Good Teacher Makes Good Student: A Discriminative-Aware Knowledge Preservation Approach for Zero-Shot Sketch-Based Image Retrieval

Haifeng Zhao [1,2,3,*], Tianjian Wu [1,3,4], Yuting Tao [1,3], and Yan Zhang [1,3]

[1] School of Software Engineering, Jinling Institute of Technology, Nanjing, China
[2] Jiangsu Hoperun Software Co. Ltd., Nanjing, China
[3] Information Analysis Engineering Research Center of Jiangsu Province, Nanjing, China
[4] School of Computer and Electronic Information, Nanjing Normal University, Nanjing, China
Email: zhf@jit.edu.cn (H.Z.), wutianjian@eswincomputing.com (T.W.), tao_yuting@jit.edu.cn (Y.T.), zy@jit.edu.cn (Y.Z.)
*Corresponding author

***Abstract*—Sketch-Based Image Retrieval (SBIR) is widely used in animation, e-commerce, and security. In these real-world applications, the classes of retrieval may be very different from the training classes, making it a zero-shot SBIR problem. Most methods in the literature resort to bridging the semantic gap between the sketch and image domains by learning a common space with a pre-trained model on a large dataset as the base network, and then fine-tuning on the target SBIR datasets. In this process, the acquired knowledge of the pre-trained model may be lost, resulting in performance degradation. To tackle this problem, we propose a teacher-student network architecture, which consists of a teacher network using the pre-trained model and a student network whose output is guided by the teacher network. Instead of introducing supplementary semantics in the teacher network, we adopt a more powerful pre-trained model as the teacher network and further enhance its discriminative capability by adding a hard-coded margin based on the prediction probability. The student network is then fine-tuned by using the teacher network's output as the learning target. Experiments on two benchmark datasets show that the proposed approach outperforms the state-of-the-art methods by more than 10%, which verifies that the prior knowledge can be better preserved by a good teacher network, which can make the student network good too.**

***Keywords*—Sketch-Based Image Retrieval (SBIR), zero-shot, knowledge preservation**

I. Introduction

Sketch-Based Image Retrieval (SBIR) is widely used in many real-world applications, such as animation, E-commerce, and security [1]. It allows searching for interesting images with a free-hand drawing as the input instead of conventional text and images.

Given a sketch query, the aim of an SBIR task is to retrieve images in the target set which have similar semantics to the query. To this end, a training set with labeled sketches and images is needed for learning the semantic relationship across the sketch and image domains. Generally, the training set and target set in an SBIR task share the common class set. That is, the classes of the retrieved images have appeared in the training set [2–4]. However, in real-world applications, it is often difficult to cover all the categories in the training set. When the categories in the target set are not in the training set, the retrieval has to rely on a single sketch given the trained model, leading to a Zero-Shot SBIR (ZS-SBIR) task. Fig. 1 shows the difference between the SBIR and ZS-SBIR tasks. One of the solutions to the ZS-SBIR problem is to learn a common embedding space of the sketch and image domains with a pre-trained model on a large dataset as the base network, and then to fine-tune the target datasets [5, 6]. In this way, the domain gap can be bridged by embedding cross-modal information in an intermediate space. However, in the model fine-tuning process, the prior knowledge acquired in the pre-trained model may be lost, resulting in performance degeneration on the target datasets [7].

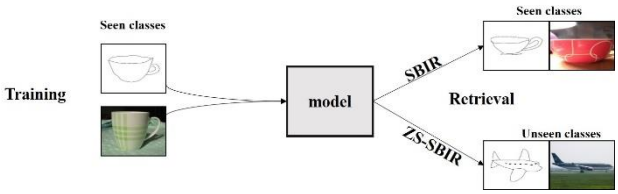

Fig. 1.The difference between SBIR and ZS-SBIR. Images retrieved in the SBIR task are in the same class set as the training set. Whereas in the ZS-SBIR, the retrieved classes are unseen in the training set.

To address this problem, Liu *et al*. [7] proposed to use a knowledge distillation [8] based approach to preserve prior knowledge from the pre-trained model by the student-

Manuscript received September 22, 2023; revised October 12, 2023; accepted October 31, 2023; published March 14, 2024.

teacher learning framework. In this framework, the student model not only learns the embedding of the sketch and image domains, but also learns prior knowledge from the teacher, which is a pre-trained model on ImageNet [9]. WordNet [10] is employed in the student-teacher learning process to provide semantic guidance.

Inspired by the idea of knowledge distillation [7], we propose a novel approach to alleviate the performance degeneration on ZS-SBIR tasks by concentrating on the discriminative ability of the model under the student-teacher learning framework. Specifically, the discriminative ability is reflected in the prediction probability distribution of the final model output. Higher prediction probability means higher discriminative ability on the target dataset. This is achieved through two phases. In the first phase, we adopt a more powerful pre-trained model as the base to improve the discriminative ability. Empirically, a stronger teacher can not always improve the ability of its students. But if the student and the teacher have similar structures, the student can learn more from the teacher. That is, more prior knowledge can be transferred from the teacher to the student [8, 11, 12]. To this end, our student model is designed to be the same as the pre-trained model except at the output layer. In the second phase, we enhance the output probability distribution by adding a hard-coded margin to improve the discriminative ability. The enhanced probability distribution then is used to guide the training of the student.

We conducted experiments on two benchmark datasets, the Sketchy Extended [2, 13] and TU-Berlin Extended [14, 15]. Our approach outperforms the State-of-the-Art (SOTA) methods by more than 10%, which verifies its effectiveness.

The contributions of this paper are threefold. Firstly, we propose a novel approach based on knowledge distillation to preserve knowledge by improving the discriminative ability of the trained model. Secondly, the discriminative ability is improved by adopting a more powerful teacher model and by adding hard-coded margins at the output probability distribution of the teacher. Thirdly, we conduct comprehensive experiments on benchmark datasets, which verifies that our method can improve the retrieval performance by a large margin compared to the SOTA methods.

The rest of our paper is organized as follows. We introduce the related work in Section II. Section III describes the details of the proposed method. The experiments and analysis are presented in Section IV. Finally, we conclude our paper in Section V.

II. Related Work

The tasks of image retrieval through sketches can be divided into three categories: Sketch-Based Image Retrieval (SBIR), Fine-Grained Sketch-Based Image Retrieval (FG-SBIR) and Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR).

SBIR is a category-level search. It can only retrieve natural images in the same category as the sketch. Besides traditional SBIR methods, deep learning methods [16, 17] better solve this problem by using common ranking loss methods.

FG-SBIR is an instance-level search. Its goal is to capture the fine-grained similarities between sketches and images, and find the exact matching images for the input sketch [13, 16]. FG-SBIR is a more challenging task compared with the SBIR task which requires only category level retrieval. Better feature extraction methods are needed to tackle this problem. Therefore, FG-SBIR tasks tend to use deep learning methods [17–19].

ZS-SBIR is a combination of the zero-shot task [20] and the SBIR task [21]. The task is challenging in two aspects. First, the categories in the retrieval phase may not be in the training phase, making it similar to the general zero-shot setting. Second, knowledge learned in the model has to cover both sketch and image domains, which is more challenging as there may be a large domain gap between sketches and images. Shen *et al*. [5] first raised this problem in 2018 and used a generative hash-based approach to model category semantics. Yelamarthi *et al*. [21] also proposed a generative model using the variational and adversarial autoencoders. As the generative models need sketches and images to be aligned as pairs, which is not always feasible, more researchers resorted to employing additional semantic information to guide the training process on the fine-tuned large model [6, 22, 23]. However, as pointed out by Liu *et al*. [7], catastrophic forgetting [24] occurs in the process of fine-tuning large pre-trained model, which can cause the performance degeneration. A knowledge distillation [8] based method is then adopted by the student-teacher learning framework using auxiliary semantics [7].

III. The Proposed Approach

In this section, we describe the details of our knowledge distillation-based approach. We first give the formal definition of the zero-shot sketch-based image retrieval task. Then we link the task to knowledge distillation.

The architecture of the proposed network is shown in Fig. 2. The sketch and image inputs are fed into two branches. The first branch is the student network which uses a CSE-ResNeXt-101 [25, 26] architecture. It learns from the true labels and predicts categorical probabilities by the benchmark loss which is the cross entropy between predicted and true labels of the original dataset. The second branch is the teacher network, which is pre-trained on ImageNet [9] with the same architecture as the student. The motivation that we choose the same model architecture for both the teacher and the student is that the student will get more knowledge from the teacher as suggested in [8, 11, 12]. That is, the student with the same teacher architecture can learn more knowledge than that of with different student-teacher architectures. Besides, to preserve more knowledge, we use a hard-coded margin to enhance the discriminative ability of the teacher output. In other words, we enhance the guidance of the teacher, making the knowledge transfer more fluent. Specifically, the outputs of the student network are compared with the enhanced soft labels of the teacher network by the

ImageNet loss to predict the ImageNet labels on the training set.

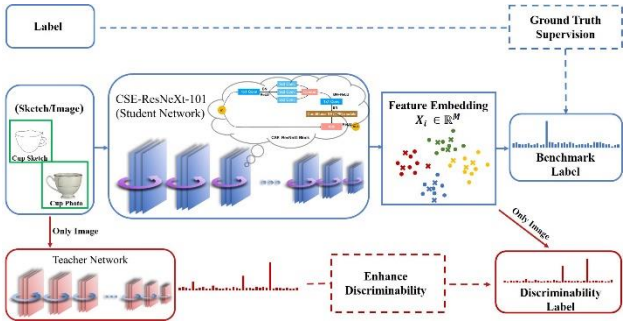

Fig. 2. The proposed network architecture. "Only Image" means we train only images on the Teacher Network, whereas we train both sketches and images on the Student Network.

A. Problem Formulation

A ZS-SBIR problem involves a source domain and a target domain. The source domain is represented by $O^S = \{P^S, S^S\}$, where P^S is an image set and S^S is a sketch set, respectively. Similarly, the target domain is represented by $O^T = \{P^T, S^T\}$, containing an image set P^T and a sketch set S^T. Let the training set of images in the source domain be

$$P^S = \{(p_i, y_i) | y_i \in C^S\}_{i=1}^{n_1} \quad (1)$$

and the training set of sketches be

$$S^S = \{(s_j, z_j) | z_j \in C^S\}_{j=1}^{n_2} \quad (2)$$

where p_i is the i-th image in the training set, y_i is the label of p_i , s_j is the j-th sketch in the training set, z_j is the label of s_j, and C^S is a set of classes in the source domain. n_1 and n_2 are the numbers of images and sketches in the training set, respectively.

In the same way, the testing set of images is

$$P^T = \{(p_i', y_i') | y_i' \in C^T\}_{i=1}^{m_1} \quad (3)$$

and the testing set of sketches is

$$S^T = \{(s_j', z_j') | z_j' \in C^T\}_{j=1}^{m_2} \quad (4)$$

where p_i', y_i', s_j', C^T, m_1 and m_2 have the same meaning as corresponding variables in O^S, but for the testing set O^T.

In the zero-shot setting

$$C^S \cap C^T = \emptyset \quad (5)$$

With the above formulation, we define the ZS-SBIR problem as follows. Given a query sample s_j' in the testing sketch set, the aim is to find the best matching p_i' in the testing image set, satisfying the constraint $z_j' = y_i'$. That is, the sketch and its corresponding image should be in the same class. This is a typical classification problem and we can train a deep network to get the discriminative features.

B. Feature Embedding

A classification system normally consists of a feature embedding module and a classifier module. In deep learning, the end-to-end scheme integrates these two modules in a whole network. For the feature embedding module, we follow [7], and adopt the CSE-ResNeXt model [25], which is based on the ResNeXt [27], as the base network. The CSE-ResNeXt model puts the data from different domains under a single framework by adding a symbol to represent which domain the data come from. This allows training to pay more attention to the common space of the sketch and image domains in order to bridge the gap between two modalities.

Let the CSE-ResNeXt model be $h(\cdot,\cdot,\theta_h)$. The sketch domain will be $f(\cdot,\theta_f) = h(\cdot, input_{domain} = 1, \theta_h)$, and the image domain will be $f(\cdot,\theta_f) = h(\cdot, input_{domain} = 0, \theta_h)$. With the CSE-ResNeXt model at hand, source features x_i can be embedded at the full-connected layers in M dimension.

C. Learning to Objectives

There are two learning objectives in our problem. One is the conventional objective in SBIR, i.e., how to learn the common space of the sketch and image modalities to bridge the domain gap. This is the benchmark classifier. The other is how to preserve pre-trained knowledge on the large fine-tuned model. This is the knowledge preserving classifier.

1) Learning in the common space

The aim to learn features in the common space is to make the examples of sketches and images in the same class close to each other. This is a classification problem. We design the learning objective by adding a SoftMax layer at the end of feature embedding layers as follows:

$$\hat{y_i} = softmax(w^T x_i + b) \quad (6)$$

where $\hat{y_i} \in \mathrm{R}^{|C^S|}$is the predicted labels in the source image domain. w^T and b are weight and bias.

The loss of the benchmark classifier is defined as the cross entropy between the predicted and true labels.

$$L_B = \frac{1}{N}\sum_i -\log\frac{\exp(w_{y_i}^T x_i + b_{y_i})}{\sum_{k \epsilon C^S}\exp(w_k^T x_k + b_k)} \quad (7)$$

where $w_{y_i}^T$ and b_{y_i} are the weight and bias of the benchmark classifier. N is the size of the training set.

2) Learning for knowledge preservation

To solve the knowledge preservation problem, we adopt a teacher-student network architecture. As the prior knowledge is in the pre-trained model, we use the pre-trained ImageNet model as the teacher network, guiding the student network by predicting the same input fed to the student network. To observe the ability of the knowledge preservation, the classifier is designed to predict the labels of the original ImageNet domain. That is, the student network is to learn the ImageNet labels for the image inputs. This is also a classification problem:

$$\tilde{y}_i = softmax(v^T x_i + c) \quad (8)$$

where $\tilde{y}_i \in \mathrm{R}^{|C^O|}$ is the predicted labels in the original ImageNet domain. v^T and c are weight and bias respectively for the ImageNet classifier.

Similar to the benchmark classifier in Section III.C.1, we also use the cross-entropy loss as the learning objective. However, the true labels of the ImageNet are not available as the sketches and images in the training set are not annotated according to the ImageNet. Therefore, the teacher network's prediction is used as the soft label indicator to guide the training of the student network. The cross-entropy loss is

$$L_T = \frac{1}{n_1}\sum_i - q_i^t \log \frac{\exp\,(v_{y_i}^T x_i + c_{y_i})}{\sum_{k\epsilon C^o} \exp\,(v_k^T x_k + c_k)} \quad (9)$$

where q_i^t is the predicted probability vector for an image sample by the teacher network, i.e., the pre-trained ImageNet model. In this way, the knowledge acquired by the teacher network can be transferred to the student network by minimizing the cross-entropy of soft ImageNet labels and predicted ImageNet labels.

3) A stronger teacher prediction model

Using the pre-trained ImageNet model as the teacher network leads to a question: which pre-trained ImageNet model should be used to guide the training? Liu *et al.* [7] used the CSE-ResNet-50 as the teacher network. This model has a top-1 accuracy of 76.71% [26]. We make the hypnosis that the more prior knowledge in the pre-trained model, the more can be preserved in the student network after fine-tuning. Here, the amount of prior knowledge can be indicated by the top-1 accuracy on the ImageNet. We use the SE-ResNet-101 [26] and SE-ResNeXt-101 [27], which have the top-1 accuracy of 77.62% and 79.30%, as the stronger teacher networks. In Section IV.E, we show that the stronger teacher leads to a better knowledge preservation ability in the student network, as it has more knowledge to preserve under the same degeneration condition.

4) Discriminative-aware prediction

Apart from a stronger teacher, we further improve the knowledge preservation ability by introducing more discriminative ability. Liu *et al*. [7] used additional semantics as the complementary for the teacher's prediction. Motivated by the idea of enhancing discriminative ability [29], we propose a new method to operate directly on the predicted vector from the ImageNet. The predicted output of the teacher network is essentially the possibility that a sample belonging to a specific class. Intuitively, the probability can be viewed as the inverse distance to the class center. The samples in the same class are supposed to be close to the class center, and far from the other class centers. In the probability vector, this is represented by a higher probability for the same class.

Cui *et al*. [1] pointed out that higher discriminative ability can push samples closer to the ones in the same class, and farther from the ones in other classes. Therefore, we propose to improve the discriminative ability of the classifier by increasing the predicted probability vector with a hard-coded margin. Specifically, we increase the maximum probability of the predicted vector by a positive margin at the factor of a, and decrease other entries of the predicted vector by a negative margin at a factor of $-b$.

Given the teacher's probability vector

$$q_i^t = [q_1^t, \cdots, q_k^t, \cdots, q_{1000}^t]^T \quad (10)$$

the enhanced probability vector is

$$\tilde{q}_i^t = [q_1^t \cdot (1-b), \cdots, q_k^t \cdot (1+a), \cdots, q_{1000}^t \cdot (1-b)]^T \quad (11)$$

After the enhanced probability vector is obtained, the cross-entropy loss of the student network becomes the discriminative vector loss

$$L_D = \frac{1}{n_1}\sum_i - \tilde{q}_i^t \log \frac{\exp\,(v_{y_i}^T x_i + c_{y_i})}{\sum_{k\epsilon C^o} \exp\,(v_k^T x_k + c_k)} \quad (12)$$

where q_i^t in Eq. (9) is replaced by $\tilde{q}_i^t$.

D. The Final Objective

The total loss of the whole student network L is the weighted sum of the benchmark loss L_B and the enhanced discriminative vector loss L_D:

$$L = L_B + \lambda_D L_D \quad (13)$$

With the total loss at hand, the whole network can be trained by minimizing the total loss of the network to produce a discriminative model.

IV. Experiments

A. Experimental Settings

We conducted experiments on two benchmark datasets, the Sketchy Extended [2, 13] and TU-Berlin Extended [5, 14].

The original Sketchy dataset [2] was collected by sketching the image objects. It consists of 12,500 images and 75,471 sketches distributed in 125 classes. The images and sketches are paired well. Liu *et al*. [13] collected another 60,502 natural images to extend the original dataset, forming a total number of 73,002 images. Examples of the dataset are shown in Fig. 3(a).

The TU-Berlin dataset [14] contains 20,000 sketches in 250 categories, with 80 sketches for each category. The extended version [15] added real images from ImageNet and Google Image search. The total collected images are 191,067 real images, and by average 764 images per category. Liu *et al*. [13] also extended it with 204,489 images[1]. Some examples are shown in Fig. 3(b). Compared to the Sketchy Extended dataset, the sketches in the TU-Berlin Extended dataset are more abstract with fewer strokes, making it a more challenging dataset.

[1]We actually found 204,070 images in the current version.

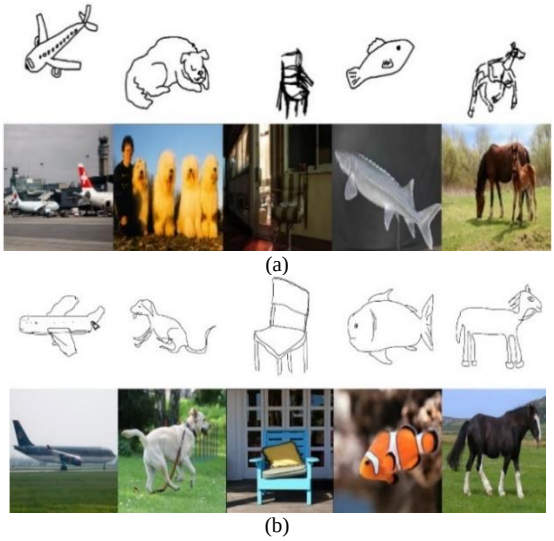

Fig. 3. Some examples from Sketchy Extended and TU-Berlin Extended datasets. (a) The Sketchy Extended; (b) The TU-Berlin Extended.

In our experiments, we followed the dataset split scheme from [5]. For the Sketchy Extended dataset, we used 100 classes for training and 25 classes for testing. We refer it as Split 1. As there may be duplicated classes in both training set and testing set, we also followed the settings from Yelamarthi *et al*. [21], where 104 classes were used for training, and the rest 21 classes were used for testing. This can make sure there are no overlapping classes in ImageNet and Sketchy Extended, making it a real zero-shot SBIR problem. We refer it as Split 2. For TU-Berlin Extended dataset, we used 30 and 120 classes for training and for testing respectively [5].

We evaluated the performance using mean average precision and precision on the top *N* retrieval examples, i.e., mAP@N and Precision@N. If the label of the retrieved image is the same as the query sketch, the retrieval is regarded as a correct query.

B. Implementation Details

We implemented our method using PyTorch [22] under Ubuntu 16.04 64-bit and CUDA 10.0. All experiments were run on a server with two Nvidia Tesla V100 cards. Each card has 32G memory.

We used Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and the loss weight $\lambda = 0.0005$. The batch size was 40, and the learning rate was set to 0.0001. The weight decay was 5×10^{-4}. Each of the input sketches and images was resized to 224×224. To feed more examples to the deep neural network, we did data augmentation. We applied affine transform to images where the transformation is determined by a random generator at the probability of 70% We also normalized each image before training by subtracting the mean and standard deviation to eliminate the effect of the variance of light conditions.

On the Sketchy Extended dataset, it took about 11 hours for 20 epochs when running on two GPUs. Each batch took about 0.75 s. On testing, it took about 10 minutes, i.e., 2.21 microseconds per sample. We trained the model on the TU-Berlin Extended dataset using two GPUs in parallel, taking about 16 h for 20 epochs, i.e., each batch took 0.75 s. The testing took about 5 min, and 4.47 ms per sample.

C. Comparison with the State-of-the-Art Methods

We compared our proposed method with the existing methods including ZSIH [5], GZS-Auto [29], GZS-VAE [29], SEM-PYPC [23], CSDB [30], OCEAN [31], PCSN [32], SAKE [7], AMDReg [33], on the Sketchy Extended Split 1 and TU-Berlin Extended datasets. Most of them are generative models. Specifically, ZSIH [5] is a zero-shot image hashing approach for generative hashing model. GZS-Auto [29] and GZS-VAE [29] are also generative models with different generators. SEM-PYPC [22] employs the idea of adversarial learning to train the model. CSDB [30] helps retrieval by generating image correspondence in the guide of image style. OCEAN [31] is an adversarial learning model in the dual learning framework. PCSN [32] is a cross-modal approach with feature alignment. SAKE [7] is the knowledge preserving model with auxiliary semantics. AMDReg [33] is an embedding framework under imbalanced conditions.

1) Results on sketchy extended split 1

As shown in Table I, our proposed method outperforms the state-of-the-art methods by at least 11.8% on the mAP@all for the Sketchy Extended Split 1. Especially, our method improves by a large margin of 12.7% compared to SAKE [7] which is the state-of-the-art knowledge distillation-based method for ZS-SBIR. Note that, methods concentrating on the feature embedding, such as PCSN [32], SAKE [7], AMDReg [33], are generally better than generative models [6, 8, 25, 36] in Table I.

TABLE I. PERFORMANCE COMPARISON WITH ALTERNATIVE APPROACHES

Method	Dimension	Sketchy Extended Split 1		TU-Berlin Extended	
		mAP@all	Precision@100	mAP@all	Precision@100
ZSIH [5]	64	0.254	0.340	0.220	0.291
GZS-Auto [30]	2048	0.253	0.305	0.187	0.281
GZS-VAE [30]	2048	0.289	0.358	0.238	0.334
SEM-PYPC [22]	64	0.349	0.463	0.297	0.426
CSDB [31]	-	0.375	0.484	0.254	0.355
OCEAN [32]	512	0.462	0.590	0.333	0.467
PCSN [33]	-	0.523	0.616	0.424	0.517
SAKE [7]	512	0.547	0.692	0.475	0.599
AMDReg [34]	512	0.551	0.715	0.447	0.574
Ours	**512**	**0.669**	**0.768**	**0.544**	**0.628**

Our approach is basically a feature embedding-based method, as improving the discriminative ability of the classifier is essentially the learning of discriminative feature embedding in the common space. Additionally, our method is based on the knowledge distillation which can transfer more knowledge to the target model making our model better than the existing models.

2) Results on TU-Berlin Extended dataset

Regarding the TU-Berlin Extended dataset, our method beats all other methods in Table I at both mAP@all and Precision@100. The improvement at mAP@all is 6.9% compared to the state-of-the-art method SAKE [7]. This improvement is less than that of on the Sketchy Extended Split 1 dataset. The reason may be that TU-Berlin Extended is more difficult than Sketchy Extended. The categories in TU-Berlin Extended are not totally independent. One category may be a generalization of another category, e.g., tree v.s. palm tree. As mentioned in Section IV.A, sketches in TU-Berlin Extended have fewer strokes to describe objects leading to more unambiguous when matching with images. The performance on existing methods in Table I also verifies this. None of them has a better mAP@all on TU-Berlin Extended than Sketchy Extended. Our approach performs best among them, which verifies the effectiveness of our proposed discriminative enhancement framework.

As Sketchy Extended Split 1 is not a truly ZS-SBIR dataset, we also ran experiments on the Sketchy Extended Split 2 dataset. The competitors are CAAE [21], CVAE [21], CSDB [30], Dooble2Search [6], SAKE [7], and SketchGCN [35]. Except for CSDB [30] and SAKE [7] which are the same as that of on Sketchy Extended Split 1, these methods are only used in true ZS-SBIR tasks. CAAE [21] and CVAE [22] are conditional generative models with variational autoencoders and adversarial autoencoders. Dooble2Search [6] employs multiple loss functions to model both the category semantics and cross-modal gap. SketchGCN [35] is a graph convolutional network based model.

3) Results on sketchy extended split 2

We compared mAP@all and Precision@100 in the same way as Sketchy Extended Split 1. As some of the methods only reported their results on mAP@200 and Precision@200, we also conducted experiments on these settings. The results are shown in Table II. It can be seen that our method outperforms the existing ones by a large margin. The only exception is SketchGCN [35] on mAP@200, in which our method is slightly lower. The reason may be that our method retrieves most best-match examples at the top 100, leaving the least mismatched examples lower score after that. Note that SAKE [7] which is a knowledge distillation-based method outperforms other generative models [21, 31] and multiple loss method [6]. Our approach outperforms SAKE [7] by a little bit at mAP@200. This verifies that the knowledge distillation is a better framework for the ZS-SBIR task. And our discriminative enhancement scheme is a better alternative under this framework.

TABLE II. RESULTS ON SKETCHY EXTENDED SPLIT 2 (21 CLASSES)

Method	Dimension	mAP@all	Precision@100	mAP@200	Precision@200
CAAE [21]	4096	-	-	0.156	0.260
CVAE [21]	4096	-	-	0.225	0.333
CSDB [30]	-	-	-	0.358	0.400
Doodle [6]	64	0.369	-	0.461	0.370
SAKE [7]	512	-	-	0.497	0.598
SketchGCN [35]	2048	0.382	0.538	**0.568**	0.487
Ours	512	**0.536**	**0.639**	0.508	**0.604**

D. Discussion on the Parameters

To measure the influence of the hard-coded margins for different values, we conducted experiments on the Sketchy Extended Split 1 dataset as [7]. As shown in Table III, With the positive margin a increasing, the mAP first rises and then drops. It is also held for the negative margin b. The reason is that the positive margin represents the intra-class similarity or inverse distance. Increasing the margin means making the example close to the class center, which makes the class more compact and improves the performance. On the other hand, increasing the negative margin is equivalent to making the example far from the class center, which results in the more intra-class distance. This can improve performance initially. But as the example is pushed too far to the region of class, a false prediction will be made, which decreases the performance.

It is important to choose a proper positive and negative margin pair. Empirically, from the results of our experiments, we can see that a small margin is enough for the performance improvement, i.e., we get the best result at (a = 0.1, b = 0.01) margin pair. We also conducted the extreme condition that the margin exceeds its ranges, i.e., a = 1. The results show that the performance drops a lot at all the negative margin values. The reason is that the examples are pushed far beyond their class regions. This also implies that the output probability distribution of the pre-trained model contains rich information of prior knowledge. Our method can persevere more knowledge from the pre-trained model.

TABLE III. MAP@ALL FOR DIFFERENT HARD-CODED MARGIN VALUES. a MEANS POSITIVE MARGIN, b MEANS NEGATIVE MARGIN

a	b			
	0	**0.01**	**0.03**	**1**
0	0.651	0.651	0.656	0.648
0.1	0.658	0.669	0.660	0.647
0.3	0.656	0.662	0.654	0.648
1	0.646	0.620	0.627	0.641

We also conducted experiments on λ_D, which controls the balance of the benchmark loss and the discriminative loss. The results are in Table IV. It can be noted that as λ_D

becomes larger but not too large, the result becomes better. This can be explained as the discriminative component has more influence on the final decision of the whole model.

TABLE IV. MAP@ALL AND P@100 FOR DIFFERENT λ_D VALUES

λ_D	map@all	P@100
0	0.485	0.637
0.1	0.518	0.665
0.3	0.588	0.721
1	0.669	0.768
3	0.662	0.747

When it becomes too large, it will make the benchmark and discriminative loss unbalanced, which leads to performance decrease.

E. Ablation Study

To analyze how much contribution the stronger teacher network makes towards the final prediction precision, we conducted an ablation analysis. Firstly, we only used the student model. As shown in Table V, the performance is the worst without any teacher information. After adding teacher information, the performance increases significantly. It implies that the teacher's information is useful for the knowledge preservation. We then added the discriminative enhancement module, which further improved the performance. This verifies that our enhancement is effective. Additionally, we tested teachers using CSE-ResNet-50 [26], CSE-ResNet-101 [26], and CSE-ResNeXt-101 [25, 36]. Table V shows that the discriminative capability of the student model follows that of the teacher model. This proves that a good teacher makes a good student.

TABLE V. ABLATION ON THE SKETCHY EXTENDED DATASET FOR MAP@ALL. CSER, SER, CSERN AND SEERN STAND FOR CSE-RESNET, SE- RESNET, CSE-RESNEXT, AND SE-RESNEXT, RESPECTIVELY

Student	Teacher	L_B	$L_B + L_T$	$L_B + L_D$
CSER-50	SER-50	0.423	0.544	0.550
CSER-101	CSER-101	0.431	0.561	0.574
CSERN-101	SERN-101	0.485	0.651	0.669

F. Qualitative Analysis

We plot some retrieval results with distance values obtained on the TU-Berline Extended dataset in Fig. 4.

Fig. 4. Examples of retrieved result of some sketches from the TU-Berline Extended dataset.

It can be seen that most of the images are in the same category as the query sketch. Some misclassified images are shown with a red box. The first four rows show the correct retrievals, and the bottom two rows show the retrievals with some errors. It can be observed that the correct ones have lower distance values, i.e., under 0.9. The false predictions have larger distances. Although incorrect, the misclassified images are very similar to the query, even they are not in the same class. For instance, the box with a cross sign looks similar to the Wind turbine.

V. CONCLUSION

In this paper, we have introduced a novel approach for zero-shot sketch-based image retrieval. In our approach, knowledge preservation plays an important role in the classification, which is implemented via a student-teacher network structure based on knowledge distillation. The stronger teacher was used to increase the amount of knowledge transfer. Moreover, a discriminative ability enhancement scheme was proposed to further improve the performance. We conducted experiments on two benchmark datasets, and showed the effectiveness of our method. In the future, we will try to use the GAN (generative adversarial network) to better bridge the cross-domain gap.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

HZ conducted the research, designed experiments, analyzed the data, and wrote the paper; TW collected data, ran the experiments, analyzed the data, and wrote the first version; YT wrote the paper and analyzed data; YZ proposed the topic, wrote the paper, provided support. All authors had approved the final version.

FUNDING

This research was funded by the International Science and Technology Cooperation Project of Jiangsu Province (BZ2020069), and the Major Program of University Natural Science Research of Jiangsu Province (21KJA520001).

REFERENCES

[1] S. Ouyang, T. M. Hospedales, Y.-Z. Song, and X. Li, "ForgetMeNot: Memory-aware forensic facial sketch matching," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[2] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: Learning to retrieve badly drawn bunnies," *ACM Transactions on Graphics (SIGGRAPH)*, 2016.

[3] J. M. Saavedra and J. M. Barrios, "Sketch based image retrieval using Learned KeyShapes (LKS)," in *Proc. the British Machine Vision Conference* (BMVC), Sep. 2015.

[4] K. Pang, K. Li, Y. Yang, H. Zhang, T. M. Hospedales, T. Xiang, and Y.-Z. Song, "Generalising fine-grained sketch-based image retrieval," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 677–686.

[5] Y. Shen, L. Liu, F. Shen, and L. Shao, "Zero-shot sketch-image hashing," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3598–3607.
[6] S. Dey, P. Riba, A. Dutta, J. Llados, and Y.-Z. Song, "Doodle to search: Practical zero-shot sketch-based image retrieval," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
[7] Q. Liu, L. Xie, H. Wang, and A. L. Yuille, "Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval," in *Proc. the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 3661–3670.
[8] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint, arXiv:1503.02531, 2015.
[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: A large-scale hierarchical image database," in *2009 Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
[10] C. Fellbaum, *WordNet: An Electronic Lexical Database*, Cambridge, MA: MIT Press, 1998.
[11] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, 2021.
[12] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation," in *Proc. the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 1921–1930.
[13] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao, "Deep sketch hashing: Fast free-hand sketch-based image retrieval," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognitio*n (CVPR), July 2017.
[14] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" *ACM Transactions on Graphics (SIGGRAPH)*, vol. 31, no. 4, pp. 1–44, 2012.
[15] H. Zhang, S. Liu, C. Zhang, W. Ren, R. Wang, and X. Cao, "SketchNet: Sketch classification with web images," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[16] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C.-C. Loy, "Sketch me that shoe," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
[17] J. Xie, G. Dai, F. Zhu, and Y. Fang, "Learning barycentric representations of 3D shapes for sketch-based 3D shape retrieval," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
[18] J. Zhang, F. Shen, L. Liu, F. Zhu, M. Yu, L. Shao, H. T. Shen, and L. V. Gool, "Generative domain-migration hashing for sketch-to-image retrieval," in *Proc. the European Conference on Computer Vision (ECCV)*, Sep 2018.
[19] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in *Proc. the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
[20] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 09, pp. 2251–2265, Sep. 2019.
[21] S. K. Yelamarthi, M. S. K. Reddy, A. Mishra, and A. Mittal, "A zero-shot framework for sketch based image retrieval," in *Proc. the European Conference on Computer Vision (ECCV)*, Springer, 2018, pp. 316–333.
[22] A. Dutta and Z. Akata, "Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
[23] X. Xu, M. Yang, Y. Yang, and H. Wang, "Progressive domain-independent feature decomposition network for zero-shot sketch-based image retrieval," in *Proc. the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, 2020, pp. 984–990.
[24] R. M French, *Catastrophic Forgetting in Connectionist Networks*, John Wiley Sons, Ltd, 2006.
[25] P. Lu, G. Huang, Y. Fu, G. Guo, and H. Lin, "Learning large Euclidean margin for sketch-based image retrieval," arXiv preprint, arXiv:1812.04275, 2018.
[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
[27] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5987–5995.
[28] S. Cui, S. Wang, J. Zhuo, L. Li, Q. Huang, and Q. Tian, "Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3941–3950.
[29] V. K. Verma, A. Mishra, A. Mishra, and P. Rai, "Generative model for zero-shot sketch-based image retrieval," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 704–713.
[30] T. Dutta, A. Singh, and S. Biswas, "Styleguide: Zero-shot sketch-based image retrieval using style-guided image generation," *IEEE Transactions on Multimedia*, 2020.
[31] J. Zhu, X. Xu, F. Shen, R. K.-W. Lee, Z. Wang, and H. T. Shen, "Ocean: A dual learning approach for generalized zero-shot sketch-based image retrieval," in *Proc. the IEEE International Conference on Multimedia and Expo (ICME)*, 2020, pp. 1–6.
[32] C. Deng, X. Xu, H. Wang, M. Yang, and D. Tao, "Progressive cross-modal semantic network for zero-shot sketch-based image retrieval," *IEEE Transactions on Image Processing*, vol. 29, pp. 8892–8902, 2020.
[33] T. Dutta, A. Singh, and S. Biswas, "Adaptive margin diversity regularizer for handling data imbalance in zero-shot sbir," in *Proc. the European Conference on Computer Vision (ECCV)*, Springer, 2020, pp. 349–364.
[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," arXiv preprint, arXiv:1912.01703, 2019.
[35] Z. Zhang, Y. Zhang, R. Feng, T. Zhang, and W. Fan, "Zero-shot sketch-based image retrieval via graph convolution network," in *Proc. the AAAI Conference on Artificial Intelligence (AAAI)*, 2020, pp. 12943–12950.
[36] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.