Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Category-Level Fine-Grained Sketch-Based 3D Shape Retrieval

Bingrui Wang
Tianjin University
Yuan Zhou
zhouyuan@tju.edu.cn
Tianjin University
Sun-Yuan Kung
Princeton University

Research Article

Keywords: Sketch, 3D shape retrieval, Fine-grained, Category-level, Cross-domain gap

Posted Date: August 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10281515/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Category-Level Fine-Grained Sketch-Based 3D Shape Retrieval

Bingrui Wang[1], Yuan Zhou[1*], Sun-Yuan Kung[2]

[1*]School of Electrical and Information Engineering, Tianjin University, Tianjin, 300072, China.
[2]Department of Electrical Engineering, Princeton University, Princeton, NJ, 08540, USA.

*Corresponding author(s). E-mail(s): zhouyuan@tju.edu.cn;

Abstract

Category-level fine-grained sketch-based three-dimensional (3D) shape retrieval (SBSR) aims to retrieve 3D shapes that belong to the same sub-category as the input sketch. It is challenging due to the abstraction of sketches, the cross-domain discrepancies between two-dimensional (2D) sketches and 3D shapes, and the fact that the sub-categories may be very similar according to the fine-grained setting. No existing category-level SBSR datasets or methods account for fine-grained scenarios. In this paper, we contribute a new fine-grained SBSR dataset containing 7,666 sketches and 20,445 3D shapes grouped into 54 sub-categories in three broad categories of airplanes, cars, and chairs. We also design two fine-grained feature extractors for sketches and 3D shapes and optimize the cross-domain alignment approach to handle the fine-grained retrieval tasks. Extensive experiments on the new fine-grained SBSR dataset and two widely used coarse-grained SBSR benchmarks demonstrate the effectiveness of our proposed method.

Keywords: Sketch, 3D shape retrieval, Fine-grained, Category-level, Cross-domain gap

1 Introduction

3D shapes are widely used in industry and entertainment. With the large amount of 3D shape data being generated, it is becoming a pressing problem to quickly find the 3D shapes people want to employ. Compared with the keyword and 3D example, free-hand sketch is an intuitive and convenient search modality in the day of ubiquitous smart

touch screens. Meanwhile, because of the human abstract mind embedded in sketch, it is extremely adaptable and can give a good indication of the overall structure and key details of the retrieve target. Therefore, SBSR is gradually becoming a popular research topic in computer vision [1–11].

However, existing SBSR methods focus almost exclusively on coarse-grained scenarios. It is difficult for the user to accurately retrieve the desired 3D shapes, since the categories at coarse-grained level may contain many distinct sub-categories, e.g., the category 'chair' has visually very different sub-categories 'cantilever', 'ladder back' and 'scissors' (Fig. 1). Thus, we need to pursue research on the fine-grained SBSR (FG-SBSR) task.

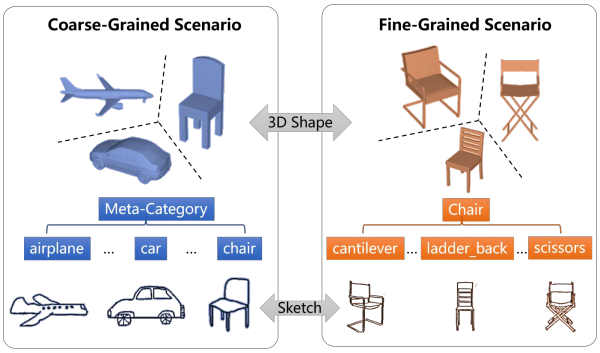


Fig. 1: Examples of SBSR in coarse- and fine- grained scenario. People may find it difficult to retrieve the exact 3D shapes they need at coarse-grained level scenarios.

FG-SBSR is extremely challenging. While fine-grained sketch-based image retrieval has advanced considerably [12, 13], these methods assume a fixed 2D pose alignment between sketches and photos—an assumption invalid for 3D objects observed from arbitrary viewpoints. Thus, 2D-aligned part-attention strategies [14] cannot handle multi-view part duplication and occlusion. Moreover, existing coarse-grained SBSR methods optimize for inter-category separation (e.g., chairs vs. cars), failing to capture the subtle intra-category structural differences required by our fine-grained setting.

Moreover, the scarcity of data is even more critical for FG-SBSR than for common SBSR. Researchers may be eager to study SBSR in fine-grained scenarios, but there is no suitable dataset yet due to the high cost and difficulty of collecting and dividing paired sketch-3D shape samples.

To address the data bottleneck, we propose a new FG-SBSR dataset. Specifically, we invite domain experts to define a unified fine-grained taxonomy, and then organize graduate students under expert supervision to re-classify samples from existing sketch/3D repositories, including Tu-Berlin [15], sketchy [16], pro-chair-3d [17] and FG3D [18]. To further mitigate category imbalance, these students also manually draw a collection of new sketches that strictly adhere to the sub-category specifications. This novel FG-SBSR dataset contains 7,666 sketches and 20,445 3D shapes grouped into 54 sub-categories in three broad categories of airplanes, cars and chairs, and the

sample size of each sub-category is guaranteed to be 12 or more. We hope that research interest in this new computer vision task will be stimulated by making the proposed FG-SBSR dataset publicly available.

To conquer the aforementioned challenges of SBSR in fine-grained scenarios, we design two feature extractors for sketch and 3D shape and optimize the cross-domain alignment approach, as shown in Fig. 2. Specifically, to better capture the detailed features of the 3D shape, we propose the key parts extractor (KPE), a deep network module that extracts and scores the features of a set of key part regions in each view of the 3D shape. Afterward, we fuse the K highest scoring part features, concatenate them with the spatial attention (SA) fine-tuned overall feature of the view, and then aggregate the feature of each view to obtain the final feature of the 3D shape. After training the fine-grained 3D shape feature extractor, we opt for the 3D shape-guided sketch feature learning approach to train the fine-grained sketch feature extractor while performing the cross-domain alignment. With our new designs, the gap between different modalities in the desired fine-grained setting can be effectively bridged.

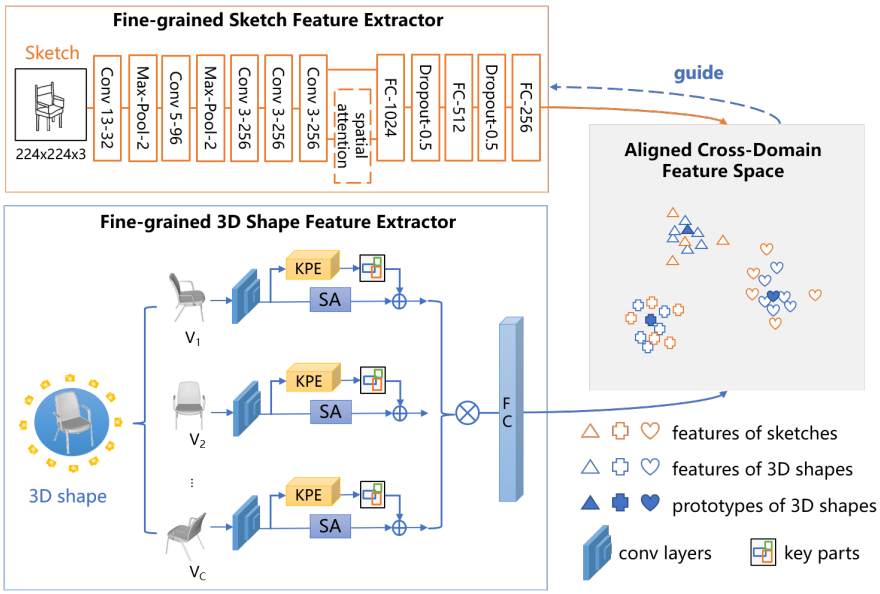


Fig. 2: Illustration of our overall framework.

In summary, the main contributions of this paper are as follows:

1. We take the first step toward category-level FG-SBSR by formally defining this practical yet underexplored task. To support it, we contribute a meticulously annotated dataset with 7,666 sketches and 20,445 3D shapes across 54 fine-grained sub-categories;
2. We propose a cross-view part-aware 3D extractor that explicitly mines and fuses discriminative local details from multi-view projections. To bridge the 2D-3D domain

gap, we further develop a 3D-shape-guided sketch learning strategy, which leverages the richer priors of 3D data to compensate for the high abstraction and data scarcity of sketches;
3. Extensive experiments demonstrate that our approach not only sets a new state-of-the-art on the proposed FG-SBSR benchmark, but also generalizes robustly to conventional coarse-grained SBSR datasets.

2 Related Work

2.1 Sketch-based 3D Shape Retrieval

In the early days, SBSR methods [19–22] had been built upon hand-crafted features, until Wang et al. [1] used a siamese neural network to bring this cross-domain retrieval problem to the deep learning era. The final goal of all deep SBSR methods [1–11] is finding a joint semantic space for the 3D shape and 2D sketch modalities, in which different inner-domain categories will be pushed far away and the same inter-domain categories will be brought closer. After years of development, the SBSR methods have made promising progress and gradually established a basic framework of employing view-based approaches (represented by multi-view (MV) convolutional neural network (CNN) [23]) to extract 3D shape features, employing mature image CNNs (like ResNet-50 [24]) to extract sketch features, and then applying fully connected networks to do the cross-domain alignment. However, these existing category-level SBSR methods only focus on coarse-grained scenarios, where is quite hard to highlight the advantages of sketch-based retrieval.

In this paper, we followed the deep SBSR tradition, but contributed a new dataset as well as a novel method based on data characteristics of 3D shapes and sketches, to extend the category-level SBSR task into the fine-grained scenario.

2.2 Fine-grained Image Retrieval

Due to the large intra-subcategory variance and the small inter-subcategory variance, fine-grained image retrieval (FIR) is a type of hard problem in computer vision and has received long-standing interest. Many researchers [25–31] have invested a great deal of effort in studying sketch-based FIR. They have designed deep networks that could capture subtle differences in images, as well as bridge the semantic gap between sketches and images. These works can largely inspire our FG-SBSR task. However, FG-SBSR is far more challenging since (1) sketch and real-world photo are both 2D images, whereas sketch and 3D shape are dimensional-mismatch, (2) the existing datasets [16, 32] for sketch-based FIR assume a same pose between sketch-photo pairs, yet poses of the 3D shape differ in different views, (3) data of photos are far easier to collect than 3D shapes. As a result, though sketch-based FIR methods [27–29] are somewhat informative, they could not be easily applied to FG-SBSR.

In this paper, we drew on some essential ideas from FIR methods and considered the specific SBSR challenges, to propose a novel method applicable to category-level FG-SBSR.

3 Category-Level FG-SBSR Dataset

3.1 Existing SBSR Benchmarks

Researchers have been working hard on building large-scale datasets for the SBSR task. At present, there are three commonly used benchmarks, SHREC'13 [21], SHREC'14 [22], and D2O [33]. They all contain thousands of sketches and 3D shape samples span dozens or even over a hundred distinct categories, as shown in Table 1. But the grain size of these categories is relatively coarse (Fig. 3). If research is done only based on such datasets, it is difficult to demonstrate the effectiveness of the proposed method in fine-grained retrieval scenarios, even if experimental metrics are high. Therefore, we need to construct a new benchmark that is more suitable for the much-needed FG-SBSR task.



Fig. 3: Examples from the 'chair' category in the existing 3 SBSR datasets, which vary greatly. It is difficult for users to find the 3D shapes they need at this level of granularity.

3.2 A Novel Dataset Oriented to FG-SBSR

In our review of existing sketch and 3D shape datasets, we observed that samples belonging to the three broad categories of airplanes, cars, and chairs are particularly abundant, largely owing to their extensive practical applications in industry and entertainment. Drawing upon domain-specific design expertise, we further refined these three coarse categories into a set of fine-grained sub-categories, as presented in Fig. 4. To ensure the reliability and consistency of the reclassification, we invited a group of graduate students, under the close supervision of domain experts, to reclassify the corresponding samples in the sketch datasets Tu-Berlin [15], sketchy [16], and pro-chair-3d [17], as well as the 3D shape dataset FG3D [18], according to our proposed taxonomy. In addition, these students manually drew a collection of new sketches that adhere to sub-category specifications, to supplement the existing data and address potential imbalances.

As a result, the novel category-level fine-grained SBSR dataset was constructed. It contains 7,666 sketches and 20,445 3D shapes, grouped into 54 sub-categories in three broad categories of airplanes, cars and chairs. The sample size of each sub-category

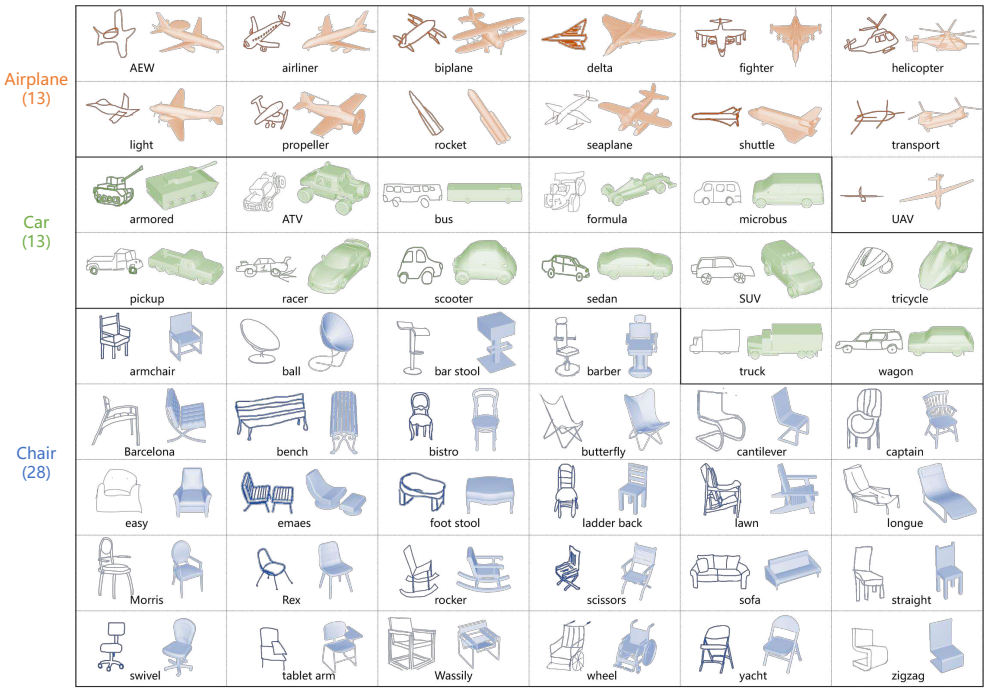


Fig. 4: Categories and some sketch-3D shape examples in the proposed FG-SBSR dataset.

is guaranteed to be 12 or greater. As shown in Table 1 as well as Fig. 3 and 4, the size, balance and sample quality of our new FG-SBSR dataset are comparable to the 3 currently used coarse-grained SBSR benchmarks introduced earlier.

To the best of our knowledge, this is the first dataset oriented to the category-level FG-SBSR task, i.e., each sketch is assigned to a fine-grained sub-category and corresponds to all 3D shapes sharing that label. We believe our production will be a notable step for SBSR tasks toward fine-grained scenarios.

Table 1: The Dataset Size Comparison Between Our FG-SBSR And 3 Coarse-grained SBSR Benchmarks

Dataset	Fine-grained	Catagories	Sketches Total / Average	3D Shapes Total / Average	Smallest Category
SHREC'13	-	90	7200 / 80	1258 / 13.98	4
SHREC'14	-	171	13680 / 80	8987 / 52.56	1
D2O	-	50	7M / 140K	8992 / 179.84	30
FG-SBSR(ours)	✓	54	7666 / 141.96	20445 / 378.61	12

4 Methodology

4.1 Problem Definition

We first give the preliminary definition of category-level FG-SBSR. Dataset $\mathcal{D}$ is defined by $\{(x_i, y_i, c_i)|c_i \in \mathcal{C}\}$, where x_i, y_i and c_i represent the sketch, 3D shape and the category label of the i^{th} sample, respectively. And here, $\mathcal{C}$ is the set of different categories in $\mathcal{D}$. We then split $\mathcal{D}$ into a training set, $\mathcal{D}_{train} = \{(x_i, y_i, c_i)|c_i \in \mathcal{C}_{train}\}$, and a testing set, $\mathcal{D}_{test} = \{(x_i, y_i, c_i)|c_i \in \mathcal{C}_{test}\}$, where $\mathcal{D}_{train} \cup \mathcal{D}_{test} = \mathcal{D}$, $\mathcal{D}_{train} \cap \mathcal{D}_{test} = \emptyset$ and $\mathcal{C}_{train} = \mathcal{C}_{test} = \mathcal{C}$.

4.2 Fine-grained 3D Shape Feature Extractor

To cope with the desired fine-grained scenario, the 3D shape feature extractor needs to be redesigned. As shown in Fig. 5, the basic architecture of our extractor follows the tradition of 3D shape feature learning [23, 34–36]: we first render a number of 2D views of the 3D shape, extract the feature of each view, and then aggregate them to obtain the final 3D shape feature. However, after the view feature is captured through the convolutional layers, we explicitly design a branch that focuses on partial details, in addition to the overall spatial branch.

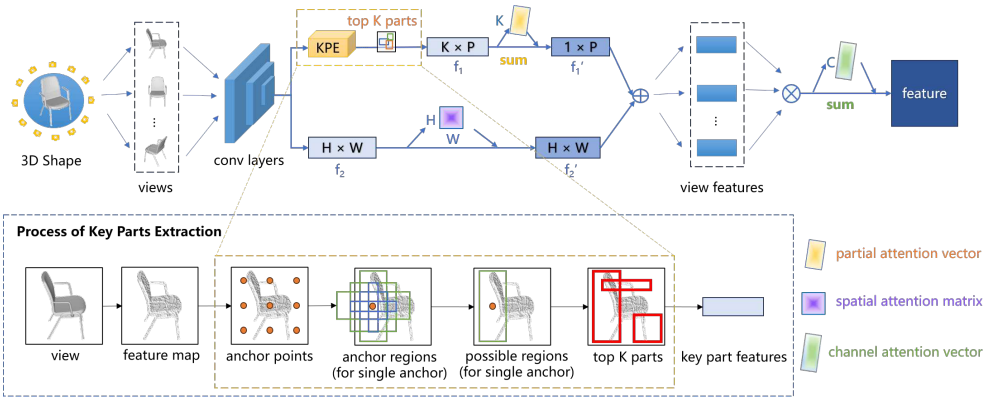


Fig. 5: Illustration of the proposed fine-grained 3D shape feature extractor.

The core of the partial branch is KPE, a deep network module that can extract and score the features of a set of key part regions in each view of the 3D shape. We illustrate the key parts extraction process in Fig. 5. It first generates a series of anchor points on the feature map derived from the view through the convolutional layers. A set of regions is then delineated around each anchor point. Regions exceeding the boundaries and redundant regions overlapping too much with others are eliminated to filter out all possible regions. Afterward, the top K scoring regions are selected and thus get the corresponding key part features of the view.

Then, we fuse the K part features via a partial attention vector, and concatenate the fine-tuned part and overall spatial features of the view. And thereafter apply a

channel-level attention vector [37] to aggregate the feature of each view to obtain the final feature of the whole 3D shape.

4.3 Fine-grained Sketch Feature Extractor

Sketches are drawn by people one stroke by one stroke, so there is vast blank space in the sketch image and huge stylistic differences between sketch samples [38–40]. Therefore, although both are 2D images, we should not simply employ the various CNNs designed for real-world photos (as many SBSR methods did), to process sketches that are generated by highly engaged human abstract minds.

As a result, we proposed a novel fine-grained sketch feature extractor, which is shown in Fig. 2 and has the following features compared to the photo-oriented CNNs: (1) increases the convolution kernel size at the first- and second-layer filters, (2) reduces the depth of the feature extractor network and (3) sets relatively high dropout values at the final fully connected layers.

4.4 3D Shape-Guided Sketch Learning

SBSR is a typical cross-domain task. We need to map the semantically equivalent sketch and 3D shape features to adjacent locations in an aligned feature space, i.e., cross-domain alignment. General SBSR methods treat the domain of sketches and the domain of 3D shapes as equal and map the two kinds of features to a common semantic space.

However, if we take a closer look at the data and related datasets for both domains, the following two facts will be found: (1) due to the production barrier, 3D shapes are always more detailed and regular than sketches, so it is easier for the neural network to learn features of 3D shapes; (2) the amount of available data in the 3D shape domain is relatively larger than that of the sketch domain.

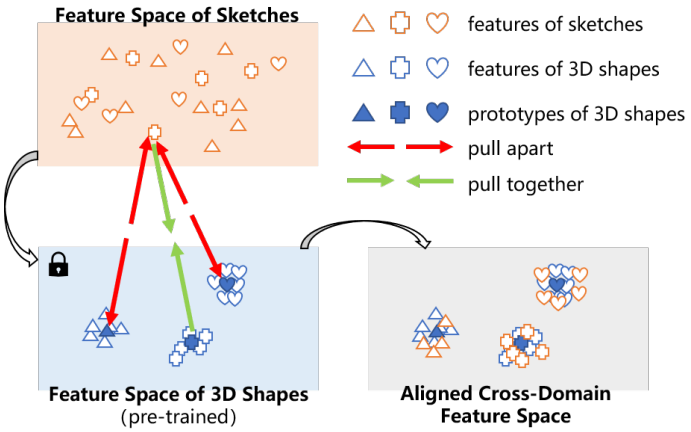


Fig. 6: Illustration of using 3D shape prototypes to guide sketches alignment.

Thus, we identified the cross-domain training approach of guiding sketch learning by 3D shapes, as illustrated in Fig. 6. Our training process is divided into two stages: (1) train a 3D shape feature extractor; (2) use the trained 3D shape extractor to guide the training of the sketch feature extractor, specific steps are given in Algorithm 1. We apply the prototype contrastive loss in both stages, which formulated as

$$\mathcal{L}_{NCE}^{i,P} = -\log \frac{\exp(sim(v_i, P_{c^+})/\tau)}{\sum \exp(sim(v_i, P_c)/\tau)}, \ P_c \in \mathcal{P}. \tag{1}$$

where sim is a similarity measure function that uses the L2 parametric, τ is the temperature coefficient (0.07), c is the category label and c^+ refers to the positive category, v is the feature extracted from the sample of 3D shape in stage 1 while of the sketch in stage 2, $\mathcal{P}$ is the set of prototypes of 3D shapes , P_c is the prototype of category c.

Algorithm 1: Steps of 3D Shape-Guided Sketch Learning

Input: sketch dataset $\mathcal{D}^x$ and sketch feature extractor $\mathcal{FE}^x$, 3D shape dataset $\mathcal{D}^y$ and the pre-trained 3D shape feature extractor $\mathcal{FE}^y$, number of memory samples per category m, number of negative categories n, max number of training epochs $MaxEpoch$

Output: sketch feature extractor $\mathcal{FE}^x$

construct 3D shape feature set $\mathcal{D}_f^y$ with $\mathcal{FE}^y$ and $\mathcal{D}^y$;
initialize 3D shape feature memory bank $\mathcal{M}$ with $\mathcal{D}_f^y$ and m;
initialize sketch feature extractor $\mathcal{FE}^x$, bind it to the optimizer $\mathcal{O}$;
for *epoch = 1:MaxEpoch* **do**
  for *sketch sample x and it's category label c in Dataloader($\mathcal{D}^x$)* **do**
    initialize the set of category labels $\mathcal{C}$;
    put the positive label c into set $\mathcal{C}$;
    randomly choose n negative category labels c^- according to c and put them into $\mathcal{C}$;
    initialize the set of prototypes $\mathcal{P}$;
    for c_i *in* $\mathcal{C}$ **do**
      calculate the prototype P of category c_i by cluster corresponding 3D shape features in $\mathcal{M}$;
      put P into set $\mathcal{P}$;
    end
    calculate loss with Eq.1;
    update sketch feature extractor $\mathcal{FE}^x$ with $\mathcal{O}$;
    randomly choose m 3D shape features with label c in $\mathcal{D}_f^y$ to update $\mathcal{M}$;
  end
end

5 Experiments

5.1 Datasets and Settings

5.1.1 Datasets Setting

To verify the efficacy of our proposed FG-SBSR method, we conducted experiments on all the 3 subsets of our new FG-SBSR dataset introduced in Section 3. In parallel, in order to verify the generalization of our method, experiments are also conducted on the 2 commonly used coarse-grained SBSR benchmarks, SHREC'14 [22] and D2O [33].

Since the data in SHREC'14 is extremely unbalanced, we chose a subset of 30 categories with the maximum sample size to do the experiments (notated as SHREC'14-SUB). To ease the replication, the names of categories involved in SHREC'14-SUB are listed in Table 2:

Table 2: Categories involved in SHREC'14-SUB.

airplane	armchair	bed	bicycle	bookshelf
bus	cabinet	cell_phone	chair	couch
door	eyeglasses	floor_lamp	guitar	helicopter
house	keyboard	knife	laptop	motorbike
piano	ship	spoon	sword	table
tablelamp	telephone	train	vase	wheel

They are all daily necessities, which happens to be a relatively fine-grained retrieval setting.

For D2O, the sample size of sketches is quite large, so we randomly took 1,000 per category for experimentation.

We randomly split the samples into a training set and a testing set according to 80%–20%, except the sketches in SHREC'14, which used the original 50–30 division.

5.1.2 Implementation Details

We implemented the proposed method using PyTorch. A ResNet-50 [24] (pre-trained on ImageNet [41]) was applied as the CNN backbone of the fine-grained 3D shape feature extractor. Memory size m per category was set to 5.

The maximum epoch number was 200, and Adam was adopted to optimize our network with a learning rate initially set to $1e-4$ and set to exponentially decay at a rate of 0.95. For testing, each positive sample tier was paired with nine randomly selected negative tiers.

5.1.3 Evaluation Metrics

To quantify the performance of our proposed method, we adopt several commonly used evaluation metrics [20] in the experiments: nearest neighbor (NN), first-tier (FT), second-tier (ST), normalized discounted cumulated gain (nDCG), E-measure

(E), mean reciprocal rank (MRR) and mean average precision (mAP) evaluators were used. Except for E, the larger the metric, the better.

5.2 Experiment Results

We compared our method results with those of state-of-the-art methods, including Siamese [1], DCHML [3], TCL [5], CGN [7], JFLN [8] and prototype learning-based simple baseline (PLSB) [11]. For fairness, all methods used the same settings described in Section 5.1. According to the original papers, the Siamese method used a three-layer CNN as its backbone, AlexNet was applied to DCHML, and the backbone of TCL, CGN, JFLN and PLSB was the same pre-trained ResNet-50. The code for the methods other than the fully open-source JFLN is our own production.

Table 3: Sketch-3D Shape Retrieval Performance on 3 Subsets of the FG-SBSR Dataset.

Subset	Method	NN ↑	FT ↑	ST ↑	nDCG ↑	E ↓	MRR ↑	mAP ↑
Airplane	siamese	0.155	0.139	0.120	0.198	0.755	0.283	0.213
	DCHML	0.098	0.092	0.100	0.151	0.763	0.175	0.192
	TCL	0.113	0.100	0.100	0.153	0.774	0.234	0.179
	CGN	0.142	0.163	0.179	0.265	0.686	0.258	0.282
	JFLN	0.161	0.165	0.165	0.253	0.699	0.279	0.274
	PLSB	0.136	0.192	0.199	0.294	0.671	0.324	0.290
	ours	**0.369**	**0.306**	**0.257**	**0.433**	**0.619**	**0.500**	**0.404**
Car	siamese	0.168	0.144	0.122	0.204	0.755	0.308	0.214
	DCHML	0.111	0.078	0.083	0.134	0.779	0.236	0.172
	TCL	0.111	0.092	0.109	0.166	0.745	0.184	0.218
	CGN	0.158	0.191	0.150	0.239	0.702	0.286	0.272
	JFLN	0.152	0.173	0.157	0.244	0.706	0.282	0.266
	PLSB	0.149	0.198	0.188	0.289	0.666	0.302	0.306
	ours	**0.369**	**0.314**	**0.275**	**0.455**	**0.601**	**0.483**	**0.426**
Chair	siamese	0.123	0.126	0.122	0.187	0.772	0.281	0.186
	DCHML	0.105	0.105	0.105	0.163	0.740	0.227	0.212
	TCL	0.098	0.098	0.103	0.156	0.773	0.206	0.181
	CGN	0.113	0.087	0.100	0.152	0.787	0.239	0.167
	JFLN	0.136	0.141	0.125	0.198	0.745	0.269	0.221
	PLSB	0.123	0.135	0.139	0.209	0.719	0.219	0.248
	ours	**0.359**	**0.298**	**0.242**	**0.414**	**0.628**	**0.499**	**0.390**

In Tables 3, the best evaluation metrics of each method on the 3 subsets of our new FG-SBSR datasets are shown in boldface, respectively. It can be seen that our proposed method has outstanding performance in fine-grained retrieval scenarios, with all metrics far exceeding the existing methods. The NN metrics of our method in these 3 fine-grained benchmarks even had an improvement of more than 100%.

To verify the generality of our methods, we also conducted experiments on SHREC'14-SUB and D2O. The results are shown in Table 4, where the best metrics are marked in bold, respectively. We can see that the proposed methods also perform well on the coarse-grained benchmarks.

Table 4: Sketch-3D Shape Retrieval Performance on 2 Coarse-grained SBSR Benchmarks.

Benchmark	Method	NN ↑	FT ↑	ST ↑	nDCG ↑	E ↓	MRR ↑	mAP ↑
SHREC'14-SUB	siamese	0.111	0.127	0.131	0.195	0.726	0.201	0.239
	DCHML	0.316	0.234	0.174	0.314	0.711	0.478	0.292
	TCL	0.304	0.164	0.141	0.254	0.750	0.449	0.235
	CGN	0.480	0.323	**0.247**	0.448	**0.638**	0.621	0.398
	JFLN	0.510	0.315	0.239	0.445	0.648	0.632	0.395
	PLSB	0.519	0.288	0.226	0.421	0.654	0.622	0.383
	ours	**0.533**	**0.327**	0.236	**0.451**	0.647	**0.660**	**0.401**
D2O	siamese	0.134	0.132	0.123	0.192	0.757	0.294	0.203
	DCHML	0.184	0.229	0.204	0.312	0.653	0.385	0.317
	TCL	0.256	0.231	0.203	0.325	0.645	0.406	0.341
	CGN	0.300	0.326	0.262	0.421	0.609	0.473	0.399
	JFLN	0.328	0.357	0.295	0.474	0.574	0.482	0.456
	PLSB	0.285	0.264	0.240	0.386	0.636	0.438	0.361
	ours	**0.451**	**0.425**	**0.313**	**0.533**	**0.543**	**0.575**	**0.518**

Fig. 7 shows some retrieval results of our FG-SBSR method on the mentioned datasets. The names of corresponding datasets are listed below the retrieval results. For each example, the query sketch is listed on the left side, and its retrieved top-five 3D shapes are listed on the right based on ranking order. The failure cases are marked in orange and boxed up, with their ground truth category labeled below.

This is only a qualitative analysis, but it can be easily seen that our method is able to efficiently retrieve 3D shape results that are visually very similar to the input sketches. Even the failure cases somehow look similar to success cases (e.g., 'microbus' and 'truck' in the FG-SBSR-Car subset).

5.3 Further Analysis

5.3.1 Ablation Study

We conducted four ablation studies on FG-SBSR-Airplane dataset to validate our proposed method. A checkmark below **FG3D** indicates the use of the proposed fine-grained 3D shape feature extractor, otherwise the original MVCNN [23]. A checkmark below **FGSK** indicates our fine-grained sketch feature extractor is used, otherwise the use of ResNet-50 [24], which is a normal CNN designed for real-world photos. A checkmark below **GA** indicates the use of our 3D shape-guided alignment approach, otherwise, we treated both domains equally and mapped sketches and 3D shapes to a common semantic space.

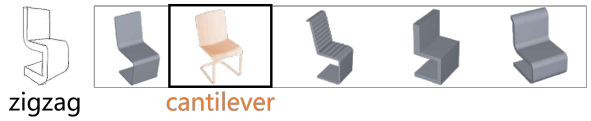
The results are shown in Table 5, where the best metrics are in boldface. It is easy to tell that the two new fine-grained feature extractors for 3D shape and sketch, as well as the 3D shape-guided sketch learning approach, all play a key role in dealing with the FG-SBSR problem. The relatively low metrics in row 2 of the table may suggest that the understanding of sketches is the critical challenge and development direction for SBSR tasks.



(a) FG-SBSR-Airplane



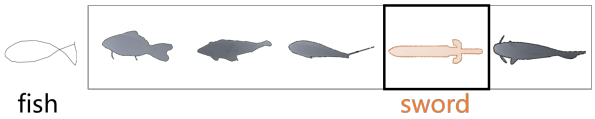
(b) FG-SBSR-Car



(c) FG-SBSR-Chair



(d) SHREC'14-SUB



(e) D2O

Fig. 7: Retrieval results obtained by the proposed FG - SBSR method on the mentioned datasets. The failure cases are marked in orange and boxed up.

Table 5: Ablation Study on FG-SBSR-Airplane.

	FG3D	FGSK	GA	NN ↑	FT ↑	ST ↑	nDCG ↑	E ↓	MRR ↑	mAP ↑
1	-	✓	✓	0.288	0.252	0.200	0.338	0.659	0.397	0.348
2	✓	-	✓	0.184	0.146	0.126	0.204	0.737	0.269	0.233
3	✓	✓	-	0.275	0.222	0.207	0.336	0.658	0.390	0.333
4	✓	✓	✓	**0.369**	**0.306**	**0.257**	**0.433**	**0.619**	**0.500**	**0.404**

5.3.2 Intra-Domain Retrieval Results

We proposed a 3D shape-guided sketch learning approach in Section 4.4. To further justify it, we conducted intra-domain retrieval experiments using sketches and 3D shapes contained in the FG-SBSR dataset, respectively. The results are shown in Table 6. The SBSR performance served as a reference and is also listed in the table in italics. It can be easily noticed that the retrieval performance in the 3D shape domain far exceeds that in the sketch domain, and the SBSR metrics are always better than those of sketches as well. This to some extent confirms that 3D shapes are more refined, more feature-centralized, and also more learnable by the deep network, while sketches are just the opposite. Therefore, our approach of training a 3D shape feature extractor first, and then using it to guide the training of the sketch feature extractor, can make sufficient use of the data characteristics of 3D shapes and achieve superior cross-domain retrieval results.

Table 6: Intra & *Cross* Domain Retrieval Performance on 3 Subsets of the FG-SBSR Dataset.

Subset	Domain	NN ↑	FT ↑	ST ↑	nDCG ↑	E ↓	MRR ↑	mAP ↑
Airplane	3D Shape	0.817	0.695	0.443	0.831	0.401	0.871	0.784
	Sketch	0.373	0.206	0.157	0.300	0.728	0.499	0.278
	SBSR	*0.369*	*0.306*	*0.257*	*0.433*	*0.619*	*0.500*	*0.404*
Car	3D Shape	0.849	0.707	0.452	0.850	0.393	0.890	0.797
	Sketch	0.358	0.215	0.171	0.311	0.722	0.502	0.279
	SBSR	*0.369*	*0.314*	*0.275*	*0.455*	*0.601*	*0.483*	*0.426*
Chair	3D Shape	0.696	0.493	0.372	0.674	0.503	0.779	0.603
	Sketch	0.294	0.212	0.174	0.303	0.711	0.447	0.282
	SBSR	*0.359*	*0.298*	*0.242*	*0.414*	*0.628*	*0.499*	*0.390*

6 Conclusion

In this paper, we extended the common category-level SBSR task to the more practical fine-grained scenario. Firstly, to enable FG-SBSR, we proposed a new dataset containing 54 sub-categories in three broad categories of airplanes, cars and chairs. Then, we designed a novel method, including two new fine-grained feature extractors and a 3D shape-guided sketch alignment approach, to handle the subtle cross-domain retrieval task. Extensive experiments on our new FG-SBSR dataset and two widely used coarse-grained SBSR benchmarks could demonstrate the efficiency of the proposed method. We believe this work will be a notable step for SBSR tasks toward real-world application. In the future, we plan to collect more data samples and conduct the FG-SBSR study in some open scenarios.

Funding Declaration

This work was supported by the National Natural Science Foundation of China under Grant No. 62171320.

References

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

[2] Dai, G., Xie, J., Zhu, F., Fang, Y.: Deep correlated metric learning for sketch-based 3d shape retrieval. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)

[3] Dai, G., Xie, J., Fang, Y.: Deep correlated holistic metric learning for sketch-based 3d shape retrieval. IEEE Transactions on Image Processing **27**(7), 3374–3386 (2018)

[4] Qi, A., Song, Y.-Z., Xiang, T.: Semantic embedding for sketch-based 3d shape retrieval. In: BMVC, vol. 3, pp. 11–12 (2018)

[5] He, X., Zhou, Y., Zhou, Z., Bai, S., Bai, X.: Triplet-center loss for multi-view 3d object retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1945–1954 (2018)

[6] Xu, Y., Hu, J., Wattanachote, K., Zeng, K., Gong, Y.: Sketch-based shape retrieval via best view selection and a cross-domain similarity measure. IEEE Transactions on Multimedia **22**(11), 2950–2962 (2020)

[7] Dai, W., Liang, S.: Cross-modal guidance network for sketch-based 3d shape retrieval. In: 2020 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2020). IEEE

[8] Zhao, Y., Liang, Q., Ma, R., Nie, W., Su, Y.: Jfln: Joint feature learning network for 2d sketch based 3d shape retrieval. Journal of Visual Communication and Image Representation **89**, 103668 (2022)

[9] Liang, S., Dai, W., Cai, Y., Xie, C.: Sketch-based 3d shape retrieval via teacher–student learning. Computer Vision and Image Understanding **239**, 103903 (2024)

[10] Su, Y., Li, W., Bai, J., Lin, G.: Skd-sbsr: Structural knowledge distillation for sketch-based 3d shape retrieval. Knowledge-Based Systems **310**, 112891 (2025)

[11] Liang, S., Dai, W., Cheng, C., Wei, Y., Cai, Y.: Rethinking sketch-based 3d shape retrieval: A simple baseline and benchmark reconstruction. Neurocomputing **618**, 128990 (2025)

[12] Yu, Q., Liu, F., Song, Y.-Z., Xiang, T., Hospedales, T.M., Loy, C.-C.: Sketch me that shoe. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

[13] Bhunia, A.K., Sain, A., Shah, P., Gupta, A., Nath Chowdhury, P., Xiang, T.,

Song, Y.-Z.: Adaptive fine-grained sketch-based image retrieval. arXiv e-prints, 2207–01723 (2022) https://doi.org/10.48550/arXiv.2207.01723 arXiv:2207.01723 [cs.CV]

[14] Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., Wang, L.: Learning to navigate for fine-grained classification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 420–435 (2018)

[15] Eitz, M., Hays, J., Alexa, M.: How do humans sketch objects? ACM Transactions on graphics (TOG) **31**(4), 1–10 (2012)

[16] Sangkloy, P., Burnell, N., Ham, C., Hays, J.: The sketchy database: learning to retrieve badly drawn bunnies. ACM Transactions on Graphics (TOG) **35**(4), 1–12 (2016)

[17] Qi, A., Gryaditskaya, Y., Song, J., Yang, Y., Qi, Y., Hospedales, T.M., Xiang, T., Song, Y.-Z.: Toward fine-grained sketch-based 3d shape retrieval. IEEE transactions on image processing **30**, 8595–8606 (2021)

[18] Liu, X., Han, Z., Liu, Y.-S., Zwicker, M.: Fine-grained 3d shape classification with hierarchical part-view attention. IEEE Transactions on Image Processing **30**, 1744–1758 (2021)

[19] Furuya, T., Ohbuchi, R.: Ranking on cross-domain manifold for sketch-based 3d model retrieval. In: 2013 International Conference on Cyberworlds, pp. 274–281 (2013). IEEE

[20] Li, B., Lu, Y., Godil, A., Schreck, T., Bustos, B., Ferreira, A., Furuya, T., Fonseca, M.J., Johan, H., Matsuda, T., *et al.*: A comparison of methods for sketch-based 3d shape retrieval. Computer Vision and Image Understanding **119**, 57–80 (2014)

[21] Li, B., Lu, Y., Godil, A., Schreck, T., Aono, M., Johan, H., et al.: Shrec'13 track: Large scale sketch-based 3d shape retrieval (2013)

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

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

[24] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[25] Wei, X.-S., Wu, J., Cui, Q.: Deep learning for fine-grained image analysis: A survey. arXiv preprint arXiv:1907.03069 (2019)

[26] Li, Y., Hospedales, T.M., Song, Y.-Z., Gong, S.: Fine-grained sketch-based image retrieval by matching deformable part models (2014)

[27] Pang, K., Yang, Y., Hospedales, T.M., Xiang, T., Song, Y.-Z.: Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10347–10355 (2020)

[28] Bhunia, A.K., Yang, Y., Hospedales, T.M., Xiang, T., Song, Y.-Z.: Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9779–9788 (2020)

[29] Bhunia, A.K., Sain, A., Shah, P.H., Gupta, A., Chowdhury, P.N., Xiang, T., Song, Y.-Z.: Adaptive fine-grained sketch-based image retrieval. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII, pp. 163–181 (2022). Springer

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

[31] Koley, S., Bhunia, A.K., Sain, A., Chowdhury, P.N., Xiang, T., Song, Y.-Z.: How to handle sketch-abstraction in sketch-based image retrieval? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16859–16869 (2024)

[32] Yu, Q., Song, Y.-Z., Xiang, T., Hospedales, T.M.: SketchX!-Shoe/Chair fine-grained SBIR dataset (2017)

[33] Wang, B., Zhou, Y.: Doodle2object: Practical zero-shot sketch-based 3d shape retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence (2023)

[34] Kanezaki, A., Matsushita, Y., Nishida, Y.: Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5010–5019 (2018)

[35] Hamdi, A., Giancola, S., Ghanem, B.: Mvtn: Multi-view transformation network for 3d shape recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1–11 (2021)

[36] Wang, W., Cai, Y., Wang, T.: Multi-view dual attention network for 3d object

recognition. Neural Computing and Applications **34**(4), 3201–3212 (2022)

[37] Roy, A.G., Navab, N., Wachinger, C.: Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 421–429 (2018). Springer

[38] Yu, Q., Yang, Y., Song, Y.-Z., Xiang, T., Hospedales, T.: Sketch-a-net that beats humans. arXiv preprint arXiv:1501.07873 (2015)

[39] Ha, D., Eck, D.: A neural representation of sketch drawings. arXiv preprint arXiv:1704.03477 (2017)

[40] Xu, P., Joshi, C.K., Bresson, X.: Multigraph transformer for free-hand sketch recognition. IEEE Transactions on Neural Networks and Learning Systems (2021)

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