Fine-Grained Sketch-Based 3D Shape Retrieval with Cross-Modal View Attention

Anran Qi, Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales

***Abstract*—We study, for the first time, the problem of fine-grained sketch-based 3D shape retrieval (FG-SBSR), where free-hand sketches are used as input for *instance-level* retrieval of 3D shapes. 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 therefore two new FG-SBSR datasets, consisting a total of 4,680 sketch-3D pairings from two object categories. Even with the datasets, FG-SBSR is still extremely challenging because the inherent domain gap between 2D sketch and 3D shape is large, and that retrieval needs to be conducted at instance-level as opposed to coarse category-level matching per traditional SBSR. The second contribution of the paper is the first cross-modal deep embedding model for FG-SBSR, that specifically tackles all unique challenges presented by this new problem. The key novelty of the model is a 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 3D printing, architectural modelling and film animation. Research on 3D shape retrieval has particularly flourished in recent years as AR/VR technologies prevail. As an input modality, sketch is advantageous over text to retrieve specific 3D shape instances. This is because it inherently encodes fine-grained shape and appearance information, whilst using text to describe a 3D shape instance is often inaccurate and ambiguous. However, existing sketch-based 3D shape retrieval (SBSR) methods [1], [2], [3], [4], [5], [6], [7], [8] predominantly focus on retrieving 3D shapes of the same category (see Fig. 1 for a comparison between category-level SBSR and instance-level SBSR). This greatly narrows the practical advantage of SBSR since text is often a simpler form of input when only category-level 3D retrieval is concerned. Type 'chair' into a 3D search engine and numerous 3D chairs will be retrieved. In contrast, we argue that it is when retrieving a particular chair within a large gallery of 3D chairs that a sketch-based query is preferred over text.

In this paper, for the first time, the problem of fine-grained instance-level SBSR (FG-SBSR) is studied. One of the key reasons for the lack of previous attempts is the lack of FG-SBSR datasets. All existing SBSR datasets such as the SHREC series of datasets [9], [3] provide only category-level pairings between sketches and 3D shapes. They are often obtained cheaply by merging existing 3D shape datasets with off-the-shelf sketch datasets that share the same categories. Similar practice however cannot be followed here, since we are faced with a much harder problem of collecting instance-level sketch and photo pairings, *i.e.*, each sketch needs to be drawn with a specific 3D shape instance as reference[1].

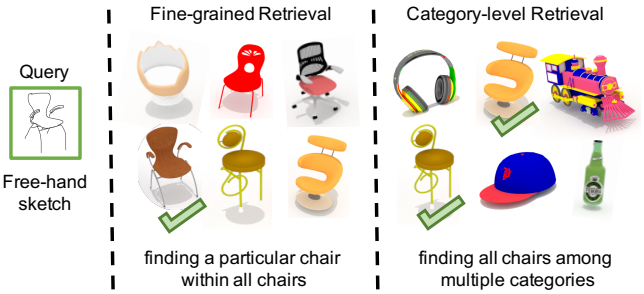


Fig. 1: Comparison between category-level and fine-grained sketch-based 3D shape retrieval.

As the first contribution of this paper, we present two FG-SBSR datasets, consisting of a total of 4,680 sketch-3D pairings (organized as quadruplets) across two categories (chair and lamp). The dataset is built via crowd-sourcing by asking users to finger sketch on a touchscreen device. A key problem that needs to be addressed is that of view ambiguity – people tend to draw sketches from different viewpoints. We address this problem by (i) first conducting a pilot study to determine salient views that ordinary users are accustomed to draw, and (ii) intentionally allowing for more than one sketched view per 3D model in our datasets. 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 greatly stimulate research interest in this new computer vision problem.

Even with the datasets, solving the FG-SBSR problem is far from being straightforward. It not only inherits all challenges brought by traditional category-level 3D shape retrieval, but also poses a few unique ones on its own. The large domain gap between sketch and 3D model data first needs to be addressed. The gap 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

Manuscript received April 19, 2005; revised August 26, 2015.

[1]Similar trend can also be observed when sketch-based image retrieval research shifted from category-level to fine-grained [10], [11].

view gap: sketches are drawn from specific view points, while 3D shape models are entirely view-independent.

As the second contribution, we propose to learn a deep joint embedding space that simultaneously address all the aforementioned gaps. In such a space, the two modalities are aligned, and sketch and 3D shape instances can be compared by simply computing their distance. More specifically, to overcome the dimensionality gap, we follow the common practice [12], [13] 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, a novel cross-modal view attention module is introduced which automatically selects the best combination of matching views for further deep alignment. The joint embedding model is trained with a triplet ranking loss, which had become the most popular choice to tackle the abstraction gap for fine-grained sketch-based image retrieval [10], [11]. With the proposed cross-modal view attention module, a novel triplet sampling strategy is further devised which greatly increases the amount of triplets we can sample, leading to better cross-modal alignment.

Extensive experiments are carried out on the two new datasets. The results show that the proposed model significantly outperforms a number of 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.

II. Related Work

3D Shape Recognition Recent deep recognition for 3D shapes can be broadly categorized into four categories, according to how 3D shapes are represented. Point cloud-based methods [14], [15], [16], [17] directly take point clouds as input while respecting the permutation invariance of points. Volumetric-based methods [18], [19], [20], [21], [22] apply 3D convolutional neural network on voxelized shapes directly. Spherical function-based methods [23], [24] 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 [12], [25], [26], [13] encode 3D models using a collection of their 2D projections. Notable works include [12], which projects 3D objects into multiple views where each view passes through a foreside network in order to learn discriminative view descriptors, followed by view-pooling to combine multiple views. very recently, [13] proposed to use bilinear pooling to effectively aggregate convolutional feature of different views. Of those four categories, view-based methods generally outperform the other three. In this paper, we also adapt a view-based approach to encode 3D shapes, but for the first time study a cross-modal retrieval problem with view attention.

Sketch-based 3D Shape Retrieval Existing sketch-based 3D shape retrieval (SBSR) all focus on category-level, *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 [2], [1], [3] have been followed by the more recent deep learning based models [7], [4], [5], [6], [8], [27]. 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) [12] approach originally designed for 3D shape recognition to project 3D shapes into 2D images of evenly distributed views, with the exception of [27] which models 3D shapes as point clouds and employs PointNet [14], [15] 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 [7], [5], [6], [8] in the use of 2D projections of 3D shape and triplet ranking loss for embedding space learning. However, there are a couple of vital differences: (1) Those category-level SBSR models rely heavily on the category-level labels induced category classification loss [27], which is not available to our FG-SBIR problem whereby we focus on retrieving instances of the same category. (2) Our model is uniquely able to select the optimal projection views for 3D shape feature extraction, and has an effective triplet sampling strategy tailor-made for our view attention module.

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 [10], [28], [29], [30]. 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 [10], [11], whereas such view correspondence has to be separately established in FG-SBSR. As a result, although the models in [10], [28], [29], [30] 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 tedious sketch-drawing process. Existing FG-SBIR datasets [10], [11] thus have moderate sizes with hundreds of sketches per object category – similar to those of our FG-SBSR datasets.

Attention Mechanism Recently, attention modules have been introduced to deep models for addressing a variety of different tasks, including but not limited to visual question answering (VQA) [31], [32], [33], image captioning [34], [35], [36], and object retrieval [29]. 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 [31], [37], which again serves a different purpose (image spatial-sentence word co-

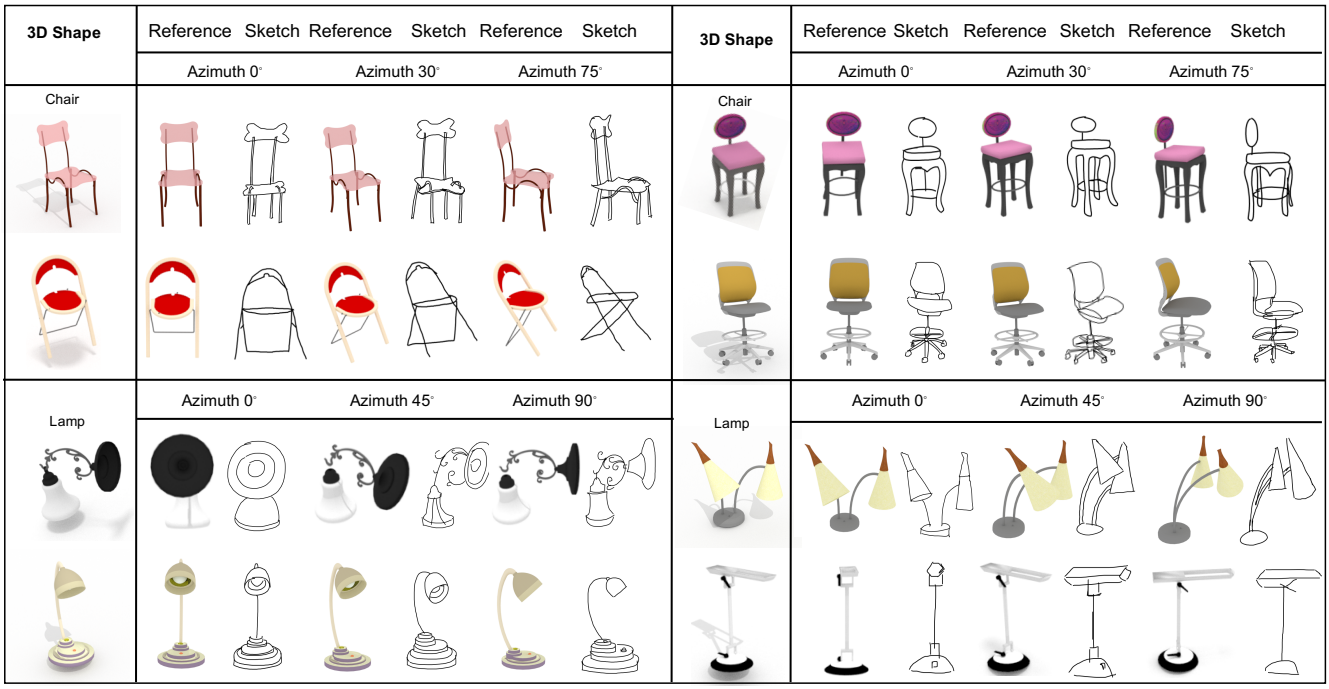


Fig. 2: Examples of the proposed chair and lamp datasets.

attention vs. view attention).

III. Fine-grained Instance-level SBSR Datasets

Pilot Study on Sketch View Ambiguity A key problem facing our sketch collection process is that of view ambiguity, *i.e.*, given a 3D model (shape), which view(s) are humans accustomed to produce a sketch for. For that, we conduct a pilot study where 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 [2]. While rotating, each is asked to choose 3 views per model that they are mostly likely to produce a sketch for. We then aggregate this view selection data (6,000=20×100×3 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.

Dataset Overview We contribute two fine-grained SBSR datasets, one for chairs and the other for lamps[3]. There are 4,680 sketch-3D shape pairs in total (organized as 1,560 quadruplets). Each quadruplet comprises one 3D shape and three free-hand sketches (*i.e.*, they are not drawn by tracing the shape images) drawn from three azimuthal angles (views) respectively, with the 2D projection/rendering of the 3D shape along the corresponding view as reference. The chair dataset has 1,005 sketch-3D shape quadruplets with azimuths $0°$, $30°$, $75°$, while the lamp dataset has 555 sketch-3D shape quadruplets with azimuths $0°$, $45°$, $90°$; Fig. 3 shows some examples. In each column, we display one specific view from various types of chair and lamp, indicating the exhaustiveness as well as highlighting the appearance difference/domain gap between 3D shapes and realistic free-hand sketches. The detailed data collection process is described below.

3D Shape Category Selection There are plenty of 3D shapes of different categories from existing 3D shape recognition datasets. The 3D shapes used in our datasets are selected from the largest 3D shape dataset ShapeNet [38]. Among the 270 object categories, chair and lamp 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. 3). 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 FG-SBSR.

3D Shape Instance Selection For each category, we manually select 3D shape instances to be used in our datasets. Inspired by [39], [40], the following criteria are used for instance selection. (i) **Representative**: There are many subcategories of chairs and lamps in ShapeNet (*e.g.*, armchair, lounge chair, Windsor chair for chairs, and floor lamp, table lamp for lamps). We make sure that representative instances from each subcategory are chosen. (ii) **Distinctness**: The selected instances in each subcategory need to be visually distinct so that it is possible that their differences can be visually depicted by sketches. (iii) **View-sensitive**: 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. (iv) **Sketchability**: The 3D shape should be easy to sketch. The free-hand sketches would

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

[3]The datasets and code of the proposed model will be made public.

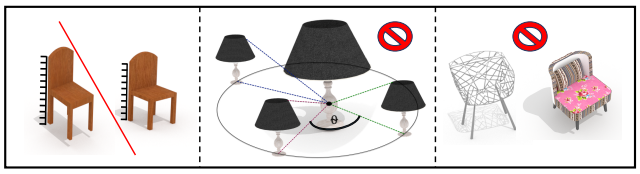


Fig. 3: Examples illustrating our 3D shape instance selection criteria. See text for details.

be 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 respectively.

Sketch Collection In this step, the rendered images of each 3D shape from the selected azimuthal angles are used as references to collect sketches. 30 volunteers are recruited to sketch the rendered images. Concretely, we show one chair/lamp 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 their fingers on a tablet/phone. Two sketches for each image (projection of the 3D shape) are drawn by different volunteers. After finishing collecting all sketches, for quality control purposes, three volunteers vote to select the best sketch out of the two. Note that none of the volunteers has had any art training, and is thus representing the general population who might use the fine-grained SBSR system. As a result, the collected sketches are nowhere near perfect (*e.g.*, lacking detail, and distorted strokes, see Fig. 3), making subsequent fine-grained SBSR task challenging. It is also noted that even when a rendered projection image is of a certain view, the corresponding sketch's view could deviate, as expected from amateur drawers.

IV. Methodology

A. Problem Definition and Model Overview

The sketches in a FG-SBSR dataset are denoted as $\{S_i\}_{i=1}^{I} \in \mathcal{S}$. Each sketch depicts an object instance whose identity is indicated as i and there are I identities in total in the dataset. Note that since each 3D shape/object identity has 3 sketches in $U = 3$ views, we further denote $S_i = \{s_i^{(u)}\}_{u=1}^{U}$ where $s_i^{(u)}$ is the uth view of the ith sketch identity. The 3D shapes are denoted as $\{X_j\}_{j=1}^{J} \in \mathcal{X}$ where j is the identity index. To reduce the domain gap, each 3D shape X_j can be represented by an arbitrary number of rendered 2D views for matching with the 2D sketches, denoted as $X_j = \{x_j^{(v)}\}_{v=1}^{V}$, where $x_j^{(v)}$ is the vth view of the jth 3D shape and V is the number of render views. Thus, the problem of fine-grained instance-level SBSR can be defined as: given a query sketch $s_i^{(u)}$, compute the similarity score between it and X_j in a gallery set of 3D shapes and use the score to rank the whole gallery set so that the true match (same instance identity) for the query sketch is ranked at the top.

To address this problem, We propose a deep multi-modal joint embedding model for cross-domain retrieval. In the learned embedding space, the similarity between a sketch and a 3D shape can be computed simply as the Euclidean distance between the two corresponding feature vectors. Our model contains multiple branches for sketch and 3D shape (projected into different views). The subnetworks in different branches have tied parameters so that the whole model is a Siamese network. Importantly a cross-modal view attention module is introduced to use a query sketch to automatically determine how the projections of different views are fused to form the final representation of the 3D shape in the joint embedding space. The model is trained with a triplet ranking loss formulation with a specifically designed triplet sampling strategy. A schematic illustration of the model can be seen in Fig. 4.

B. Learning Joint Embedding

Siamese networks have been shown to be effective in instance-level sketch-based image retrieval [10], and we adopt a similar architecture in our model to learn a joint embedding space for the sketch and 3D shape modalities. For the backbone network that extracts the feature for both sketch and 3D shape (projections), we employ VGG-16 (config. E) [41], and remove the final class label prediction layer. For each branch, the output of final layer (without ReLU function) is used as the deep feature.

The sketch branch is set as the anchor branch, where the input is a single view sketch $s_i^{(u)}$. The other two branches, positive and negative 3D shape branches encode the information of 3D shapes. Specifically, our 3D shape representation starts from multiple projections of 3D shapes $X_j = \{x_j^{(v)}\}_{v=1}^{V}$. After 2D projections of 3D shapes are obtained following the multi-view CNN [12], each projected 2D image is fed into the first part of the network $F(\cdot)$ (13 convolutional layers) separately. View features are then aggregated to one feature vector via cross-modal view attention, to be detailed in Sec. IV-C. Eventually the aggregated 3D shape features pass through the second part of the network $G(\cdot)$ (2 fully connected layers). All three branches (*i.e.*, $G(F(\cdot))$) share the parameters, hence the whole network is Siamese. To train the network, we form sketch, positive and negative 3D shape triplets and use a triplet ranking loss. The triplet construction strategy is detailed in Sec. IV-D.

C. Cross-Modal View Attention

As shown in Fig. 4, even though all V views of the positive 3D shapes contain the same object instance, their visual appearance can be drastically different from that of the anchor sketch. It is thus important to dynamically (anchor sketch-specifically) determine the relevance of each view before the features extracted from each view fused to represent the 3D shape. To this end, we propose a cross-modal view attention module, which generates a view selection vector used for guiding the fusion of the V views.

Concretely, given an anchor sketch $s_i^{(u)}$, we denote the sketch feature vector as $f_i^{(u)} \in \mathbb{R}^{4096}$ which is the final output of VGG-16 in the sketch branch, where i is the sketch identity index and u is the sketch view index. We can then obtain the

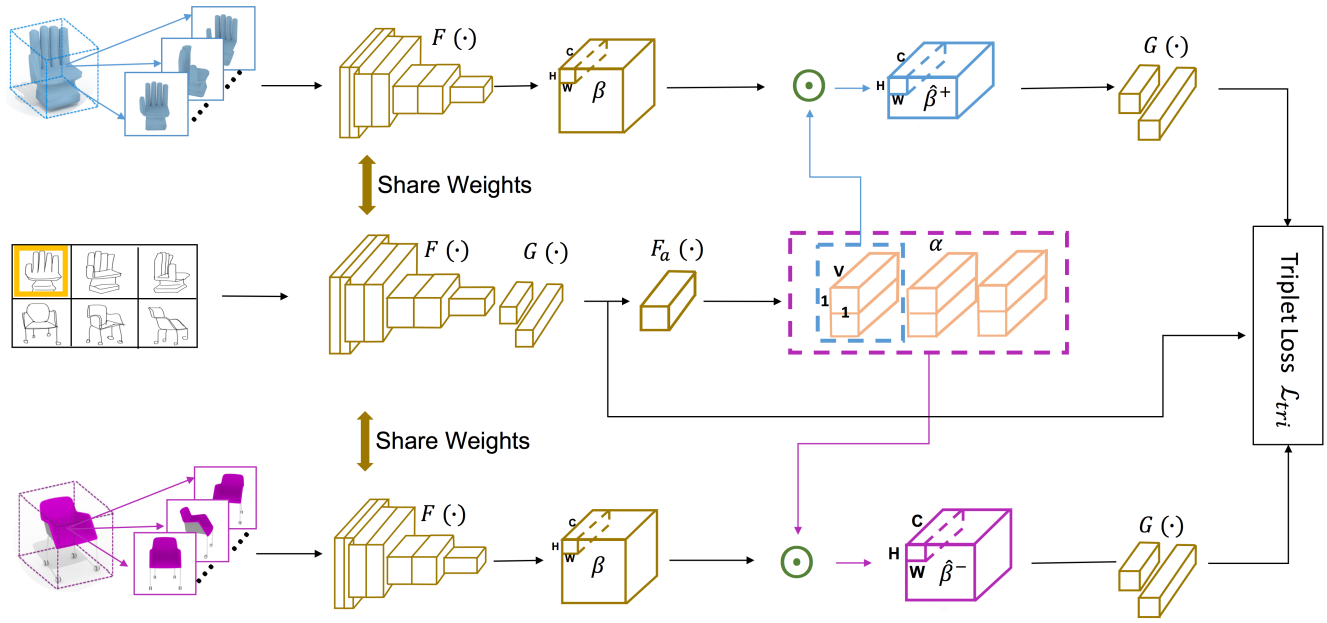


Fig. 4: An illustration of the proposed fine-grained SBSR with cross-modal view attention. Given an anchor sketch in the yellow box, the positive 3D shape is in blue boxes and the negative 3D shape is in magenta boxes. $F(\cdot)$ is the first part of the network and $G(\cdot)$ is the second part of the network. $F_a(\cdot)$ is the view attention module where α is the output view attention vector. The vectors in blue/magenta dashed boxes are used to generate positive/negative 3D feature $\hat{\beta}^+/\hat{\beta}^-$. $\odot$ represents dot product.

attention vector $g_i^{(u)} \in \mathbb{R}^V$ by feeding the feature vector into the attention module:

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

where $F_a(\cdot)$ is the view attention function learned by the view attention module and 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 ($4096 \rightarrow V$). Specifically, the vth element in the attention vector $g_i^{(u)}$ represents its view attention score corresponding to view v in 2D projections of 3D shape, denoted as $g_{i,u,v}$.

We then design a specific normalization scheme to refine the attention vector. The final view attention score, $\alpha_{i,u,v}$ can be calculated following:

$$\begin{aligned}\alpha_{i,u,v} &= \ell_2\_\text{softmax}(g_{i,u,v}; \tau) \\ &= \frac{\exp(\tau^{-2} g_{i,u,v} \|g_{i,u,\cdot}\|_2^{-1})}{\sum_{v=1}^{V} \exp(\tau^{-2} g_{i,u,v} \|g_{i,u,\cdot}\|_2^{-1})}\end{aligned} \tag{2}$$

Note that the original attention score $g_{i,u,v}$ is normalized twice. First, it is normalized by an ℓ_2 norm: this helps the numerical stability as the dimension changes significantly ($4096 \rightarrow V$) which leads to very large logit values. Second, it is normalized by a softmax function, by which a valid probability vector is produced. Note that, in the softmax function, we introduce a trainable temperature variable, τ, to further help the training. τ is designed to adjust the logits again. The motivation behind this design is that ℓ_2 norm may lead to overly flat probability meaning all views will be selected with a similar weight, and τ can help sharpen it to focus on a small number of views.

The fusion of the 2D projections of the 3D shape is then computed by view-wise dot product of the view attention vector to the 3D shape embedding as follows:

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

where $\beta_j^{(v)}$ is the vth view projection feature of the jth 3D shape extracted using $F(\cdot)$, and $\hat{\beta}_{j,i,u}$ is the attended view feature of jth 3D shape when a sketch indexed by i (identity) and u (view) is used to produce the attention vector. I.e., Eq. 3 produces a re-weighted sum (the weight is attention) of V projected 3D shape images' convolutional features. After that, we feed the attended feature into the second part of the network ($G(\cdot)$) to generate the final feature, which is of the same dimensionality as the sketch branch, *i.e.*, 4096. That is, after $G(\cdot)$, both 3D shape and sketch are represented in the same joint embedding space.

D. Triplet Sampling Strategy

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 identities (*i.e.*, $B \cdot U$ sketches and B 3D shape with $B \cdot V$ rendered 2D views). The intuition is that we assume the view attention vector is only relative to sketch view regardless of sketch identity. In other words, the second index of 3D view attended feature $\hat{\beta}_{j,i,u}$, *i.e.*, the sketch identity i indicating which sketch delivers the attention, is a dummy variable. Based on this

assumption, given an anchor sketch $s_i^{(u)}$, the positive and negative condition for a view attended 3D feature $\hat{\beta}_{j,i',u'}$ in the triplet tuple is determined only by 3D identity j and sketch view u', regardless of the sketch identity i'.

Formally, given an anchor sketch $s_i^{(u)}$, the positive 3D feature set $\mathcal{P}$ is defined as:

$$\mathcal{P} = \left\{ G(\hat{\beta}^{+}_{j,i',u'}) \right\} = \left\{ G\Big(\sum_{v=1}^{V} \alpha_{i',u',v} \cdot \beta_j^{(v)} \Big), \; j = i, u' = u, i' \in [1, 2, ...B] \right\} \tag{4}$$

This set $\mathcal{P}$ is based on the features of projections $\{\beta_j^{(v)}\}_{v=1}^{V}$ from the 3D identity j which is in the same identity with anchor sketch identity i. Then each projection feature $\beta_j^{(v)}$ is further attended by view attention score $a_{i',u',v}$ which can be produced by any sketches that are from the same view with the anchor sketch, *i.e.*, $u' = u$. As a result, for each anchor sketch $s_i^{(u)}$ in one mini-batch, there are B positive samples augmented by B view attention vectors, where B is the number of sketch identities in the batch.

The negative 3D feature set $\mathcal{N}$ is formed as:

$$\mathcal{N} = \left\{ G(\hat{\beta}^{-}_{j,i',u'}) \right\} = \left\{ G\Big(\sum_{v=1}^{V} \alpha_{i',u',v} \cdot \beta_j^{(v)} \Big), \; j \neq i, u' \in [1, 2, ..., U], i' \in [1, 2, ..., B] \right\}. \tag{5}$$

In contrast to the positive set $\mathcal{P}$, the negative set $\mathcal{N}$ is based on the features of projections $\{\beta_j^{(v)}\}_{v=1}^{V}$ from the different identities with the anchor sketch identity i. Since the sketch and 3D shape identity are different fundamentally, no matter what view attention vectors are used to attend the projections features, the attended feature should be in the negative set. In other words, all the view attention vectors in one batch can be used to attend the negative projection features, which results in $(B-1) \cdot B \cdot U$ negative samples.

We can now define our view-aware triplet loss as:

$$\mathcal{L}_{tri} = \sum_{i}^{B} \sum_{u}^{U} \sum_{p}^{|\mathcal{P}|} \sum_{n}^{|\mathcal{N}|} \max \Big(0, \Delta + D\big(f_i^{(u)}, h_p^{+}\big) - D\big(f_i^{(u)}, h_n^{-}\big) \Big), \tag{6}$$

where h_p^{+} and h_n^{-} denotes the pth/nth 3D shape feature 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.*, $\|G(\cdot)\| = 1$.

The overall pipeline of training the proposed FG-SBSR model is summarized in Alg. 1. We show in our experiments (see Sec. V-C) that the proposed view-identity hybrid sampling strategy is much more effective than the standard sampling strategy whereby the positive 3D shapes are selected only according to identity.

Algorithm 1 Training of the proposed FG-SBSR model.

Input:
Sketch, S, 3D Shape, X in a sampled batch of size B;
for $t = 1$ to max-iteration **do**
  Sample a batch data, $\mathcal{S} \in S$, $\mathcal{X} \in X$
  for $s_i^{(u)} \in \mathcal{S}, x_j^{(v)} \in \mathcal{X}, i, j \in [1, 2, \ldots, B], u \in [1, 2, \ldots, U], v \in [1, 2, \ldots, V]$ **do**
    $f_i^{(u)} = G(F(s_i^{(u)}))$ and $\beta_j^{(v)} = F(x_j^{(v)})$;
    $g_i^{(u)} = F_a(f_i^{(u)}; W_a)$
    $\alpha_{i,u,v} = \ell_2\_\text{softmax}(g_{i,u,v}; \tau)$
    $\hat{\beta}_{j,i,u} = \sum_{v=1}^{V} \alpha_{i,u,v} \cdot \beta_j^{(v)}$
  end for
  for $s_i^{(u)} \in \mathcal{S}$ **do**
    Pos. Set $\mathcal{P}$: $\left\{ G(\hat{\beta}^{+}_{j,i',u'}) \right\}, |\mathcal{P}| = B$;
    Neg. Set $\mathcal{N}$: $\left\{ G(\hat{\beta}^{-}_{j,i',u'}) \right\}, |\mathcal{N}| = (B-1) \cdot B \cdot U$;
  end for
  Optimize $\mathcal{L}_{tri}$;
end for

V. Experiments

A. Experiment Settings

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. Following [12], we put the centroid of the shape at the origin of the spherical coordinate system and translate camera uniformly so that $V = 24$ view projections are rendered with model fitted within frame, though our model is not constrained to these 24 consecutive views. We resize all sketches/3D views to the same size of 224×224.

Implementation Details The model is implemented on Tensorflow. The initial learning rate is set to 0.0001. And the batch size is 3, which means that each batch contains 3 sketch identities, each containing 3 sketch views, and 3×24 2D view projections of the 3 corresponding 3D shape identities. The margin Δ in the triplet loss is 0.3 (see Eq. 6). The model is pretrained on ImageNet [42] then trained for 50 epochs for each dataset. The trainable temperature variable, τ (see Eq. 2) is initialized to 2.0.

Evaluation Metrics During testing, given a query sketch, the gallery of 3D shapes are ranked based on the distance to the query sketch in the joint embedding space. For our task of fine-grained instance-level 3D shape retrieval, the cumulative matching accuracy acc@K is used for evaluation, which is calculated as the percentage of query sketches whose true-match 3D shapes are ranked in the top K.

B. Baselines

As discussed in Sec. II, there are no existing FG-SBSR models, as the problem is studied for the first time in this paper. Furthermore, neither the existing category-level SBSR models, nor the FG-SBIR models can be used directly for comparison: the former need category labels and the latter do not handle views. Therefore, the baselines compared here

	Chair Dataset		Lamp Dataset	
Method	acc.@1	acc.@5	acc.@1	acc.@5
SBSVSR	0.4494	0.7794	0.4805	0.7987
FG-T-M	0.4760	0.8126	0.4925	0.8348
FG-T-A-M	0.4710	0.7910	0.4895	0.8168
FG-T-P	0.0050	0.0448	0.0090	0.0360
FG-T-S	0.1177	0.4013	0.1261	0.3844
Synthetic	0.3201	0.6517	0.3303	0.6456
Our model	**0.5672**	**0.8706**	**0.5766**	**0.8739**

TABLE I: Comparative results against baselines.

	Chair Dataset		Lamp Dataset	
Method	acc.@1	acc.@5	acc.@1	acc.@5
w/o τ	0.5108	0.8391	0.5315	0.8468
VIIST	0.4975	0.8242	0.5345	0.8318
VSIST	0.5589	0.8507	0.5465	0.8498
Heterogeneous	0.2670	0.7380	0.2883	0.7147
Our model	**0.5672**	**0.8706**	**0.5766**	**0.8739**

TABLE II: Contributions of the different components

are designed by us 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 point cloud CNN [14] and spherical CNN [24].

Sketch-based Single View 3D Shape Retrieval (SBSVSR) This model essentially follows the FG-SBIR model in [10] but with the same branch subnet architecture for fair comparison. To avoid dealing with the view problem, each 3D shape is rendered to 3 views in accordance with the three sketch views. To form the triplets, the positive 3D shape will have the same identity and same view as the anchor sketch whilst the negative 3D shape has a different identity. In addition, we jointly (multi-task) train a sketch view classifier to classify each sketch into the three views. During testing for each query sketch, we predict the view class first, then select the projected 3D shape for that view to extract features in the joint embedding space for matching.

Fine-grained Triplet based on MVCNN [12] (FG-T-M) In this model, following most category-level SBSR models [7], [5], [6], [8], we first project the 3D shapes into 24 views for feature extraction as in MVCNN [12], in each view. 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 still resembles that of [10] and a triplet loss is also adopted as supervision. In summary, the main difference between this model and ours is the cross-modal view attention module.

Fine-grained Triplet with Spatial Attention (FG-T-A-M) [29] 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 in [29] to FG-T-M. Concretely, two convolutional layers with kernel size 1 are added to the output of the final convolutional+pooling layer of the CNN in each branch and the two attention modules do not share parameters.

Fine-grained Triplet based on Non-projection Based 3D Deep Embeddings Here we benchmark against alternative deep 3D representations, as opposed to 2D projection based. More specifically, PointNet++ [14] and Spherical CNN [24] are used in place of MVCNN on the 3D branch of our network, to form the **FG-T-P** and **FG-T-S** baselines, respectively. Since the 3D shape branch is now view-independent, no view fusion is necessary.

C. Results

Comparisons against Baselines Table I shows the results of our model and the five baselines. The following observations can be made: (1) Our model performs significantly better than all baselines on both datasets. (2) The single view model SBSVSR is clearly inferior to all models that project 3D shapes to 2D views and then fuse them. This is despite the use of ground truth view information of the sketch. (3) Among the baselines, FG-T-M is the most competitive one. But the significant performance gaps (around 9% lower on acc.@1 for both datasets) indicate that performing view selection rather than simply max-pooling the features of different views makes a big 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 the FG-SBSR task. (5) Non-projection based methods are considerably inferior than projection-based methods. In particular, the FG-T-S model captures little fine-grained detail 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 [27]. This result highlights the vital difference between category-level and instance-level SBSR: with category labels, it is most effective to align the two modalities in the semantic (category label) space, which bypasses the view dimensionality mismatch. In contrast, for instance-level retrieval, the view gap cannot be avoided. And our results show that 2D projection of 3D shapes is necessary.

Qualitative Results In Fig. 5, we show some examples of fine-grained SBSR results obtained using our model. The first 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.

Ablation Studies Here we further evaluate the effectiveness of a number of design choices. (i) *Effectiveness of view-identity hybrid triplet sampling strategy:* The triplet sampling strategy described in Sec. IV-D is unique to our model because the 3D shapes are sampled using identity but merged with view-based sketch selection to compute identity-insensitive view attention vectors to greatly increase the number of triplets, leading to more effective model training. To verify this, we compare with another two sampling strategies: a model employing the conventional sampling strategy, termed as view insensitive and identity sensitive triplet sampling or **VIIST** and a model employing a stricter sampling strategy than VIIST, termed as view sensitive and identity sensitive triplet sampling or **VSIST**. Tab. II shows that the proposed sampling strategy alone brings about 7% and 4% increase in acc.@1 respectively for chair and lamp datasets when compared with VIIST,

Fig. 5: Qualitative results. For each query sketch, the top 6 ranked 3D shapes in the gallery are shown in each row. See details in text.

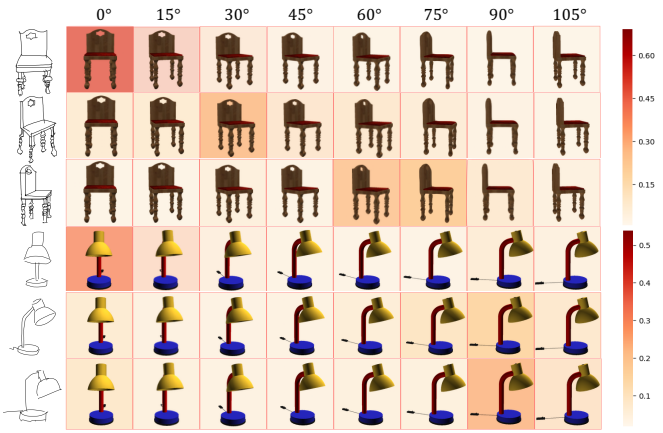


Fig. 6: Examples of attention distribution. We show 8 evenly distributed view angles which the 3D shapes are projected to. Each projection is colour coded with the corresponding attention values indicating which view is attended more for fusion. Warmer colour means higher attention value.

and about 1% and 3% increase in acc.@1 respectively for chair and lamp datasets than VSIST. (ii) *Effectiveness of 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. Tab. II shows that without making the temperature trainable, the model is clearly worse as the view attention is less effective. (iii) *Effectiveness of Siamese network:* We implement a Heterogeneous alternative of our network (termed **Heterogeneous** in Tab. II). It can be seen that the Heterogeneous architecture yields much lower performance, indicating that it suffers severely from overfitting as the number of parameters doubles. This also echoes findings from the related task of SBIR [10], which usually chooses Siamese as well.

How about the synthetic edges rather than sketches? There is dramatically domain gap between synthetic edges and real free-hand sketches. Sketch is first abstracted based on human visual understanding, and deformed in different levels as drawn by free hand in a more flexible way; different annotators might have different sketch drawing styles. Beyond, sketches are collected by asking the volunteers to sketch the 3D shapes according to the memory rather than tracing the edges. Therefore the consistency is largely eliminated by the abstractness and deformation of free-hand sketches. The result shown in Tab. II, Synthetic that sketch are replaced by synthetic data (canny edge detector on depth maps), again demonstrates the satisfying performance of the proposed method, and also give insights that there are large domain gap between real sketch data and synthetic drawings.

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 clearly that our attention module is able to identify the correct view angle of the sketch and gives the biggest weight to corresponding 2D projection. More 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 obviously visually similar but not identical, thus offering some complementary information. It is critical to note that this view attention is dynamic, *i.e.*, instance-dependent: exactly which nearby views should be used and by what weighting factor, depends on the object instance at hand, as well as how well the query sketch is drawn. This is why, as shown in Tab. 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.

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 was introduced with a novel cross-modal view attention module. Extensive experiments have shown the proposed model is superior to a number of baselines and the introduced view attention module is the key reason for the performance improvement.

REFERENCES

[1] S. M. Yoon, M. Scherer, T. Schreck, and A. Kuijper, "Sketch-based 3D model retrieval using diffusion tensor fields of suggestive contours," in *ACMMM*, 2010.

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

[3] B. Li, Y. Lu, C. Li, A. Godil, T. Schreck, M. Aono, M. Burtscher, H. Fu, T. Furuya, H. Johan *et al.*, "Shrec'14 track: Extended large scale sketch-based 3D shape retrieval," in *3DOR*, 2014.

[4] F. Wang, L. Kang, and Y. Li, "Sketch-based 3D shape retrieval using convolutional neural networks," in *CVPR*, 2015.

[5] F. Zhu, J. Xie, and Y. Fang, "Learning cross-domain neural networks for sketch-based 3D shape retrieval." in *AAAI*, 2016.

[6] G. Dai, J. Xie, F. Zhu, and Y. Fang, "Deep correlated metric learning for sketch-based 3D shape retrieval." in *AAAI*, 2017.

[7] J. Xie, G. Dai, F. Zhu, and Y. Fang, "Learning barycentric representations of 3D shapes for sketch-based 3d shape retrieval," in *CVPR*, 2017.

[8] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, "Triplet-center loss for multi-view 3D object retrieval," in *CVPR*, 2018.

[9] B. Li, Y. Lu, A. Godil, T. Schreck, B. Bustos, A. Ferreira, T. Furuya, M. J. Fonseca, H. Johan, T. Matsuda *et al.*, "A comparison of methods for sketch-based 3D shape retrieval," *CVIU*, vol. 119, pp. 57–80, 2014.

[10] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C.-C. Loy, "Sketch me that shoe," in *CVPR*, 2016.

[11] 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, p. 119, 2016.

[12] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *ICCV*, 2015.

[13] T. Yu, J. Meng, and J. Yuan, "Multi-view harmonized bilinear network for 3D object recognition," in *CVPR*, 2018.

[14] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," *CVPR*, 2017.

[15] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *NIPS*, 2017.

[16] J. Li, B. M. Chen, and G. H. Lee, "So-net: Self-organizing network for point cloud analysis," in *CVPR*, 2018.

[17] C. Wang, B. Samari, and K. Siddiqi, "Local spectral graph convolution for point set feature learning," *arXiv preprint arXiv:1803.05827*, 2018.

[18] Y. Li, S. Pirk, H. Su, C. R. Qi, and L. J. Guibas, "Fpnn: Field probing neural networks for 3D data," in *NIPS*, 2016.

[19] 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 *CVPR*, 2016.

[20] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D shapenets: A deep representation for volumetric shapes," in *CVPR*, 2015.

[21] D. Maturana and S. Scherer, "Voxnet: A 3D convolutional neural network for real-time object recognition," in *IROS*, 2015.

[22] A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "Generative and discriminative voxel modeling with convolutional neural networks," *arXiv preprint arXiv:1608.04236*, 2016.

[23] T. S. Cohen, M. Geiger, J. Köhler, and M. Welling, "Spherical cnns," *arXiv preprint arXiv:1801.10130*, 2018.

[24] C. Esteves, C. Allen-Blanchette, A. Makadia, and K. Daniilidis, "Learning so (3) equivariant representations with spherical cnns," *ECCV*, 2018.

[25] S. Bai, X. Bai, Z. Zhou, Z. Zhang, and L. Jan Latecki, "Gift: A real-time and scalable 3D shape search engine," in *CVPR*, 2016.

[26] E. Johns, S. Leutenegger, and A. J. Davison, "Pairwise decomposition of image sequences for active multi-view recognition," in *CVPR*, 2016.

[27] A. Qi, Y. Song, and T. Xiang, "Semantic embedding for sketch-based 3D shape retrieval," in *BMVC*, 2018.

[28] J. Song, Y.-Z. Song, T. Xiang, T. M. Hospedales, and X. Ruan, "Deep multi-task attribute-driven ranking for fine-grained sketch-based image retrieval," in *BMVC*, 2016.

[29] 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 *ICCV*, 2017.

[30] P. Xu, Q. Yin, Y. Huang, Y.-Z. Song, Z. Ma, L. Wang, T. Xiang, W. B. Kleijn, and J. Guo, "Cross-modal subspace learning for fine-grained sketch-based image retrieval," *Neurocomputing*, vol. 278, pp. 75–86, 2018.

[31] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *NIPS*, 2016.

[32] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *ECCV*, 2016.

[33] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *CVPR*, 2016.

[34] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.

[35] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *CVPR*, 2016.

[36] J. Mun, M. Cho, and B. Han, "Text-guided attention model for image captioning." in *AAAI*, 2017.

[37] J. Liang, L. Jiang, L. Cao, L.-J. Li, and A. Hauptmann, "Focal visual-text attention for visual question answering," in *CVPR*, 2018.

[38] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, "Shapenet: An information-rich 3D model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[39] 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.

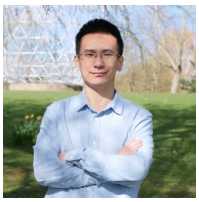
[40] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" *ACM Trans. Graph.*, vol. 31, no. 4, pp. 44–1, 2012.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

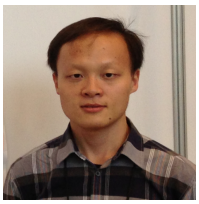
[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

Anran Qi is a PhD student of 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 recustruction.

Jifei Song is a PhD student of 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 make fine-grained retrieval based on the learned representation.

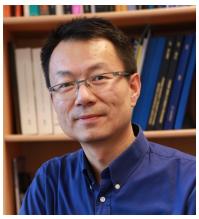


Yongxin Yang is a Lecturer at University of Surrey. He received his Ph.D. from Queen Mary University of London. His research is in the area of multi-task learning, transfer learning, and meta-learning. He has broad interests in applications of machine learning, e.g., computer vision, medical informatics,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

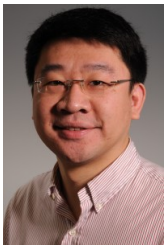


Yonggang Qi is an assistant professor at BUPT. Previously, I was a PhD student at Pattern Recognition and Intelligent Systems (PRIS) laboratory at BUPT. I was also a joint PhD at SketchX lab headed by Dr. Yi-Zhe Song at the Centre for Vision Speech and Signal Processing (CVSSP) in University of Surrey. His research interests include perceptual contour grouping and sketch-based machine vision algorithms and applications.

Timothy M. Hospedales is a Reader within IPAB in the School of Informatics at the University of Edinburgh, and Visiting Reader at Queen Mary University of London. His research focuses on machine learning, particularly life-long transfer and active learning, with both probabilistic and deep learning approaches. He has looked at a variety application areas including computer vision (behaviour understanding, person re-identification, attribute and zero-shot learning), robotics, sensor fusion, novel human-computer interfaces, computational social sciences, theoretical neuroscience and business data analytics.



Tao Xiang is a Professor of Computer Vision and Machine Learning and Distinguished Chair at Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey. His interests include computer vision, image processing, sequential data analysis, pattern recognition, machine learning, and data mining.



Yi-Zhe Song is a Reader, and the founding Director of SketchX Research Lab, Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey. He is interested in all problems associated with understanding human sketches, and how such understanding can be transferred into commercial applications. Prior to Queen Mary, he has worked as Research and Teaching Fellow at University of Bath, where he researched into problems such as perceptual grouping, image segmentation, cross-domain image analysis and non-photorealistic rendering.