter. Such parameter was used in many previous works [63], [64]. In our work, we use an unconventional design with τ being a trainable parameter. In our experiments, the predicted values are greater than one and result in smoother distribution vectors. We hypothesize that such vectors allow to obtain more accurate feature vectors of 3D shapes, leveraging information from multiple neighboring views around the sketch viewpoint.

The fusion of the feature vectors of 2D projections of the j-th 3D shape into the *3D shape feature vector* $\hat{\beta}_{j,i}^u \in \mathbb{R}^{4096}$, *conditioned on the input sketch of the shape i, sketched from the view u*, is then computed as a weighted sum:

$$\hat{\beta}_{j,i}^u = \sum_{v=1}^{V} (\alpha_i^u)^v \beta_j^v, \tag{3}$$

where $\beta_j^v \in \mathbb{R}^{7\times7\times512}$ is the v-th projection feature of the j-th 3D shape, extracted using $F(\cdot)$. Thus, Eq. 3 is a weighted sum (the weight is attention) of V projected 3D shape images' convolutional features. After that, we feed the attended feature $\hat{\beta}_{j,i}^u$ into the network $G(\cdot)$ to generate the final feature vector $h_{j,i}^u$, which is of the same dimensionality as the sketch feature vector, namely, 4096 (Fig. 2b.).

D. Triplet Sampling Strategy

To train our model we use the triplet ranking loss, commonly used to train for cross-domain retrieval tasks [1], [17].

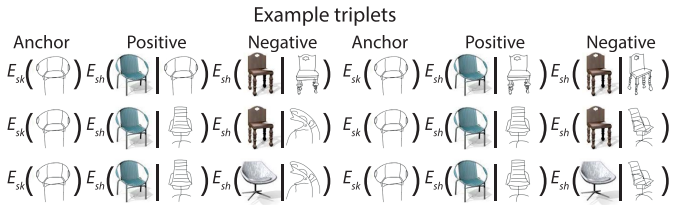


Fig. 3. Example triplets. E_{sk} denotes the sketch encoder and E_{sh} denotes the shape encoder. 3D shapes are represented via their multi-view projections. Note that a shape encoding is conditioned on a sketch view, which provides attention view scores for the 3D shape views pooling.

The query is referred to as an anchor – in our case a sketch view is used as an anchor. The corresponding ground-truth from a different domain is referred to as a positive sample – in our case it is a paired 3D shape. In addition it exploits the notion of a negative sample, which in our case is any 3D shape different from the ground-truth paired 3D shape. The goal of the triplet loss is to ensure that the anchor-negative distances are larger than the anchor-positive distances by a given margin.

We propose a triplet sampling strategy tailored for our model with view attention. The objective is to greatly increase the number of triplets that can be formed from a mini-batch of sketch and 3D shapes. Each mini-batch consists of B shape identities, comprising 3 B sketches and B 3D shape with $V\,B$ rendered 2D views (Fig. 2 c.).

Intuitively, the view attention scores should be independent to the shape/sketch identities, and dependent only on the viewpoint of the anchor sketch. Since our shape encoder is conditioned not only on a 3D shape, but also on a sketch from a certain viewpoint, we can obtain multiple positive and negative 3D shape encodings, by reusing the attention weights computed from sketches with different identities. Fig. 3 shows example triplets.

1) Positive Samples: For the positive shape encoding we use the attention weights from all sketches in a batch that have the same viewpoint as an anchor sketch. Given an anchor sketch s_i^u of a shape with an identity i, drawn from a viewpoint u, the positive 3D shape feature instance is defined as:

$$h^{+u}_{i,i'} = G(\hat{\beta}^u_{i,i'}) = G\left(\sum_{v=1}^{V} (\alpha_{i'}^u)^v \beta_i^v\right), \quad i' \in [1, 2, \ldots B], \tag{4}$$

where $\{\beta_i^v\}_{v=1}^V$ are features of multi-view projections of a 3D shape with the same identity i as the anchor sketch. Each projection feature β_i^v is attended by a view attention score $(\alpha_{i'}^u)^v$, which is predicted from a sketch $s_{i'}^u$ from the same view u as the anchor sketch of a shape with any identity i'.

As a result, for each anchor sketch s_i^u in one mini-batch, there are B positive samples, due to augmentation by B view attention vectors, where B is the number of sketch/shape identities in the batch. These positive samples form a set $\mathcal{P}_i^u = \{h^{+u}_{i,i'},\ i' \in [1, 2, \ldots B]\}$.

2) Negative Samples: The negative 3D shape feature is formed as:

$$h^{-u'}_{j\neq i,i'} = G(\hat{\beta}^{u'}_{j,i'}) = G\Bigg(\sum_{v=1}^{V}(\alpha^{u'}_{i'})^v \beta^v_j\Bigg), \quad j \neq i, \quad u' \in [1,2,\ldots,U],\ i' \in [1,2,\ldots,B]. \tag{5}$$

As per conventional triplets construction, the anchor sketch and the negative sample 3D shape identities, i and j, respectively, are different $j \neq i$. Thus, any attention vectors from any shape identity i' or sketch view u' can be used to attend the projections' features of a negative 3D shape sample. Namely, all the view attention vectors in one batch can be used to attend the negative projection features. This results in $(B-1)BU$ negative samples. The negative samples form a negative 3D feature set for an anchor sketch s^u_i:

$$\mathcal{N}^u_i = \left\{h^{-u'}_{j\neq i,i'},\ u' \in [1,2,\ldots,U],\ i' \in [1,2,\ldots,B]\right\} \tag{6}$$

3) Triplet Loss: Our view-aware triplet loss is defined as:

$$\mathcal{L}_{tri} = \sum_{i=1}^{B}\sum_{u=1}^{U}\sum_{p=1}^{|\mathcal{P}^u_i|}\sum_{n=1}^{|\mathcal{N}^u_i|}\max\Big(0,\ \Delta + D\big(f^u_i, h^+_p\big) - D\big(f^u_i, h^-_n\big)\Big), \tag{7}$$

where h^+_p and h^-_n denote the p-th/n-th 3D shape features from the positive set $\mathcal{P}$ and negative set $\mathcal{N}$, respectively; Δ is the margin, $D(\cdot)$ is the ℓ_2 distance function. Note that we constrain both sketch and 3D embedding such that they live on the multi-dimensional hypersphere, *i.e.*, $\|G(\cdot)\| = 1$. See Fig. 2d for the schematic illustration of the triplet construction process, and Fig. 3 visualizes example triplets.

We show in our experiments (see Sec. V-C) that the proposed sampling strategy is much more effective than the more naive approach of computing a 3D shape feature vector using only the attention scores computed from the sketches which have the same identity as a 3D shape.

V. Experiments

A. Experiment Settings

1) Dataset Splits and Pre-Processing: There are 1,005 and 555 sketch-3D shape quadruplets in the introduced chair and lamp datasets, respectively. Of these, we use 804 and 444 quadruplets respectively (*i.e.*, 80%) for training, and the rest for testing. Recall that each sketch-3D shape quadruplet contains three sketches of different views and one 3D shape.

We obtain multi-view shape representation, as in [15], by putting the centroid of the shape at the origin of the spherical coordinate system and translate camera uniformly. For our main experiments, we render $V = 24$ view projections with model fitted within an image frame.

We resize all sketches/3D views to the same size of 224×224.

2) Implementation Details: The model is implemented using Tensorflow. We use the Adam optimizer with an initial learning rate of 0.0001. The batch size B is set to 3, meaning that each batch contains 3 shape identities with 3 sketch views each and 3×24 2D view projections. The margin Δ in the triplet loss is 0.3 (see Eq. 7). The model is pretrained on ImageNet [65], then trained for 50 epochs for each dataset. All the methods in our experiments are trained for 50 epochs. The trainable temperature variable τ (see Eq. 2) is initialized to 2.0. We train our model on a GTX FORCE 1080. The training takes approximately 4 hours. At test time the retrieval results are obtained in under 30 seconds.

3) Evaluation Metric: To evaluate the retrieval accuracy, we use a commonly used metric *Top-K retrieval accuracy (acc@K)*, which is calculated as the percentage of query sketches whose true-match 3D shapes are among the top K ranked retrieval results. The retrieval results are ranked according to a Euclidean distance between the feature vector f^u_i of the query sketch s^u_i and the feature vector $h^u_{j,i}$ of a 3D shape X_j in the shapes gallery.

B. Baselines

As discussed in Sec. II, there are no existing FG-SBSR models, as we study this problem for the first time in this paper. Therefore, we design several baselines by merging existing category-level SBSR models with FG-SBIR models. Besides, we compare with alternative 3D shape representation learning networks that do not require 2D projection, including a point cloud CNN [19] and a spherical CNN [29]. In total, we evaluate 6 different baselines described below.

1) Sketch-Based Single View 3D Shape Retrieval (SBSVSR): Our first baseline builds on the FG-SBIR model proposed by Yu *et al.* [1]. This model is trained with the triplet loss, and assumes that for each sketch there is one paired image. In our case each sketch is paired with V views of a 3D shape, and the appearance between a sketch and individual shape views can differ drastically.

To address this problem, each 3D shape is rendered to 3 views in accordance with the 3 sketch views. To form the triplets, each of the three sketch views are selected as anchors. As the positive sample we select the view of the 3D shape with the same identity as the 3D shape in the selected anchor sketch, and the viewpoint matching the anchor sketch. As the negative sample we use any view of the 3D shape with a different identity from the one in the sketch. For a fair comparison with our method, we use $F(G(\cdot))$ as sketch and 3D shape view encoders. In addition, we jointly train a sketch view classifier to classify each sketch into one of the three views.

During testing for each query sketch, we predict the view class first. We then select the 3D shape view of the same class, which is used to obtain a 3D shape feature vector.

2) Fine-Grained Triplet Based on MVCNN [15] (FG-T-M): In this baseline, we follow most category-level SBSR models [11]–[14] approach to obtain a 3D shape feature vector. We first project 3D shapes into $V = 24$ views, as in MVCNN [15], and extract each view feature vector with $F(\cdot)$. The resultant feature vectors are then fused

by max-pooling, *i.e.*, without assigning different weights to different views as our model does. The overall network architecture resembles that of [1] and a triplet loss is also adopted as a supervision. In summary, the main difference between this model and ours is the cross-modal view attention module.

3) Fine-Grained Triplet With Spatial Attention (FG-T-A-M): [5] proposed an improved FG-SBIR model which includes a soft-attention module in both sketch and photo branches. Here, we introduce the same spatial attention module, as in [5], to FG-T-M.

The attention module $\tilde{F}_a(\cdot)$ is modeled with two convolutional layers. We train attention modules for a sketch and 3D shape view input separately. Thus, the sketch s_i^u feature vector is obtained as $f_i^u = F(\tilde{F}_a(G(s_i^u); \theta_1))$, where θ_1 are the parameters of a sketch attention module. The 3D shape feature vector is obtained as $h_i^u = F(\tilde{F}_a(\text{maxpool}(G(x_i^{v_1}), \ldots, G(x_i^{v_V})); \theta_2))$, where θ_2 are the parameters of a 3D shape attention module, and x_i^v is the v-th view of the i-th 3D shape, 'maxpool' denotes max pooling.

4) Fine-Grained Triplet Based on VNN [66](FG-T-V): Reference [66] proposed a framework named View N-gram Network (VNN) which divides the rendered view sequence into a set of visual n-grams to help capturing spatial information among multiple views. Our FG-T-V baseline adopts VNN with n-gram sizes of 3 instead of a MVCNN encoder of 3D shapes. The network design otherwise is the same as of the FG-T-M baseline (Sec. V-B.2).

5) Fine-Grained Triplet Based on Non-Projection Based 3D Deep Embeddings (FG-T-P) and (FG-T-S): The general architecture of these two baselines is the same as of the FG-T-M baseline (Sec. V-B.2). As a 3D shape encoder we exploit non-view based encoders: PointNet++ [19] and Spherical CNN [29], to form the FG-T-P and FG-T-S baselines, respectively.

C. Results

1) Comparisons Against Baselines: Table I shows the Top-K retrieval accuracy (acc@K) of our model and the six baselines, introduced in the previous section. Namely, acc@1 shows the percentage of query sketches for which the true-match 3D shape is the first in the ranked list of returned retrieval results; acc@5 shows the percentage of query sketches for which the true-match 3D shape is among the first 5 retrieved shapes in the ranked list of returned retrieval results. The following observations can be made: (1) Our model outperforms all baselines on both datasets. (2) The single view model SBSVSR is inferior to all models that project 3D shapes to 2D views and then fuse them. This is despite we use the ground truth view information of the sketch. (3) Among the baselines, FG-T-V is the most competitive. Yet, the consistent performance gaps (around 2% lower on acc.@1 for both datasets, and around 3.8% and 1.8% lower on acc.@5 for chairs and lamps, respectively) indicate that performing view selection rather than capturing spatial information among different views makes a difference in the model performance. (4) The spatial attention module in FG-T-A-M failed to improve the performance of FG-T-M. This suggests that view-attention is more effective than spatial attention for FG-SBSR. (5) Non-projection based methods are considerably inferior to projection-based methods on the task of FG-SBSR. In particular, FG-T-S captures little fine-grained information due to spectral pooling. The FG-T-P model completely failed, despite that the same PointNet++ has been used in the state-of-the-art category-level SBSR model [33]. It is better in capturing global shape information than particular shape details, and is thus more suitable for category-level retrieval tasks.

TABLE I
COMPARATIVE RESULTS AGAINST BASELINES DESCRIBED IN SEC. V-B

Method	Num. param.	Chair Dataset		Lamp Dataset	
		acc.@1	acc.@5	acc.@1	acc.@5
SBSVSR[1]	134M	44.94	77.94	48.05	79.87
FG-T-M[15]	134M	47.60	81.26	49.25	83.48
FG-T-A-M[5]	134M	47.10	79.10	48.95	81.68
FG-T-V[66]	184M	54.56	83.25	54.05	85.58
FG-T-P[19]	136M	0.50	4.48	0.90	3.60
FG-T-S[29]	135M	11.77	40.13	12.61	38.44
Our model	134M	**56.72**	**87.06**	**57.66**	**87.39**

These results highlight the vital difference between category-level and instance-level SBSR: namely, for instance-level retrieval, it is important to take into account the view information of the input sketch. Our results further show that multi-view 3D shape representation is more advantageous than the 3D-based shape representations in conjunction with existing shape encoders for the task of SBSR.

2) Qualitative Results: In Fig. 4, we show some examples of fine-grained SBSR results obtained using our model. The second column is the query sketch and next sequentially lists the top 6 retrieval results, where the true matches are highlighted in green. We can see that our model is capable of capturing subtle differences between similar 3D shapes, and is robust to sketch inaccuracies and viewpoints changes.

3) Method Convergence: Fig. 5, shows the convergence curve of the triplet loss. It can be seen that the loss does not change much from 40-th to 50-th epochs.

4) Which Views Are Attended to?: In order to understand why the cross-modal view attention module helps the FG-SBSR model, some examples of view attention vector α are visualized in Fig. 6. It can be seen that our attention module is able to identify the correct view angle of the sketch and gives the biggest weight to corresponding 2D projections. Importantly, other views are also given some weights, with the nearby views given more weight than faraway views. This is expected because nearby views are visually similar but not identical. Thus additional views offer some complementary information. It is critical to note that the view attention is instance-dependent: which nearby views should be used and with which weighting factors depends on the object instance and the sketch. This is why, as shown in Table I, when the baseline SBSVSR uses only one view, even though the view is selected correctly in most cases, the performance is drastically worse. Fusing multiple views, and fusing them intelligently, is thus the key for learning an effective FG-SBSR model.

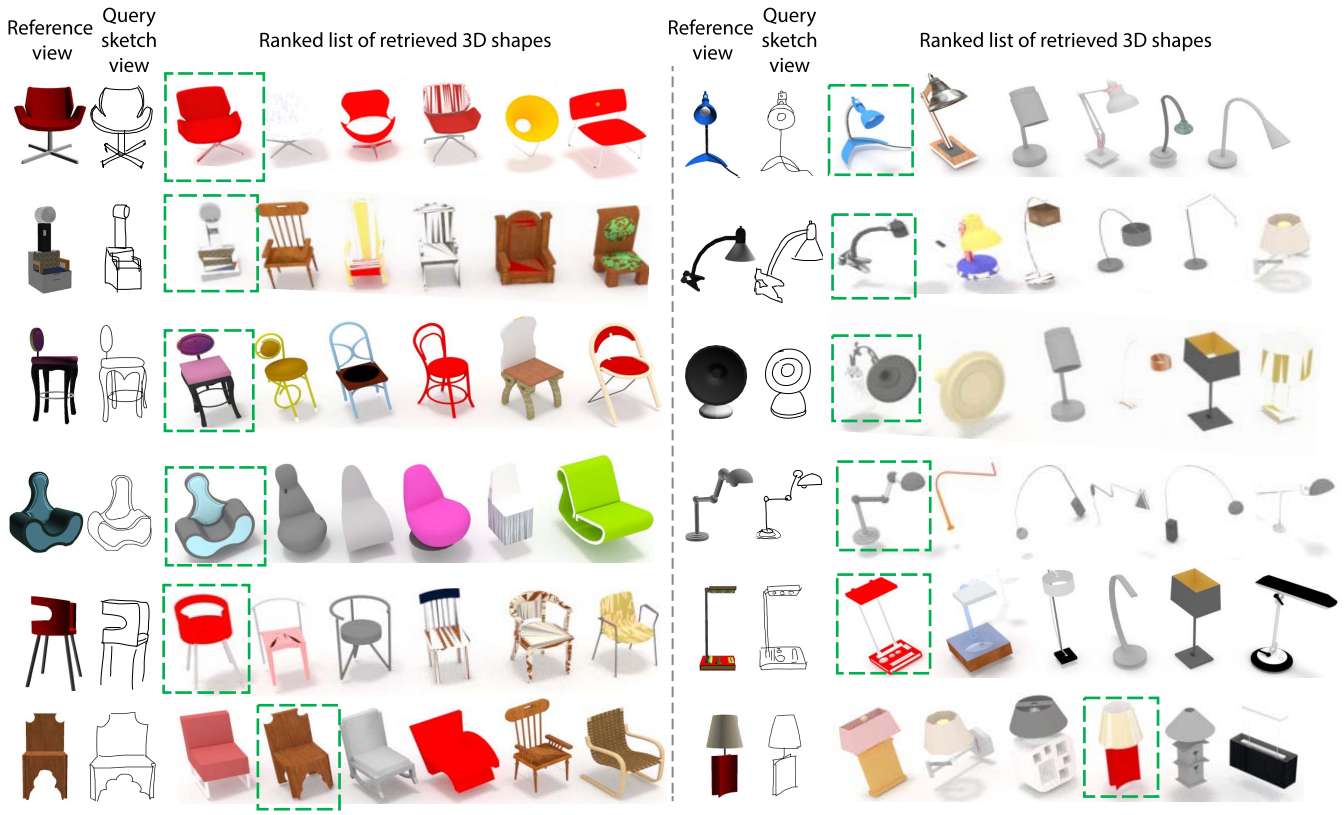


Fig. 4. Qualitative results of our method, detailed in Sec. IV. For each query sketch, the top 6 ranked 3D shapes in the gallery are shown in each row. The green-dashed line rectangles highlights a true-match 3D shape for each sketch.

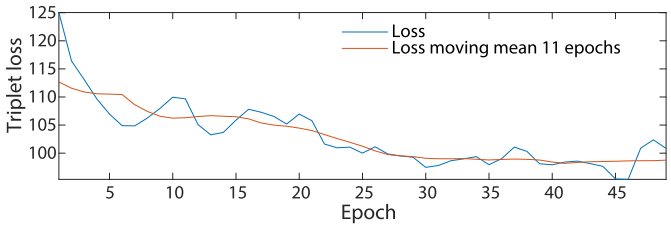


Fig. 5. Triplet loss convergence curve.

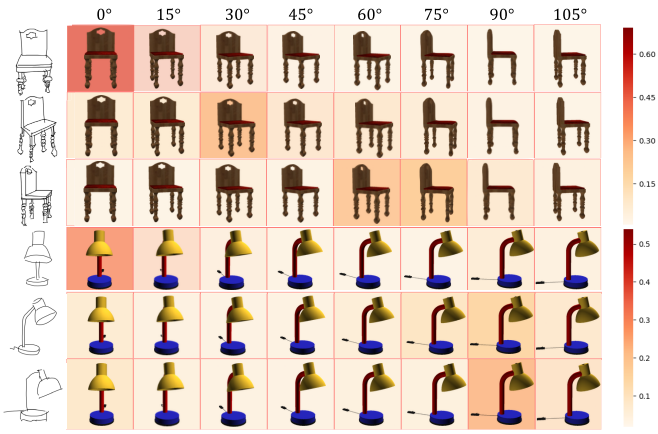


Fig. 6. Examples of attention distribution. We show 8 evenly distributed view angles which the 3D shapes are projected to. Each projection is colour coded with the corresponding attention values indicating which view is attended more for fusion. Warmer colour means higher attention value.

D. Ablation Studies

In this section we provide an evaluation of the effectiveness of a number of design choices.

1) Triplets Sampling Strategy: The triplet sampling strategy described in Sec. IV-D is unique to our model, where the positive and negative 3D shapes are sampled using shapes identities, but their feature representations are obtained by using view-attention weights from sketches of the shapes with different identities, as described in detail in Sec. IV-D. Such strategy allows us to have large number of triplets within one batch, resulting in more efficient training.

We compare this strategy with a more naive sampling strategy: The positive and negative 3D shapes are sampled using shapes identities, but when constructing a positive 3D shape feature vector, only the attention vector obtained from the anchor sketch is used. We refer to this strategy as *Naive Triplet Sampling (NTS)*.

Table II shows that our proposed sampling strategy brings about 1% and 3% increase in acc.@1 for the chair and lamp datasets, respectively, over the more naive sampling strategy.

2) Trainable Temperature Variable τ: The trainable temperature variable, τ in Eq. 2 is an unconventional design. Typically in a softmax formulation the temperature is either fixed or tuned using a validation set. Table II shows that when we omit the temperature parameter, by setting it to

TABLE II

NUMERICAL EVALUATION OF OUR DESIGN CHOICES, SEE SEC. V-D FOR THE DETAILS OF EACH OF THE EXPERIMENTS

Method	Chair Dataset		Lamp Dataset	
	acc.@1	acc.@5	acc.@1	acc.@5
NTS	55.89	85.07	54.65	84.98
$\tau = 1$	51.08	83.91	53.15	84.68
$\tau = 2$	54.38	85.90	55.56	85.59
Heterogeneous	26.70	73.80	28.83	71.47
Our model	**56.72**	**87.06**	**57.66**	**87.39**

TABLE III

CHOICE OF THE MARGIN VALUE IN THE TRIPLET LOSS

Method	Chair Dataset		Lamp Dataset	
	acc.@1	acc.@5	acc.@1	acc.@5
$\Delta 0.1$	55.56	83.08	56.76	86.49
$\Delta 0.5$	55.06	82.92	56.45	85.89
Our model ($\Delta 0.3$)	**56.72**	**87.06**	**57.66**	**87.39**

TABLE IV

NUMERICAL EVALUATION OF OUR METHOD WITH DIFFERENT NUMBER OF VIEWS, USED TO REPRESENT A 3D SHAPE

Method	Chair Dataset		Lamp Dataset	
	acc.@1	acc.@5	acc.@1	acc.@5
Our 6 view	53.06	83.75	55.56	83.78
Our 12 view	55.06	84.58	56.76	85.28
Our model (24)	56.72	87.06	**57.66**	87.39
Our 36 view	**57.21**	**87.23**	**57.66**	**87.69**

1.0, the model performance degrades as the view attention is less effective. Our method estimates the optimal values of this parameter for the chair and lamp datasets as $\tau = 1.6$ and $\tau = 1.7$, respectively. When we fix $\tau = 2.0$ value the performance is inferior to the performance with trainable parameter (Table II), further demonstrating the efficiency of our strategy.

3) Siamese vs. Heterogeneous Network: We implement a Heterogeneous alternative of our network (termed *Heterogeneous* in Table II). It can be seen that the Heterogeneous architecture yields much lower performance, indicating that it suffers severely from over-fitting as the number of parameters doubles. This also echoes findings from the related work on SBIR [1], where Siamese networks are commonly used.

4) Choice of the Margin Value in the Triplet Loss: As mentioned in Sec. V-A.2, we set a margin parameter Δ to 0.3, which is a common choice of this parameter for the fine-grained retrieval tasks. Table III shows that both increasing and decreasing its values leads to an inferior performance.

5) Number of Used Views to Represent 3D Shape: Following MVCNN [15], we used $V = 24$ views to represent 3D shapes. Table IV shows that the more views is used the higher the retrieval accuracy is. Thus, using 36 views further improves the performance of our method.

E. Additional Experiments

1) Training on Synthetic Sketches: There is a large domain gap between synthetic edges and real free-hand sketches – sketches are highly abstract, and are subject to different styles and drawing skills, whereas synthetic edges are truthful 2D representations of 3D geometry. As can be seen in Table V, model (Synthetic) trained using synthetic sketches (canny edge detection [67] on depth maps) yields far worse performance when tested using real sketches. This again confirms the large domain gap between sketch and synthetic edges.

TABLE V

COMPARISON WITH THE MODEL TRAINED ON SYNTHETIC SKETCHES (SYNTHETIC) AND THE RETRIEVAL MODEL, BASED ON THE PRIOR SKETCH TO 3D SHAPE RECONSTRUCTION (GENERATION)

Method	Chair Dataset		Lamp Dataset	
	acc.@1	acc.@5	acc.@1	acc.@5
Synthetic[67]	32.01	65.17	33.03	64.56
Generation[56]	0.50	4.15	0.90	3.90
Our model	**56.72**	**87.06**	**57.66**	**87.39**

Fig. 7. 3D reconstruction results using [56].

2) 3D Shape Generation vs. Retrieval: Sketch-based 3D reconstruction is much more challenging than retrieval, and existing works on image-based 3D reconstruction cannot be easily adapted. In particular, (i) abstraction of hand-drawn sketch causes serious misalignment between sketch and 3D shape, and (ii) a lack of color and texture makes foreground/background hard to distinguish. We fine-tuned one of the state-of-art sketch reconstruction models [56] on our dataset, those generation results are shown in Fig. 7. As can be seen, the current results have low quality. Yet, we evaluate the potential of such reconstruction results for retrieval. We adopt a triplet network, where PointNet [19] is used as a 3D shapes encoder. We use a 3D shape generated from the input sketch as an anchor, a ground-truth 3D model as a positive sample, and the other 3D models as negative samples. As can be seen in Table I, model (Generation) performs much worse than ours, due to lack of fine-grained details in the reconstruction results.

VI. CONCLUSION

We introduced the novel task of fine-grained instance-level SBSR (FG-SBSR). This task is more challenging than the well-studied category-level SBSR task, but is also more useful in real-world applications. To enable FG-SBSR study, we contributed two large-scale datasets. A deep joint embedding learning based model is introduced with a novel cross-modal view attention module. Extensive experiments show that the proposed model is superior to a number of baselines and the introduced view attention module is the key reason for the performance improvement. We hope that by contributing the FG-SBSR datasets and the proposed model as a strong baseline more researchers will start investigating this challenging, yet practical problem.

REFERENCES

[1] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C. C. Loy, "Sketch me that shoe," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 799–807.

[2] Y. Xia, S. Wang, Y. Li, L. You, X. Yang, and J. J. Zhang, "Fine-grained color sketch-based image retrieval," in *Proc. Comput. Graph. Int. Conf.*, 2019, pp. 424–430.

[3] J. Song, Y.-Z. Song, T. Xiang, T. Hospedales, and X. Ruan, "Deep multi-task attribute-driven ranking for fine-grained sketch-based image retrieval," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1–11.

[4] J. Xue, Y. Zhou, Z. Jiang, Y. Xie, and X. Li, "A multiple triplet-ranking model for fine-grained sketch-based image retrieval," in *Proc. IEEE Vis. Commun. Image Process.*, Dec. 2019, pp. 1–4.

[5] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5551–5560.

[6] Y. Wang, F. Huang, Y. Zhang, R. Feng, T. Zhang, and W. Fan, "Deep cascaded cross-modal correlation learning for fine-grained sketch-based image retrieval," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107148.

[7] S. M. Yoon, M. Scherer, T. Schreck, and A. Kuijper, "Sketch-based 3D model retrieval using diffusion tensor fields of suggestive contours," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 193–200.

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

[9] B. Li *et al.*, "SHREC'14 track: Extended large scale sketch-based 3D shape retrieval," in *Proc. Eurograph. Workshop 3D Object Retr.*, 2014, pp. 121–130.

[10] F. Wang, L. Kang, and Y. Li, "Sketch-based 3D shape retrieval using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1875–1883.

[11] F. Zhu, J. Xie, and Y. Fang, "Learning cross-domain neural networks for sketch-based 3D shape retrieval," in *Proc. Assoc. Adv. Artif. Intell.*, 2016, pp. 3683–3689.

[12] G. Dai, J. Xie, F. Zhu, and Y. Fang, "Deep correlated metric learning for sketch-based 3D shape retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 3374–3386.

[13] J. Xie, G. Dai, F. Zhu, and Y. Fang, "Learning barycentric representations of 3D shapes for sketch-based 3D shape retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5068–5076.

[14] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, "Triplet-center loss for multi-view 3D object retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1945–1954.

[15] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 945–953.

[16] B. Li *et al.*, "A comparison of methods for sketch-based 3D shape retrieval," *Comput. Vis. Image Understand.*, vol. 119, pp. 57–80, Feb. 2014.

[17] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: Learning to retrieve badly drawn bunnies," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–12, Jul. 2016.

[18] T. Yu, J. Meng, and J. Yuan, "Multi-view harmonized bilinear network for 3D object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 186–194.

[19] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 652–660.

[20] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, Jun. 2017, pp. 5099–5108.

[21] J. Li, B. M. Chen, and G. H. Lee, "SO-Net: Self-organizing network for point cloud analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9397–9406.

[22] C. Wang, B. Samari, and K. Siddiqi, "Local spectral graph convolution for point set feature learning," 2018, *arXiv:1803.05827*. [Online]. Available: http://arxiv.org/abs/1803.05827

[23] Y. Li, S. Pirk, H. Su, C. R. Qi, and L. J. Guibas, "FPNN: Field probing neural networks for 3D data," in *Proc. Adv. Neural Inf. Process. Syst.*, May 2016, pp. 307–315.

[24] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas, "Volumetric and multi-view CNNs for object classification on 3D data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5648–5656.

[25] Z. Wu *et al.*, "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1912–1920.

[26] D. Maturana and S. Scherer, "VoxNet: A 3D convolutional neural network for real-time object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 922–928.

[27] A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "Generative and discriminative voxel modeling with convolutional neural networks," 2016, *arXiv:1608.04236*. [Online]. Available: http://arxiv.org/abs/1608.04236

[28] T. S. Cohen, M. Geiger, J. Koehler, and M. Welling, "Spherical CNNs," 2018, *arXiv:1801.10130*. [Online]. Available: http://arxiv.org/abs/1801.10130

[29] C. Esteves, C. Allen-Blanchette, A. Makadia, and K. Daniilidis, "Learning SO(3) equivariant representations with spherical CNNs," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 52–68.

[30] S. Bai, X. Bai, Z. Zhou, Z. Zhang, and L. J. Latecki, "GIFT: A real-time and scalable 3D shape search engine," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5023–5032.

[31] E. Johns, S. Leutenegger, and A. J. Davison, "Pairwise decomposition of image sequences for active multi-view recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3813–3822.

[32] L. Li *et al.*, "Model-driven sketch reconstruction with structure-oriented retrieval," in *Proc. SIGGRAPH ASIA Tech. Briefs*, 2016, pp. 1–4.

[33] A. Qi, Y.-Z. Song, and T. Xiang, "Semantic embedding for sketch-based 3D shape retrieval," in *Proc. Brit. Mach. Vis. Conf.*, 2018, p. 8.

[34] P. Xu *et al.*, "Cross-modal subspace learning for fine-grained sketch-based image retrieval," *Neurocomputing*, vol. 278, pp. 75–86, Feb. 2018.

[35] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, May 2016, pp. 289–297.

[36] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 451–466.

[37] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 21–29.

[38] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," *Comput. Sci.*, vol. 2015, pp. 2048–2057, Feb. 2015.

[39] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4651–4659.

[40] J. Mun, M. Cho, and B. Han, "Text-guided attention model for image captioning," in *Proc. Assoc. Adv. Artif. Intell.*, 2017, pp. 4233–4239.

[41] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[42] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3588–3597.

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

[44] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[45] N. Parmar *et al.*, "Image transformer," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4055–4064.

[46] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," 2019, *arXiv:1904.10509*. [Online]. Available: http://arxiv.org/abs/1904.10509

[47] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*. [Online]. Available: http://arxiv.org/abs/2010.11929

[48] B. Wu *et al.*, "Visual transformers: Token-based image representation and processing for computer vision," 2020, *arXiv:2006.03677*. [Online]. Available: http://arxiv.org/abs/2006.03677

[49] K. Han *et al.*, "A survey on vision transformer," 2020, *arXiv:2012.12556*. [Online]. Available: http://arxiv.org/abs/2012.12556

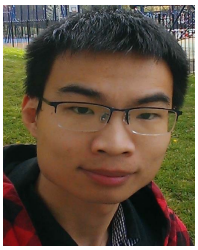
[50] J. Liang, L. Jiang, L. Cao, L.-J. Li, and A. G. Hauptmann, "Focal visual-text attention for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jan. 2018, pp. 6135–6143.

[51] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "An evaluation of descriptors for large-scale image retrieval from sketched feature lines," *Comput. Graph.*, vol. 34, no. 5, pp. 482–498, 2010.

[52] S. Antol, C. L. Zitnick, and D. Parikh, "Zero-shot learning via visual abstraction," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 401–416.
[53] K. Eissen and R. Steur, "Sketching: The basics; the prequel to sketching: Drawing techniques for product designers," BIS, Amsterdam, The Netherlands, Tech. Rep. OCLC 756275344, 2011.
[54] S.-H. Bae, R. Balakrishnan, and K. Singh, "ILoveSketch: As-natural-as-possible sketching system for creating 3D curve models," in *Proc. 21st Annu. ACM Symp. User Interface Softw. Technol. (UIST)*, 2008, pp. 151–160.
[55] B. Xu, W. Chang, A. Sheffer, A. Bousseau, J. McCrae, and K. Singh, "True2Form: 3D curve networks from 2D sketches via selective regularization," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 1–13, Jul. 2014.
[56] Z. Lun, M. Gadelha, E. Kalogerakis, S. Maji, and R. Wang, "3D shape reconstruction from sketches via multi-view convolutional networks," in *Proc. 3D Vis.*, 2017, pp. 67–77.
[57] C. Li, H. Pan, Y. Liu, X. Tong, A. Sheffer, and W. Wang, "BendSketch: Modeling freeform surfaces through 2D sketching," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Jul. 2017.
[58] Y. Gryaditskaya, M. Sypesteyn, J. W. Hoftijzer, S. C. Pont, F. Durand, and A. Bousseau, "OpenSketch: A richly-annotated dataset of product design sketches," *ACM Trans. Graph.*, vol. 38, no. 6, pp. 1–232, 2019.
[59] A. X. Chang *et al.*, "ShapeNet: An information-rich 3D model repository," 2015, *arXiv:1512.03012*. [Online]. Available: http://arxiv.org/abs/1512.03012
[60] J. P. McIntire, P. R. Havig, and E. E. Geiselman, "Stereoscopic 3D displays and human performance: A comprehensive review," *Displays*, vol. 35, no. 1, pp. 18–26, 2014.
[61] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–10, 2012.
[62] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[63] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3733–3742.
[64] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[65] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[66] X. He, T. Huang, S. Bai, and X. Bai, "View N-gram network for 3D object retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 7515–7524.
[67] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

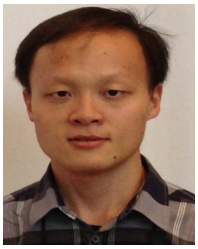
Anran Qi is currently pursuing the Ph.D. degree with the SketchX Lab, Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey. Her sketch focus on sketch-oriented or -aided 3D shaped research topic, including sketch-based 3D shape retrieval and 3D shape reconstruction.

Yulia Gryaditskaya received the master's degree from Lomonosov Moscow State University, Moscow, Russia, and the Ph.D. degree in computer vision and graphics from the Max Planck Institute for Informatics, Saarbruecken, Germany, in 2017. She is currently a Senior Research Fellow in computer vision and machine learning with the Centre for Vision Speech and Signal Processing (CVSSP). Previously, she was a Postdoctoral Researcher with Inria, Sophia Antipolis, France.



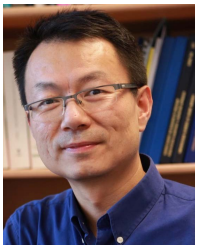
Jifei Song is currently pursuing the Ph.D. degree with the SketchX Lab, Vision Group, Queen Mary University of London. He is interested in sketch-based image retrieval using deep learning. He is developing framework learning cross-domain representation for sketch and photo modality, and then makes fine-grained retrieval based on the learned representation.

Yongxin Yang received the Ph.D. degree from the Queen Mary University of London. He is currently a Lecturer with the University of Surrey. His research is in the area of multi-task learning, transfer learning, and meta-learning. He has broad interests in applications of machine learning, such as computer vision, medical informatics, and finance.



Yonggang Qi (Member, IEEE) received the Ph.D. degree from the Pattern Recognition and Intelligent System (PRIS) Laboratory, BUPT, and the joint Ph.D. degree from the SketchX Lab (headed by Dr. Yi-Zhe Song), Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey. He is currently an Assistant Professor with BUPT. His research interests include perceptual contour grouping and sketch-based machine vision algorithms and applications.

Timothy M. Hospedales (Member, IEEE) is currently a Professor with the School of Informatics, The University of Edinburgh. He is also a Principal Researcher with the Samsung AI Centre, Cambridge, where he leads the Machine Learning Group. He has broad interests in machine learning and computer vision. He is also an Associate Editor of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, and serves as the Area Chair for several major events, including ICCV, CVPR, ECCV, AAAI, and ACL, and the Program Chair for BMVC 2018.



Tao Xiang received the Ph.D. degree in electrical and computer engineering from the National University of Singapore in 2002. He is currently a Full Professor and the Distinguished Chair of the Department of Electrical and Electronic Engineering, University of Surrey. He is also the Research Scientist Manager with Facebook AI. He has published over 200 papers in international journals and conferences with over 21 k citations. His research interests include computer vision and machine learning. He served as the Area Chair for CVPR, ECCV, and ICCV, and serves as an Associate Editor for the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE.

Yi-Zhe Song (Senior Member, IEEE) received the degree (Hons.) from the University of Bath in 2003, the M.Sc. degree from the University of Cambridge in 2004, and the Ph.D. degree in computer vision and machine learning from the University of Bath in 2008. He is currently a Professor of computer vision and machine learning, and the Director of the SketchX Lab, Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey. Previously, he was a Senior Lecturer with the Queen Mary University of London, and a Research and Teaching Fellow with the University of Bath. He is a fellow of the Higher Education Academy and a Full Member of the EPSRC Review College, the U.K.'s main agency for funding research in engineering and the physical sciences. He received the Best Dissertation Award for his M.Sc. degree. He is also the Program Chair for the British Machine Vision Conference (BMVC) in 2021, and regularly serves as the Area Chair (AC) for flagship computer vision and machine learning conferences, most recently at ICCV 2021, and CVPR 2022.