Doodle to Object: Practical Zero-Shot Sketch-Based 3D Shape Retrieval

Bingrui Wang, Yuan Zhou[*]

School of Electrical and Information Engineering, Tianjin University, Tianjin, China
{wangbingrui, zhouyuan}@tju.edu.cn

Abstract

Zero-shot (ZS) sketch-based three-dimensional (3D) shape retrieval (SBSR) is challenging due to the abstraction of sketches, cross-domain discrepancies between two-dimensional sketches and 3D shapes, and ZS-driven semantic knowledge transference from seen to unseen categories. Extant SBSR datasets suffer from lack of data, and no current SBSR methods consider ZS scenarios. In this paper, we contribute a new Doodle2Object (D2O) dataset consisting of 8,992 3D shapes and over 7M sketches spanning 50 categories. Then, we propose a novel prototype contrastive learning (PCL) method that effectively extracts features from different domains and adapts them to unseen categories. Specifically, our PCL method combines the ideas of contrastive and cluster-based prototype learning, and several randomly selected prototypes of different classes are assigned to each sample. By comparing these prototypes, a given sample can be moved closer to the same semantic class of samples while moving away from negative ones. Extensive experiments on two common SBSR benchmarks and our D2O dataset demonstrate the efficacy of the proposed PCL method for ZS-SBSR. Resource is available at https://github.com/yigohw/doodle2object.

Introduction

Free-hand sketches comprise a universal modality of human artistic communications across society. Throughout history, people of all backgrounds have used sketches to tell stories, express needs, and transcend language barriers. In modern times, with the pervasive nature of touchscreen devices and the rapid development three-dimensional (3D) sensing and modeling technologies, sketch-based 3D shape retrieval (SBSR) has attracted widespread attention for these and other reasons (Wang, Kang, and Li 2015; Dai et al. 2017; Dai, Xie, and Fang 2018; Qi, Song, and Xiang 2018; He et al. 2018; Xu et al. 2020; Dai and Liang 2020).

However, hand sketches must be drawn asynchronously by humans, at the same time, there is a certain threshold for 3D sensing and modeling. Therefore, data scarcity is a major bottleneck faced by almost all researches on both sketches and 3D shapes, when attempting to intuit meaning from adjacent domains. Compared with computer vision tasks that rely on large-scale image datasets (Li et al. 2013, 2014b), SBSR tasks are dwarfed in both variety and volume, wherein the maximum number of classes number just over 100 with no more than 80 samples per class in most cases. This predicament seems to make the task of SBSR nigh impossible, thus largely motivates the task of zero-shot SBSR (ZS-SBSR), which effectively identifies unseen but related classes during training to promote practical SBSR knowledge expansion applications.

ZS-SBSR is extremely challenging based on the field's highly abstract sketches and huge cross-domain discrepancies between two-dimensional (2D) drawing sketches and 3D shapes. Nevertheless, for ZS learning, ZS-SBSR requires the transference of semantic knowledge from seen to unseen classes. However, this describes a capability area in which traditional SBSR methods generally cannot be applied.

To handle these challenges, we propose a new prototype contrastive learning (PCL) method (Fig. 1) whose training process consists of two stages, intra-domain feature extraction and cross-domain feature alignment, upon which our novel prototype contrastive loss is framed. After training, samples from different domains are clustered near their corresponding prototype to effectively distinguished positive and negative samples, including unseen ones that appear during retrieval. Accordingly, our PCL method can handle the ZS-SBSR problem.

Furthermore, to advance the practical adaptation of ZS-SBSR, it is necessary to build a suitable dataset. Thus, we start with the Doodle2Object (D2O) dataset, which can be augmented to meet the challenges of ZS-SBSR with the help of ModelNet40 (Wu et al. 2015) and QuickDraw (Ha and Eck 2017). D2O consists of 8,992 3D shapes and more than 7M sketches spanning 50 categories. The dataset guarantees a sample size of at least 30 items per 3D shape class and contains the temporal information of sketches. We believe that the proposed D2O dataset has the potential to mimic the real-world semantic gap between sketches and the larger domain of 3D shapes.

Extensive experiments are conducted in this study using the two most common SBSR benchmarks (i.e., SHREC'13 (Li et al. 2013) and SHREC'14 (Li et al. 2014b)), showing that the proposed PCL method applied to ZS-SBSR outperforms contemporary methods on temporal sketch information, demonstrating PCL's potential for the practical appli-

[*] Corresponding author.

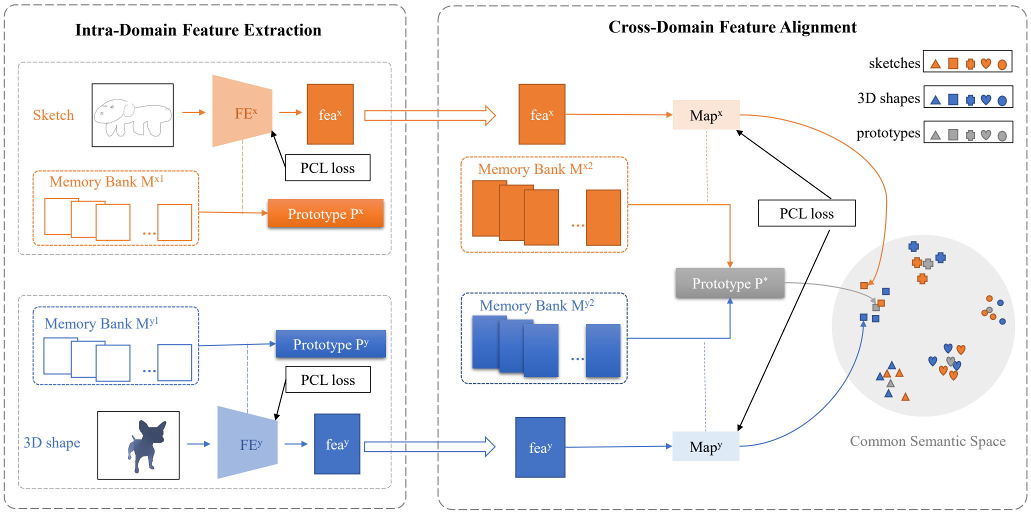


Figure 1. Proposed prototype contrastive learning method for zero-shot sketch-based three-dimensional shape retrieval.

cation of ZS-SBSR. In summary, the main contributions of this paper are as follows:

1. An extended SBSR to the more practical ZS-SBSR using a novel PCL method to improve effective retrieval for cross-domain data as well as unseen classes;
2. PCL-clustered samples of sketches or 3D shapes near the corresponding prototype that take advantage of both prototype and contrastive learning to effectively distinguish between negative and unseen samples;
3. An augmented D2O dataset that alleviates the data scarcity of existing SBSR data and provides sketches that are closer to real scenarios;
4. A method that outperforms state-of-the-art SBSR methods on two widely used benchmarks and our augmented D2O dataset.

Related Work

SBSR

For decades, many SBSR methods (Li et al. 2013, 2014b; Furuya and Ohbuchi 2013; Li et al. 2014a) have been built upon handcrafted features. Then, (Wang, Kang, and Li 2015) used a Siamese deep learning network to address cross-domain retrieval problems. Deep correlated (Dai et al. 2017) and holistic metric learning (Dai, Xie, and Fang 2018) methods both use discriminative losses to increase the distinguishability of different inner-domain classes and correlation losses to minimize the distances between inter-domain classes. For the first time, structural equation modeling (Qi, Song, and Xiang 2018) aligns sketches and 3D shapes in a common semantic space rather than the usual joint feature space. The triplet-center loss (He et al. 2018) improves the triplet loss using the combination of center loss. Best view selection (Xu et al. 2020) selects the best perspective view of 3D shapes according to training sketches before adopting a multiview convolutional neural network (Su et al. 2015). Based on the idea of knowledge distillation, the cross-modal guidance network (Dai and Liang 2020) first employs a 3D shape feature extractor and considers it as a teacher to guide the feature learning of sketches.

We adopt a deep metric method based on prototype learning in this paper in which the training process is an instance of online distillation. Sketches and 3D shapes are used together to construct prototypes in the common semantic space.

ZS Retrieval Learning

ZS learning is widely used in various computer vision applications, such as the related ZS sketch-based image retrieval (SBIR) task. (Yelamarthi et al. 2018) used conditional variational and adversarial autoencoders to associate the visual information of a sketch to that of an image. (Dey et al. 2019) applied domain and triplet ranking losses to learn a common embedding space in which the distances between sample pairs in the same class were smaller than pairs from different classes. The SEM-paired cycle-consistent generative model (Dutta and Akata 2019) maps sketch and image characters to a common semantic space while preserving the back-translation ability. Semantic-aware knowledge preservation (Liu et al. 2019) generates fake images from sketches for the cross-modal retrieval task. The progressive cross-modal semantic network (Deng et al. 2020) cross-reconstructs semantic features extracted from sketches and images. The prototype- and memory-enhanced joint distribution optimal transport method (Hu et al. 2021) introduces trainable cluster prototypes and feature memory banks for unsupervised SBIR tasks.

In this paper, we map sketch and 3D shape features adjacent to the corresponding prototype, built by the memory banks of both domains in the common semantic space.

A More Practical SBSR Dataset

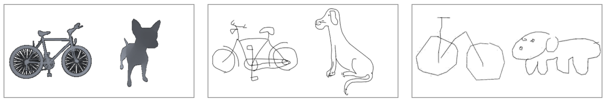

(a) (b) (c)

Figure 2. Samples from sketch or 3D shape datasets, take categories of bicycle and dog as examples: (a) SHREC'13/14 shapes; (b) SHREC'13/14 sketches; (c) QuickDraw sketches. QuickDraw sketches are noticeably more abstract but are still recognizable.

SHREC'13 (Li et al. 2013) and SHREC'14 (Li et al. 2014b) are the most widely used benchmark datasets for SBSR tasks. Figs. 2(a) and 2(b) show examples of 3D shapes and sketches from the datasets. SHREC'13 contains 7,200 hand-drawn sketches and 1,258 3D shapes divided into 90 classes, each containing 50 sketches for training and 30 for testing. However, the number of distinct classes differs for the 3D sketches, where 23 of the 90 classes have sample sizes no greater than five. SHREC'14 is larger than SHREC'13, which contains 13,680 sketches and 8,987 3D shapes grouped into 171 classes. Like SHREC'13, this dataset contains 80 sketches per class: 50 for training and 30 for testing. 38 of the total 171 classes have samples of 3D less than five shapes.

The scarcity of data in some classes complicates network learning. Moreover, the sketch samples of SHREC'13/14 are normal Portable Network Graphics formats that lack natural temporal sketching information. These have relatively fine styles compared with an actual user's free-hand doodling, which can be described by comparing Figs. 2(b) and 2(c). Accounting for the limitations of the benchmark datasets, we contribute a new D2O dataset for the ZS-SBSR task consisting of 8,992 3D shapes and more than 7M sketches divided into 50 classes.

For the 3D shapes, we extend the SHREC'14 datasets with ModelNet40 (Wu et al. 2015), one of the most widely used for 3D shape recognition. The sample size of each class is guaranteed to be 30 or greater. Nearly 60% of the classes contain more than 100 3D shapes, and five classes have more than 500 samples.

To gain a better insight into the actual SBSR performance, we use the QuickDraw (Ha and Eck 2017) dataset, which contains 50M drawings sorted into 345 classes collected from the "Quick, Draw!" game in which users sketched objects of a given class in 20 s. The dataset contains a large variety of temporal data based on the accumulation of strokes. Samples tend to match real SBSR scenarios in which non-expert artists search for 3D shapes by sketching. Examples are shown in Fig. 2(c).

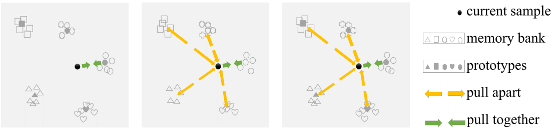
To the best of our knowledge, this is the first dataset oriented to an SBSR task containing many sketches created by real users, guaranteeing sample sizes for each 3D shape class. Our D2O production is a notable step toward the practical application of ZS-SBSR techniques.

Methodology

Problem Definition

In describing ZS-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 class label of the i^{th} sample, respectively. Here, $\mathcal{C}$ is the set of all possible classes in $\mathcal{D}$. In our ZS setting, the model is required to test data whose classes have never been used for training; thus, $\mathcal{C}$ is split into $\mathcal{C}_s$(seen)and $\mathcal{C}_u$(unseen), where $\mathcal{C}_s \cup \mathcal{C}_u = \mathcal{C}$ and $\mathcal{C}_s \cap \mathcal{C}_u = \emptyset$. According to $\mathcal{C}_s$ and $\mathcal{C}_u$, we define the training set, $\mathcal{D}_{train} = \{(x_i, y_i, c_i)|c_i \in \mathcal{C}_s\}$, and the testing set, $\mathcal{D}_{test} = \{(x_i, y_i, c_i)|c_i \in \mathcal{C}_u\}$.

PCL



(a) PL (b) CL (c) PCL

Figure 3. Prototype contrastive learning (PCL).

Retrieval methods usually apply triplet loss training, which pulls the anchor closer to the positive sample and moves it farther away from negative ones. Hence, models trained in this fashion show good retrieval performance. Nonetheless, this does not well-match the real-world situation. Specifically, the triplet loss-based approach assumes an equal proportion of positive and negative samples. However, when we conduct real retrieval tasks, many more negative than positive samples exist.

We apply the PCL method to cope with this imbalance using prototype learning (PL) (Snell, Swersky, and Zemel 2017; Ji et al. 2020) to aggregate samples of the same category while using contrastive learning (CL) (Wu et al. 2018; He et al. 2020) to move the sample away from negatives. Our PCL draws on the main ideas of both methods, as shown in Fig. 3. Moreover, it is not bound to any particular modality and has the potential to transfer knowledge from seen to unseen classes. Thus we apply it to the ZS cross-modality retrieval scenario.

The PCL process contains two steps. First, we randomly choose the negative classes based on the training sample and calculate prototypes (including positives) by concentrating the samples in memory. Second, we update the network parameters in memory according to the prototype contrastive loss, $L_{NCE}^{i,P}$ (Eq. 1), which references the instance-wise contrastive loss function (Oord, Li, and Vinyals 2018), formulated as

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

where v is the feature extracted from the sample, sim is a similarity measure function that uses the L2 parametric, τ is the temperature coefficient (0.07), $\mathcal{M}$ is the memory bank in which v is stored, P_c is the clustering center prototype of the class-c features in $\mathcal{M}$, c^+ is the positive class, and $\mathcal{P}$ is the set of prototypes updated with $\mathcal{M}$.

PCL-Based Intra-Domain Feature Extraction

Influenced by the Siamese network, some models (Wang, Kang, and Li 2015; Dai et al. 2017; Dai, Xie, and Fang 2018; Dai and Liang 2020; Dey et al. 2019) that deal with the cross-domain retrieval task use structurally similar feature extractors on both source and target domains and design a complex cross-domain loss function to reduce the inter-domain gap. This is useful for free-hand sketches and real-life photos, which are both 2D images.

However, there is a huge semantic gap between sketches and 3D shapes. Much work has been performed on recognition tasks, and we should not ignore them. Hence, this paper refers to more recent feature extraction networks in both sketching and 3D shape domains and uses the PCL method for training. A PCL pseudocode feature extraction method is given in Algorithm 1.

Algorithm 1: Pseudo-code for PCL-Based Feature Extraction

```
Input: feature extractor $FE_\theta$, training dataset $\mathcal{D}$,
       memory size for each class $k$, number of
       negative classes $n$, $MaxEpoch$
Output: feature extractor $FE_\theta$
binding $\theta$ to the optimizer $\mathcal{O}$;
initialize sample memory bank $\mathcal{M}$ with $\mathcal{D}$ and $k$;
for epoch = $1{:}MaxEpoch$ do
    for training sample $s$ and it's class label $c$ in
      Dataloader($\mathcal{D}$) do
        initialize the set of class labels $\mathcal{C}$;
        put the positive class label $c$ into $\mathcal{C}$;
        randomly choose $n$ negative class 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 class $c_i$ by
              cluster samples in $\mathcal{M}$;
            put $P$ into set $\mathcal{P}$;
        end
        calculate loss with Eq.1;
        optimize $\theta$ with $\mathcal{O}$;
        randomly choose a sample with label $c$ in $\mathcal{M}$,
          replace it with the new sample $s$;
    end
end
```

Network Structure of the Sketch Feature Extractor Humans draw sketches stroke-by-stroke. Thus, they contain both spatial and temporal information. Therefore, we adopt a cascaded convolutional neural network (CNN)–recurrent neural network (RNN) structure to extract their features (Fig. 4).

Sketch-a-net (Yu et al. 2015) was the first specially oriented deep CNN for free-hand sketching. Compared with classic CNN photo-recognition architectures, Sketch-a-net has a larger convolution kernel (15 × 15) for its first-layer filters to capture sparse low-texture features and remove local response normalization layers (Krizhevsky, Sutskever, and Hinton 2012), because most sketches are binary images. The CNN part of our sketch feature extractor uses Sketch-a-net as the backbone and adds a spatial attention branch (Hu, Shen, and Sun 2018) for recalibration.

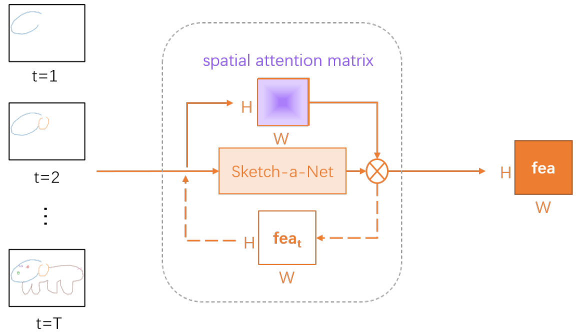


Figure 4. Sketch feature extractor architecture: Sketch-a-net with spatial attention network; T is set to 1 if the input sample lacks temporal information.

We use the RNN to simulate stroke accumulations. However, as each SHREC'13/14 sketch sample contains only one static image, no SBSR method (Wang, Kang, and Li 2015; Dai et al. 2017; Dai, Xie, and Fang 2018; Qi, Song, and Xiang 2018; He et al. 2018; Xu et al. 2020; Dai and Liang 2020; Xu et al. 2022) makes use of the sketch's temporal information. Thus, unless specified, the experiment in this paper is set to $T = 1$ so that features are extracted without entering the RNN loop.

Network Structure of the 3D Shape Feature Extractor Depending on the data representation form, one can extract voxel-, point-cloud-, view-, or hybrid-based methods. Among these, the view-based method has excellent performance and high applicability; thus, we refer to it for 3D shape feature extraction (Su et al. 2015).

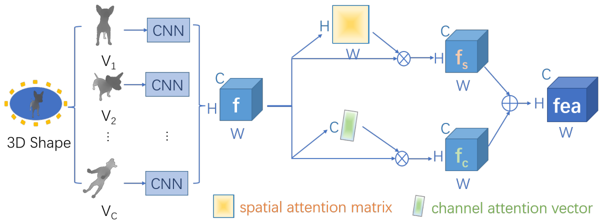


Figure 5. 3D shape feature extraction architecture with multi-view dual attention network.

As shown in Fig. 5, we first render 2D views from multiple 3D object angles and use a shared ResNet-50 CNN (He et al. 2016) as the backbone to extract the features of each view. After aggregating the multiple-view features, we refer to the multiview dual attention network (Wang, Cai, and

Wang 2022) and concurrently apply spatial and view channel attention branches (Roy, Navab, and Wachinger 2018) to obtain the final feature.

PCL-Based Cross-Domain Feature Alignment

Thus far, two single-domain feature extractors have been trained by the PCL method. However, as required by our cross-domain retrieval task, the semantically equivalent sketch and 3D shape features must be mapped to adjacent locations in the common semantic space (i.e., cross-domain alignment).

Because the prototype concept does not require a specific modality, we adopt the PCL method to solve the cross-domain alignment problem. We build sketch-based 3D-shape sample pairs using their class labels and their domains to cluster prototypes into common semantic spaces while assigning several common classes to each pair to train three fully connected network layers. Therefore, each sample, regardless of modality, can be near the prototype of the same class and far from others in the common semantic space. The specific steps are given in Algorithm 2.

Experiments

Datasets and Settings

Dataset Settings To verify the efficacy of our PCL method, we conducted experiments on SHREC'13 (Li et al. 2013), SHREC'14 (Li et al. 2014b), and D2O. For SHREC'13, 23 of the 90 classes had less than five 3D shapes, and we classified them as unseen for testing; the other 67 were classified for training. For SHREC'14, we classified 38 classes with a 3D-shape sample number of five or less as unseen for testing; the other 133 were classified for training. For D2O, we randomly selected 80% of the classes as seen for training and the remaining 20% as unseen. Owing to the extremely large number of sketches in D2O, we randomly took 1,000 per class for experimentation.

To test the performance of the normal SBSR task, we split the seen samples into training and testing sets. For SHREC'13/14, the sketches used the original 50–30 division, whereas the 3D shapes were divided randomly according to 80–20%. For D2O, we randomly chose 80% for training and 20% for testing.

Implementation Details We implemented our PCL using PyTorch. A ResNet-50 (He et al. 2016) pre-trained on ImageNet was used as the 3D shape feature extractor backbone, whereas the sketch feature extractor applied Sketch-a-net (Yu et al. 2015). The size of two alignment maps was 2048-1024-256. Memory size k per category was set to 10.

The max epoch number was 50. Adam was adopted as optimizer with a learning rate initially set to 1e$-$4 with exponentially decay at rate 0.95. For testing, each positive sample tier was paired with nine randomly selected negative tiers.

Evaluation Metrics 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. Apart from E, the larger the metric, the better.

Algorithm 2: Steps of Cross-Domain Alignment

```
Input: pre-trained feature extractor $FE^x$ for sketch
       and $FE^y$ for 3D shape, dataset of sketches
       $\mathcal{D}^x$ and dataset of 3D shapes $\mathcal{D}^y$, memory
       size for each class $k$, number of negtive
       classes $n$, $MaxEpoch$
Output: mapping network $Map^x$ for feature from
        sketch to common space, mapping network
        $Map^y$ for feature from 3D shape to
        common space
construct sketch feature set $\mathcal{D}^x_f$ with $FE^x$ and $\mathcal{D}^x$;
construct 3D shape feature set $\mathcal{D}^y_f$ with $FE^y$ and $\mathcal{D}^y$;
construct dataset of sketch-"3D shape" feature pairs
  $\mathcal{D}^*$ with $\mathcal{D}^x_f$ and $\mathcal{D}^y_f$;
initialize feature memory banks $\mathcal{M}^x$ and $\mathcal{M}^y$ by the
  given size $k$, with $\mathcal{D}^x_f$ and $\mathcal{D}^y_f$ respectively;
initialize mapping network $Map^x$ and $Map^y$;
for epoch = 1:$MaxEpoch$ do
    for feature pair $f^x$-$f^y$ and corresponding class
      label $c$ in Dataloader($\mathcal{D}^*$) do
        initialize the set of class labels $\mathcal{C}$;
        put the positive class label $c$ into $\mathcal{C}$;
        randomly choose $n$ negative class 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 common prototype $P^*$ of $c_i$ by
              clustering features in $\mathcal{M}^x$ and $\mathcal{M}^y$;
            put $P^*$ into $\mathcal{P}$;
        end
        calculate loss with Eq.1;
        update $Map^x$ and $Map^y$ concurrently;
        update $\mathcal{M}^x$ and $\mathcal{M}^y$ with current features $f^x$
          and $f^y$ respectively;
    end
end
```

Experiment Results on Three Benchmarks

We compared our PCL method results to those of state-of-the-art methods, including Siamese (Wang, Kang, and Li 2015), DCHML (Dai, Xie, and Fang 2018), TCL (He et al. 2018), and CGN (Dai and Liang 2020), on SHREC'13, SHREC'14, and D2O using the same settings for fairness. 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 and CGN was the same pre-trained ResNet-50. All codes used for these methods were our production, and we did not employ temporal sketch information.

In Tables 1, 2 and 3, the best retrieval performances of each method on SHREC'13, SHREC'14, and D2O are shown in boldface, respectively. Accordingly, our PCL method outperformed most extant methods on all metrics with SBSR and ZS-SBSR tasks. Our PCL performed particularly well on D2O compared with others due to the high ab-

Method	SBSR							ZS-SBSR						
	NN	FT	ST	nDCG	E	MRR	mAP	NN	FT	ST	nDCG	E	MRR	mAP
siamese	19.0	14.2	12.5	20.8	76.1	33.2	20.8	13.7	12.4	11.7	18.6	76.2	25.9	20.1
DCHML	41.7	43.3	29.5	51.0	54.7	48.9	52.5	18.3	21.1	13.6	23.8	63.2	27.8	33.9
TCL	43.3	47.7	33.7	57.2	51.5	51.9	56.6	25.0	17.8	20.3	31.4	65.0	39.0	31.7
CGN	**65.0**	52.0	30.8	60.8	53.4	**77.5**	58.8	33.3	29.4	22.5	39.8	63.4	48.4	38.6
PCL(ours)	55.0	**54.7**	**41.7**	**71.9**	**45.4**	68.7	**65.9**	**38.3**	**38.9**	**26.1**	**47.0**	**54.9**	**52.4**	**48.0**

Table 1: Sketch-3D shape retrieval performance (%) on SHREC'13

Method	SBSR							ZS-SBSR						
	NN	FT	ST	nDCG	E	MRR	mAP	NN	FT	ST	nDCG	E	MRR	mAP
siamese	18.7	12.9	12.0	20.0	76.1	31.2	20.7	10.0	8.6	8.7	13.6	78.4	21.4	17.0
DCHML	30.7	32.5	**24.6**	41.4	61.2	39.7	41.9	16.7	17.7	17.9	27.3	67.3	27.8	30.0
TCL	33.3	31.7	20.9	38.0	63.6	40.7	40.4	21.3	21.6	19.1	30.6	66.5	31.4	32.8
CGN	42.0	35.0	22.2	41.7	61.6	46.8	44.2	26.3	25.0	21.4	35.1	65.0	36.8	35.6
PCL(ours)	**43.0**	**40.3**	24.3	**45.5**	**58.1**	**48.2**	**49.1**	**33.3**	**32.8**	**22.8**	**39.9**	**57.8**	**45.5**	**43.4**

Table 2: Sketch-3D shape retrieval performance (%) on SHREC'14

Method	SBSR							ZS-SBSR						
	NN	FT	ST	nDCG	E	MRR	mAP	NN	FT	ST	nDCG	E	MRR	mAP
siamese	15.4	13.1	13.0	20.5	75.7	30.4	20.8	14.6	10.5	9.8	16.2	78.1	28.5	18.0
DCHML	21.0	24.4	23.8	36.1	62.2	34.2	36.5	12.5	14.1	15.4	22.4	72.4	27.0	23.2
TCL	20.2	25.5	23.9	36.7	63.9	37.7	34.6	16.9	15.5	14.9	23.4	72.5	29.8	24.3
CGN	32.1	33.9	27.9	45.8	60.5	50.3	41.5	21.5	14.9	**15.7**	25.0	72.2	33.7	25.0
PCL(ours)	**51.6**	**40.6**	**29.2**	**53.0**	**55.6**	**60.6**	**51.3**	**32.3**	**21.6**	12.7	**28.0**	**71.4**	**47.2**	**29.9**

Table 3: Sketch-3D shape retrieval performance (%) on D2O

straction of the QuickDraw style sketches, which were difficult to capture by classical CNN architectures designed for common photos.

Fig. 6 shows some retrieval results of our PCL method on all three datasets with both SBSR and ZS-SBSR tasks. Query sketches are listed on the left, and their retrieved top-five 3D shapes are listed on the right based on ranking order.

Further Analysis

Ablation Study We conducted four ablation studies to validate the ZS retrieval performance of our PCL method:

1. Cluster(W/A): Learn prototypes via dynamic clustering with cross-domain alignment.
2. Anchor (W/A): Fix anchor points for prototypes with cross-domain alignment.
3. PCL (O/A): PCL without cross-domain alignment.
4. PCL (W/A): PCL with cross-domain alignment.

The results are shown in Table 4, where the best metrics are shown in boldface. It is easy to see that apart from the ST on D2O, the PCL method performed far better than the cluster and fixed-anchor-point methods on all metrics on all three benchmark datasets. Furthermore, the semantic alignment was crucial for cross-domain tasks.

Temporal Sketch Information Temporal information is naturally embedded in sketches. However, no known SBSR methods (Wang, Kang, and Li 2015; Dai et al. 2017; Dai, Xie, and Fang 2018; Qi, Song, and Xiang 2018; He et al. 2018; Xu et al. 2020; Dai and Liang 2020; Xu et al. 2022)

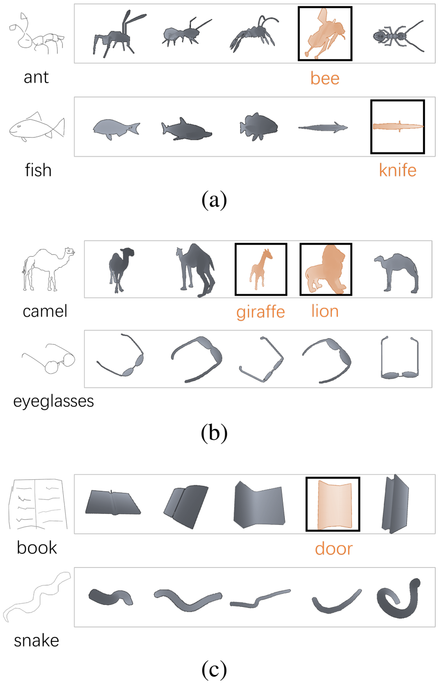


Figure 6. Retrieval results obtained by PCL: (a) SBSR and ZS-SBSR examples on SHREC'13; (b) SBSR and ZS-SBSR examples on SHREC'14; and (c) SBSR and ZS-SBSR examples on D2O. The failure cases are marked in orange and boxed up. They somehow look similar to success cases.

Dataset	Method	NN	FT	ST	nDCG	E	MRR	mAP
SHREC'13	cluster(W/A)	28.3	22.5	17.9	31.5	62.1	46.7	34.7
	anchor(W/A)	19.7	18.7	14.5	24.5	69.3	25.8	29.8
	PCL(O/A)	9.7	9.5	11.1	16.4	72.1	17.0	23.2
	PCL(W/A)	**38.3**	**38.9**	**26.1**	**47.0**	**54.9**	**52.4**	**48.0**
SHREC'14	cluster(W/A)	21.7	22.1	20.7	32.4	65.7	32.6	33.3
	anchor(W/A)	17.7	18.7	18.8	28.5	67.3	28.9	30.3
	PCL(O/A)	9.7	10.3	7.9	13.6	68.9	29.1	24.1
	PCL(W/A)	**33.3**	**32.8**	**22.8**	**39.9**	**57.8**	**45.5**	**43.4**
D2O	cluster(W/A)	21.1	18.6	15.5	25.6	71.7	32.7	26.4
	anchor(W/A)	18.0	17.9	**16.1**	25.5	71.5	28.8	26.7
	PCL(O/A)	10.1	12.4	9.4	15.1	78.8	28.0	16.5
	PCL(W/A)	**32.3**	**21.6**	12.7	**28.0**	**71.4**	**47.2**	**29.9**

Table 4: ZS-sketch 3D shape retrieval performance (%)

Task	Stroke Accumulation	NN	FT	ST	nDCG	E	MRR	mAP
SBSR	30%(W/R)	28.7	28.4	24.0	39.4	61.9	43.5	38.7
	50%(W/R)	35.4	31.5	25.5	42.5	60.1	50.9	40.6
	80%(W/R)	**52.9**	**40.7**	**30.2**	**53.8**	**55.4**	**62.2**	50.8
	100%(W/R)	49.8	**42.6**	**31.2**	**55.7**	**54.1**	60.5	**53.5**
	100%(O/R)	**51.6**	40.6	29.2	53.0	55.6	**60.6**	**51.3**
ZS-SBSR	30%(W/R)	19.5	17.8	15.4	24.9	72.7	33.5	24.8
	50%(W/R)	21.0	20.8	**16.4**	27.5	**70.9**	31.4	28.7
	80%(W/R)	**32.7**	**21.6**	14.2	**29.8**	73.4	**48.4**	28.7
	100%(W/R)	**36.9**	**23.7**	**20.3**	**33.6**	**67.9**	45.3	**30.4**
	100%(O/R)	32.3	**21.6**	12.7	28.0	71.4	**47.2**	**29.9**

Table 5: Influence (%) of temporal sketch information on D2O

use it, as sketch samples in SHREC'13 (Li et al. 2013) and SHREC'14 (Li et al. 2014b) contain only static images. Fortunately, with the help of QuickDraw (Ha and Eck 2017), the D2O dataset is available to provide temporal information of sketches. Thus we conducted experiments on D2O with employing an RNN to mimic stroke accumulation features.

Sketches were re-plotted into several cumulative stroke sub-pictures (30, 50, 80, and 100%) according to their temporal information, feeding into the RNN after producing spatial features through the CNN. To compare the "with-RNN" version (W/R), the retrieval performance of the network without entering the RNN (O/R) loop trained only by the final sketch image (stroke accumulation 100%) is also listed. Table 5 presents the results, where the top two of each metric is bolded. Owing to arbitrary free-hand sketching, 80% of the stroke accumulation is reliable at 100% and performs well enough on the network (W/R) to beat the without-RNN version (O/R), which required 100% final images. Faster retrieval of the desired 3D shape is obviously attractive to amateur users with little patience in practical situations. A retrieval example employing temporal sketch information is visualized in Fig. 7.

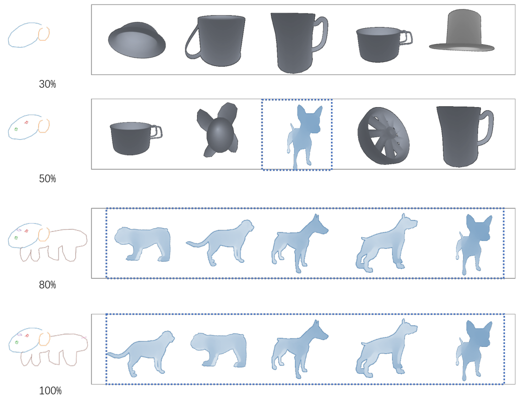


Figure 7. Example of a retrieval task employing temporal sketch information. The correct cases are marked in blue and boxed by a dashed line. An interesting point worth noting: the order of the first two retrieval results is swapped after adding the dog's tail as the last stroke.

Conclusion

This paper represents a positive step toward achieving practical ZS-SBSR tasks. Previous SBSR datasets are too scarce in some 3D shape classes and do not provide temporal sketch information. To overcome these limitations, we provided a new D2O dataset using ModelNet40 and QuickDraw and proposed a novel PCL method to increase the efficacy of cross-domain and unseen class data retrieval. Extensive experiments on the SHREC'13, SHREC'14, and the augmented D2O dataset demonstrated the power of our proposed PCL-enabled ZS-SBSR method. In a future work, we plan to conduct more experiments with data collected from real environments, considering the use of AutoML feature extractors to ease the task of customizing networks for specific modalities.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2020YFC1523204) and National Natural Science Foundation of China (U2006211 and 62171320).

References

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

Dai, G.; Xie, J.; Zhu, F.; and Fang, Y. 2017. Deep correlated metric learning for sketch-based 3d shape retrieval. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Dai, W.; and Liang, S. 2020. Cross-modal guidance network for sketch-based 3D shape retrieval. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Deng, C.; Xu, X.; Wang, H.; Yang, M.; and Tao, D. 2020. Progressive cross-modal semantic network for zero-shot sketch-based image retrieval. *IEEE Transactions on Image Processing*, 29: 8892–8902.

Dey, S.; Riba, P.; Dutta, A.; Llados, J.; and Song, Y.-Z. 2019. Doodle to search: Practical zero-shot sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2179–2188.

Dutta, A.; and Akata, Z. 2019. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5089–5098.

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

Ha, D.; and Eck, D. 2017. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, X.; Zhou, Y.; Zhou, Z.; Bai, S.; and Bai, X. 2018. Triplet-center loss for multi-view 3d object retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1945–1954.

Hu, C.; Yang, Y.; Li, Y.; Hospedales, T. M.; and Song, Y.-Z. 2021. Towards Unsupervised Sketch-based Image Retrieval. *arXiv preprint arXiv:2105.08237*.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Ji, Z.; Chai, X.; Yu, Y.; Pang, Y.; and Zhang, Z. 2020. Improved prototypical networks for few-Shot learning. *Pattern Recognition Letters*, 140: 81–87.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Li, B.; Lu, Y.; Godil, A.; Schreck, T.; Aono, M.; Johan, H.; Saavedra, J. M.; and Tashiro, S. 2013. *SHREC'13 track: large scale sketch-based 3D shape retrieval*.

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

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

Liu, Q.; Xie, L.; Wang, H.; and Yuille, A. L. 2019. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3662–3671.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Qi, A.; Song, Y.-Z.; and Xiang, T. 2018. Semantic Embedding for Sketch-Based 3D Shape Retrieval. In *BMVC*, volume 3, 11–12.

Roy, A. G.; Navab, N.; and Wachinger, C. 2018. Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In *International conference on medical image computing and computer-assisted intervention*, 421–429. Springer.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Su, H.; Maji, S.; Kalogerakis, E.; and Learned-Miller, E. 2015. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, 945–953.

Wang, F.; Kang, L.; and Li, Y. 2015. Sketch-based 3d shape retrieval using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1875–1883.

Wang, W.; Cai, Y.; and Wang, T. 2022. Multi-view dual attention network for 3D object recognition. *Neural Computing and Applications*, 34(4): 3201–3212.

Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; and Xiao, J. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1912–1920.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.

Xu, P.; Hospedales, T. M.; Yin, Q.; Song, Y.-Z.; Xiang, T.; and Wang, L. 2022. Deep learning for free-hand sketch: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

Yelamarthi, S. K.; Reddy, S. K.; Mishra, A.; and Mittal, A. 2018. A zero-shot framework for sketch based image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 300–317.

Yu, Q.; Yang, Y.; Song, Y.-Z.; Xiang, T.; and Hospedales, T. 2015. Sketch-a-net that beats humans. *arXiv preprint arXiv:1501.07873*.