Progressive Cross-Modal Semantic Network for Zero-Shot Sketch-Based Image Retrieval

Cheng Deng, *Senior Member, IEEE*, Xinxun Xu, Hao Wang, Muli Yang, and Dacheng Tao, *Fellow, IEEE*

***Abstract*— Zero-shot sketch-based image retrieval (ZS-SBIR) is a specific cross-modal retrieval task that involves searching natural images through the use of free-hand sketches under the zero-shot scenario. Most previous methods project the sketch and image features into a low-dimensional common space for efficient retrieval, and meantime align the projected features to their semantic features (e.g., category-level word vectors) in order to transfer knowledge from seen to unseen classes. However, the projection and alignment are always coupled; as a result, there is a lack of explicit alignment that consequently leads to unsatisfactory zero-shot retrieval performance. To address this issue, we propose a novel progressive cross-modal semantic network. More specifically, it first explicitly aligns the sketch and image features to semantic features, then projects the aligned features to a common space for subsequent retrieval. We further employ cross-reconstruction loss to encourage the aligned features to capture complete knowledge about the two modalities, along with multi-modal Euclidean loss that guarantees similarity between the retrieval features from a sketch-image pair. Extensive experiments conducted on two popular large-scale datasets demonstrate that our proposed approach outperforms state-of-the-art competitors to a remarkable extent: by more than 3% on the Sketchy dataset and about 6% on the TU-Berlin dataset in terms of retrieval accuracy.**

***Index Terms*— Zero-shot learning, sketch-based image retrieval, progressive generation.**

I. Introduction

DUE to explosive growth of image content on the Internet, image retrieval has come to play an important role in many fields, including e-commerce, medical diagnosis, and remote sensing. Conventional image retrieval methods [1]–[3] require providing textual descriptions that are in many cases difficult to obtain. On mobile devices, image retrieval with free-hand sketches, where target candidates are illustrated visually and concisely, has attracted widespread attention and formed the basis of Sketch-Based Image Retrieval (SBIR). It is difficult to guarantee that all categories are able to be trained in realistic scenarios, which results in unsatisfactory performance when tested on unseen categories. Therefore, a more realistic setting has emerged, namely ZS-SBIR, which combines zero-shot learning (ZSL) and SBIR. However, ZS-SBIR is extremely challenging, since it deals simultaneously with the significant domain gap, large intra-class variances, and limited knowledge about the unseen classes.

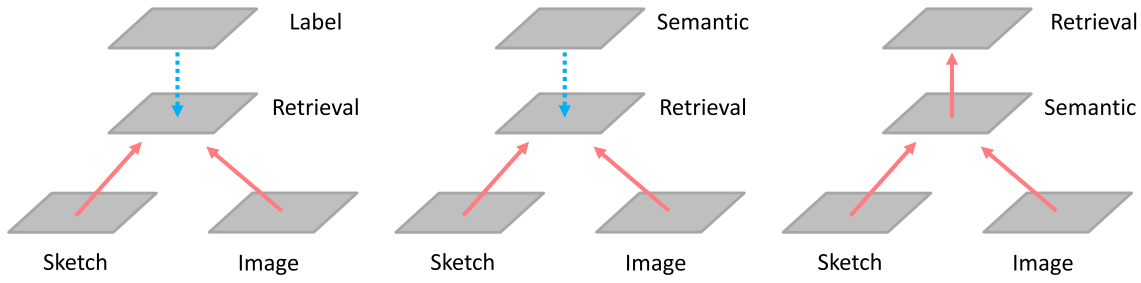
This kind of problem (involving multi-modal data) is of great research significance in the field of computer vision. One popular solution is projecting multi-modal data into a common space. Following this vein, a number of research works have been proposed. For example, [4]–[6] sought to achieve better metric distance between multi-modal data. I2LT [7] learned an effective and robust projection by jointly considering intermodal and intramodal label transfers, which builds a bridge to align different modalities for extremely rare or unseen classes. Similarly, previous ZS-SBIR works have attempted to solve this problem through projecting sketch and image features to a low-dimensional common space for retrieval, then adopting label information or category-level word vectors in order to constrain the relationship between the projected features, as shown in Figure 1(a) and Figure 1(b). Among them, label information lacks the ability to conduct relation modeling among categories, meaning that semantic knowledge cannot be bridged from training categories to test ones. Therefore, word vectors, with their ability to model the relationships among categories, have attracted ever-increasing attention [8]–[10]. In order to constrain the relationships between the projected features in a low-dimensional space, most existing methods simultaneously project sketch/image features and label/semantic supervision to a low-dimensional common space. However, this type of operation deteriorates the original semantic knowledge, since the low-dimensional projection lacks explicit alignment; that is, the projected features are not aligned to the original word vectors. Moreover, previous works have guaranteed only the mapping of the sketch or image modality to a common semantic space and their translation back to the original modality, while ignoring the relationship between the projected features of the current modality and the corresponding modality, thereby rendering the knowledge of projected features insufficient. To overcome the above drawbacks, we argue that only when the projected features have been explicitly aligned in semantic space will they be beneficial to the generation of more effective retrieval features under the zero-shot scenario. Hence, we present our progressive solution characterized by first aligning and then decoding, as shown in Figure 1(c). Moreover, since the projected features should possess the knowledge of two modalities, we propose a cross-reconstruction loss to ensure that the projected features can reconstruct not only their own modality but also the corresponding modality.

Manuscript received December 2, 2019; revised May 10, 2020 and August 24, 2020; accepted August 25, 2020. Date of publication September 10, 2020; date of current version September 18, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFE0104100 and Grant 2016YFE0200400. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Guo-Jun Qi. *(Corresponding author: Cheng Deng.)*

Cheng Deng, Xinxun Xu, Hao Wang, and Muli Yang are with the School of Electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: chdeng.xd@gmail.com; xinxun.xu@gmail.com; haowang.xidian@gmail.com; muliyang.xd@gmail.com).

Dacheng Tao is with the Faculty of Engineering, School of Computer Science, The University of Sydney, Sydney, NSW 2008, Australia (e-mail: dacheng.tao@sydney.edu.au).

Digital Object Identifier 10.1109/TIP.2020.3020383



(a) (b) (c)

Fig. 1. Three different ways to align sketch and image features into the common space. The methods illustrated in Figure 1(a) and Figure 1(b) project sketch and image features into a low-dimensional common space for efficient retrieval, while also utilizing the label or word vectors to constrain the relation between the projected features. Figure 1(c) illustrates our method, which first explicitly aligns the sketch and image features to word vectors, then projects them into a common retrieval space.

Accordingly, in this article, we propose a two-branch progressive cross-modal semantic network for the ZS-SBIR task: here, each branch first explicitly aligns the sketch and image features to the semantic features in an adversarial manner, then utilizes the decoder to project them into a common retrieval space. This strategy can align sketch and image features with the word vectors, which is conducive to the transfer of knowledge under the zero-shot scenario. Moreover, since the semantic features should encode complete knowledge about the two modalities, the cross-reconstruction loss is introduced to ensure that the semantic features can reconstruct the input of not only the current original modality but also the corresponding modality. In addition, imposing classification loss on retrieval features ensures that the retrieval features will be discriminative, while the multi-modal Euclidean loss can ensure that the retrieval features of the same class from different modalities will be more similar. It should be noted that the parameters of decoders in each branch are shared in order to alleviate over-fitting.

The main contributions of this work can be summarized as follows:

- We propose a progressive projection to effectively solve the knowledge loss problem that arises due to the lack of explicit alignment, which is conducive to knowledge transfer.
- The cross-reconstruction loss ensures that the semantic features possess complete knowledge concerning the modalities of sketch and image, which enables the problem of limited knowledge to be solved.
- Extensive experiments on two popular large-scale datasets demonstrate that our proposed approach greatly outperforms state-of-the-art methods by more than 3% on Sketchy [11] and about 6% on TU-Berlin [12] in terms of retrieval accuracy.

II. Related Work

In this section, we briefly review the prior literature in the fields of SBIR, ZSL and ZS-SBIR.

A. Sketch-Based Image Retrieval

The domain gap between the sketch and image representations is the main challenge in SBIR. Accordingly, existing approaches mostly focus on bridging the domain gap between sketch and image. Approaches to this can be broadly divided into two categories: hand-crafted features-based methods and deep learning-based ones. The hand-crafted features-based methods mostly attempt to bridge the domain gap by using edge-maps extracted from images; examples include the gradient field HOG descriptor [13], histogram of oriented edges [14], and Learned Key Shapes (LKS) [15]. Moreover, with the development of deep learning, convolutional neural networks (CNNs) have become widely utilized in computer vision fields. Yu *et al.* [16] first attempted to use CNN for sketch classification, obtaining better feature representation for both sketch and image. Moreover, Siamese architecture [17] can obtain a better metric of retrieval distance, which is beneficial for conducting retrieval. Triplet ranking loss [11] has also been adopted for coarse-grained SBIR, which obtains an effective metric by pulling the same category closer and pushing the different categories further away. Recently, CPRL [18] proposed a novel cross-modality representation learning paradigm with coupled dictionary learning to obtain robust cross-domain representations for SBIR. For instance-level SBIR, moreover a multi-scale, multi-channel, deep neural network framework [19] has been specifically designed to accommodate the unique characteristics of sketches, including multiple levels of abstraction. In this article, we propose a two-branch progressive cross-modal semantic network, in which each branch projects features to a common semantic space via adversarial training. In addition, the projected features are regularized to translate back to not only their own modality, but also the other corresponding modality, which is an effective means of reducing the domain gap between sketch and image.

B. Zero-Shot Learning

Due to the high cost of data collection and annotation, zero-shot learning has attracted extensive attention in many fields, such as image tagging [20], cross-modal retrieval [21], and action recognition [22]. Existing zero-shot approaches can be classified into two categories: namely, embedding-based and generative-based approaches. Some approaches in the first category learn non-linear multi-modal embedding [23]–[27];

most of these focus on learning non-linear mapping from the image space to the semantic space. As for generative-based approaches, a conditional generative moment matching network [28] has been proposed to synthesize the features of unseen categories. Semantics-Preserving Adversarial Embedding Network (SP-AEN) [10] has been proposed to preserve semantic information during image feature synthesis, which is conducive to ensuring that the synthesized features possess more knowledge. Moreover, side information is required under the zero-shot scenario, such that the knowledge can be learned from seen classes and transferred to unseen classes. One popular form of side information is attributes [29], which require costly expert annotation. Therefore, some studies [26], [30], [31] have utilized other auxiliary information, such as text-based [32] or hierarchical models [33] for label embedding. In this work, the side information is combined from two different semantic knowledge models; this ensures that the whole model can more effectively transfer knowledge from seen to unseen classes under the zero-shot scenario.

C. Zero-Shot Sketch-Based Image Retrieval

In recent years, ZS-SBIR, a combination of ZSL with SBIR, has attracted an ever-increasing amount of research interest. The first ZS-SBIR work [34] consists of sketch and image binary encoders, utilizing a multi-modal learning network to mitigate heterogeneity between two different modalities. CVAE [35] attempted to synthesize the image features of corresponding sketches by utilizing conditional variational auto-encoders, then conducted retrieval on the image aspect. A recent work, SEM-PCYC [36], proposed a semantically aligned cycle-consistent generative model that maintains a cycle consistency that only requires supervision at category levels. Moreover, the content-style decomposition-based model [37] utilized the concept of content-style separation for the ZS-SBIR task, disentangling the original data representations into semantic-aware domain-invariant content and data-specific variations. By contrast, we propose a progressive way of mapping visual features to a common semantic space, first through training with a common discriminator, then by decoding the semantic features to obtain the retrieval features. The cross-reconstruction loss imposed on each branch allows the semantic features to possess more complete knowledge.

III. Methodology

Given a particular sketch, our goal is to retrieve the corresponding images from the natural image gallery under the zero-shot setting; that is, the sketch categories of the training set and the test set are disjoint.

We first provide a formal definition of the ZS-SBIR task. Let $D_{tr} = \{(x_i^{ske}, x_i^{img}, s_i^{seen}, y_i)\}_{i=1}^{N_s}$ be the training set with N_s samples and $D_{te} = \{(x_i^{ske}, y_i)\}_{i=N_s+1}^{N_s+N_u}$ be the test set with N_u samples, where x_i^{ske}, x_i^{img}, s_i^{seen} and y_i are the sketch, natural image, semantic knowledge and label respectively. Their corresponding label spaces are $\mathcal{Y}_{train} = \{1, 2, 3, \ldots, C_1\}$ and $\mathcal{Y}_{test} = \{C_1 + 1, C_1 + 2, \ldots, C_1 + C_2\}$, which satisfy the zero-shot setting $\mathcal{Y}_{train} \cap \mathcal{Y}_{test} = \varnothing$. Therefore, the sketch and image data from the seen categories are used only for training. Moreover, $\mathcal{S} = \{s_i^{seen}\}_{i=1}^{N_s}$ is the set of side information. At the test stage, given an x_i^{ske} taken from D_{te}, the objective of ZS-SBIR is to retrieve the corresponding natural images from the test image retrieval gallery.

The architecture of our proposed model is illustrated in Figure 2. Our model consists of a semantic knowledge embedding for providing side information and a progressive cross-modal network to synthesize retrieval features. In order to solve the problem of knowledge loss arising due to a lack of explicit alignment, the progressive cross-modal network first obtains the semantic features by aligning the projected features to the word vectors in an adversarial fashion, then decodes the semantic features to obtain the retrieval features. The cross-cycle consistency constraint on each branch ensures that the sketch or image modality are mapped to a common semantic space, then subsequently translated back to not only the original modality but also the corresponding modality; this ensures that the semantic features possess more complete knowledge concerning both modalities. Moreover, sharing the parameters between retrieval feature decoders can alleviate over-fitting. In addition, imposing a classification loss and multi-modal Euclidean loss on the retrieval features allows for the generation of highly discriminative features. The main goal of our model is to learn three mapping functions: $G_{\theta_{ske}}(\cdot)$, $G_{\theta_{img}}(\cdot)$ and the retrieval feature decoder function $\hat{D}_{\theta_{\hat{D}}}(\cdot)$.

A. Semantic Feature Generation

In zero-shot learning, it is important to provide semantic information that can act as knowledge supervision when learning semantic features. Our proposed model utilizes text-based embedding and hierarchical embedding to provide such supervision.

1) Semantic Knowledge Embedding: In this article, we adopt two widely-used text-based models to obtain text representations: namely, Word2Vec [32] and GloVe [38]. Word2Vec can map words into vector space, where the relationships between words can be built. Compared with the bag-of-words model and TF-IDF model, Word2Vec can better capture the semantic information of words and measure the similarity between words. Moreover, GloVe constructs a word co-occurrence matrix based on a corpus, then learns word vectors based on a word co-occurrence matrix. For hierarchical embedding, we adopt the hierarchical dictionary WordNet[1] to obtain the semantic similarity. Under the zero-shot setting, we only consider the seen classes when measuring the semantic similarity between words in the hierarchy model. The hierarchical embedding for the Sketchy and TU-Berlin datasets therefore contains 354 and 664 nodes, respectively.

2) Generative Adversarial Mechanism: As illustrated in Figure 2, each branch contains a generator and a common discriminator. Taking a training sketch-image pair as an example, their features are extracted from a VGG16 [39] network pre-trained on ImageNet [40] dataset (before the last pooling layer). The goal of adversarial learning is to learn semantic features in an adversarial fashion, which means

[1] https://wordnet.princeton.edu/

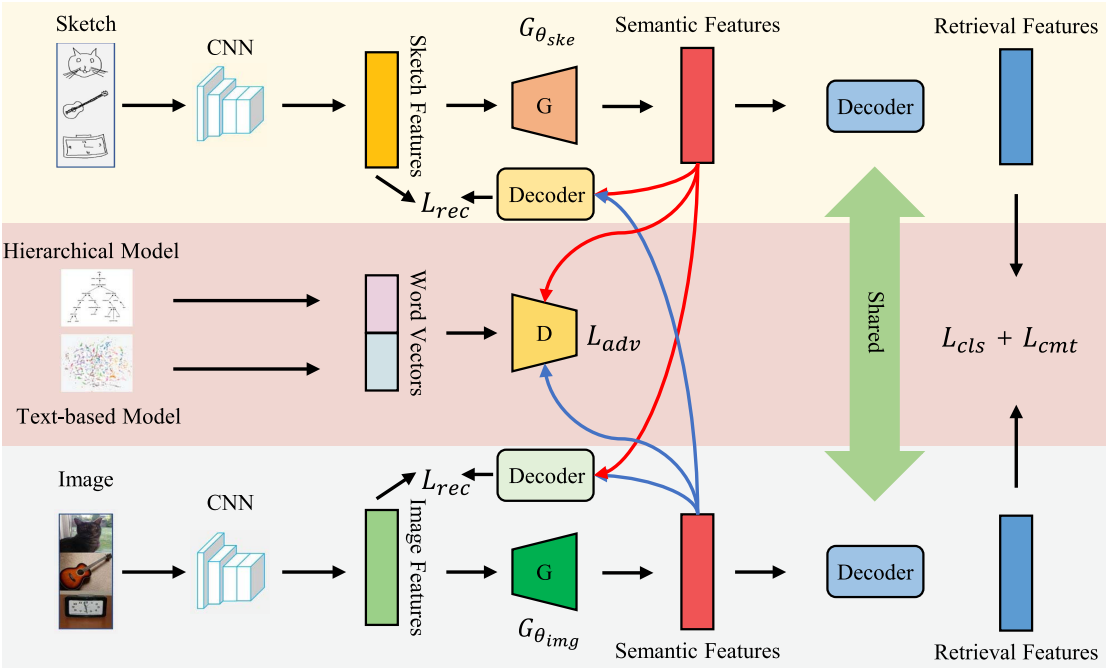


Fig. 2. Our network generates retrieval features in a progressive way. First, each branch aligns the sketch and image features to the semantic space in an adversarial fashion. The word vectors based on text and hierarchical models produce a semantic representation that serves as a true example to the discriminator. Meanwhile, the cross-reconstruction loss is beneficial to improving the high-level knowledge representation of the semantic features. Subsequently, the decoders with shared parameters take the semantic features as input and obtain the retrieval features. The classification loss and multi-modal Euclidean loss are utilized to regularize the generation of retrieval features.

that the semantic features are expected to be similar to the word vectors by 'fooling' the discriminator D_{θ_D}. Specifically, the objective can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{adv} = 2 \times & \mathbb{E}_{s^{seen}}(\log D_{\theta_D}(s^{seen})) \\ & + \mathbb{E}_{x^{ske}}(\log[1 - D_{\theta_D}(G_{\theta_{ske}}(x^{ske}))]) \\ & + \mathbb{E}_{x^{img}}(\log[1 - D_{\theta_D}(G_{\theta_{img}}(x^{img}))]),\end{aligned} \tag{1}$$

where x^{ske}, x^{img}, s^{seen}, $G_{\theta_{ske}}(\cdot)$, $G_{\theta_{img}}(\cdot)$ and $D_{\theta_D}(\cdot)$ denote the sketch features, image features, word vectors with semantic knowledge, sketch semantic generation function, image semantic generation function and discriminator function, respectively. Moreover, the sketch semantic generation network $G_{\theta_{ske}}(\cdot)$, image generation network $G_{\theta_{img}}(\cdot)$ and discriminator network $D_{\theta_D}(\cdot)$ are parameterized by θ_{ske}, θ_{img} and θ_D. Here, $G_{\theta_{ske}}(\cdot)$ and $G_{\theta_{img}}(\cdot)$ minimize the objective, while the opponent $D_{\theta_D}(\cdot)$ tries to maximize it.

3) Cross-Reconstruction Constraint: Although mapping sketch features and image features to a common semantic space by means of a generative adversarial mechanism effectively reduces the domain gap and the intra-class variances, the semantic features do not guaranteed that the input x^{ske}/x^{img} and the output x^{sem} are matched well, which is not conducive to knowledge transfer. Since the semantic features learned from each branch belong to the same semantic space, they ought to be able to reconstruct both the original sketch and the image features well. To this end, two decoders are designed to decode the semantic features in order to reconstruct both the original sketch and image features. Subsequently, we introduce the cross-reconstruction loss to ensure that the reconstructed features will be similar to the original features. Therefore, the cross-reconstruction losses in sketch and image branches can be formulated as follows:

$$\mathcal{L}_{rec\_ske} = ||\tilde{x}^{ske}_{ske} - x^{ske}||_2^2 + ||\tilde{x}^{ske}_{img} - x^{ske}||_2^2, \tag{2}$$

$$\mathcal{L}_{rec\_img} = ||\tilde{x}^{img}_{ske} - x^{img}||_2^2 + ||\tilde{x}^{img}_{img} - x^{img}||_2^2, \tag{3}$$

where x^{ske} denotes the sketch features; x^{img} denotes the natural image features; $\tilde{x}^{ske}_{ske}$ denotes the reconstructed sketch features based on the semantic features learned from the sketch branch; $\tilde{x}^{ske}_{img}$ denotes the reconstructed sketch features based on the semantic features learned from the image branch; $\tilde{x}^{img}_{ske}$ denotes the reconstructed image features based on the semantic features learned from the sketch branch, and $\tilde{x}^{img}_{img}$ denotes the reconstructed image features based on the semantic features learned from the image branch. The above features can be depicted as follows:

$$\tilde{x}^{ske}_{ske} = R_{\theta'_{ske}}(x^{sem}_{ske}), \tag{4}$$

$$\tilde{x}^{ske}_{img} = R_{\theta'_{ske}}(x^{sem}_{img}), \tag{5}$$

$$\tilde{x}^{img}_{ske} = R_{\theta'_{img}}(x^{sem}_{ske}), \tag{6}$$

$$\tilde{x}^{img}_{img} = R_{\theta'_{img}}(x^{sem}_{img}), \tag{7}$$

where $R_{\theta'_{ske}}(\cdot)$ and $R_{\theta'_{img}}(\cdot)$ denote the reconstruction function on the sketch branch and the image branch respectively; x^{sem}_{ske} and x^{sem}_{img} stand for the semantic features learned from the sketch branch and image branch respectively. The total cross-reconstruction loss can be formulated as follows:

$$\mathcal{L}_{rec} = \mathcal{L}_{rec\_ske} + \mathcal{L}_{rec\_img}. \tag{8}$$

B. Retrieval Feature Generation

1) Classification Constraint: It should be noted that the semantic features learned from the two branches are constrained only by adversarial loss and cross-reconstruction loss;

these can only provide sufficient semantic knowledge, but cannot ensure that the features will be class-discriminative. However, whether or not the retrieval features are discriminative affects the metric in distance space, which is extremely important for the retrieval task. In order to alleviate this issue, two decoders with shared parameters are trained to decode semantic features to obtain retrieval features. Meanwhile, the category classifiers are introduced after the two branches to generate the retrieval features. By implementing this method, our approach makes retrieval features more discriminative and also alleviates over-fitting by sharing the parameters of the retrieval feature decoders. The loss can be written as follows:

$$\mathcal{L}_{cls} = -\mathbb{E}[log\, P(y|x^{ret}_{ske})] - \mathbb{E}[log\, P(y|x^{ret}_{img})], \tag{9}$$

where y is the category label of x^{ske} and x^{img}, while x^{ret}_{ske} and x^{ret}_{img} denote the retrieval features generated by the sketch and image branches respectively. The generation of these two features can be expressed as follows:

$$x^{ret}_{ske} = \hat{D}_{\theta_{\hat{D}}}(x^{sem}_{ske}), \tag{10}$$

$$x^{ret}_{img} = \hat{D}_{\theta_{\hat{D}}}(x^{sem}_{img}), \tag{11}$$

where $\hat{D}_{\theta_{\hat{D}}}(\cdot)$ is the decoder function that takes the semantic features as inputs and the retrieval features as outputs.

2) Multi-Modal Euclidean Loss: Although sketch-image pairs are drawn from different modalities, their categories are the same, meaning that the retrieval features generated by the network should be similar. To this end, a multi-modal Euclidean loss is introduced to increase the similarity of the retrieval features of the same class from different modalities. The loss can be formulated as follows:

$$\mathcal{L}_{cmt} = ||x^{ret}_{ske} - x^{ret}_{img}||^2_2. \tag{12}$$

C. Objective and Optimization

The full objective of our proposed model can be expressed as follows:

$$\mathcal{L} = \lambda_{adv} \times \mathcal{L}_{adv} + \lambda_{rec} \times \mathcal{L}_{rec} + \lambda_{cls} \times \mathcal{L}_{cls} + \lambda_{cmt} \times \mathcal{L}_{cmt}, \tag{13}$$

where the first and second terms are combined to indicate the loss of generating semantic features, while the last two terms are utilized to regularize the generation procedure of the retrieval features. Moreover, λ_{adv}, λ_{rec}, λ_{cls} and λ_{cmt} are coefficients used to balance the overall performance. The whole model is optimized with Adam [41] in PyTorch; details of the optimization are presented in Algorithm 1.

Since the generators $G_{\theta_{ske}}(\cdot)$ and $G_{\theta_{img}}(\cdot)$ minimize $\mathcal{L}_{adv}$ against an opponent discriminator $D_{\theta_D}(\cdot)$ that tries to maximize $\mathcal{L}_{adv}$, the objective $\mathcal{L}_{adv}$ can be divided into the loss of the generator (denoted as $\mathcal{L}_{gen}$), and the loss of the discriminator (denoted as $\mathcal{L}_{dis}$). During the optimization process, we minimize $\mathcal{L}_{gen}$ to make the generated semantic features more realistic, as well as minimize $\mathcal{L}_{dis}$ to make the discriminator more discriminative. In this article, we optimize our proposed model by first updating the discriminator's network parameters with $\mathcal{L}_{dis}$, then updating the network parameters with the sum of $\mathcal{L}_{gen}$, $\mathcal{L}_{rec}$, $\mathcal{L}_{cls}$ and $\mathcal{L}_{cmt}$.

Algorithm 1 Training Procedure for Our Model

Input: Dataset $D_{tr} = \{(x^{ske}_i, x^{img}_i, s^{seen}_i, y_i)|y_i \in \mathcal{Y}_{train}\}$, max training iteration M, batch size N_B, λ_{adv}, λ_{rec}, λ_{cls} and λ_{cmt}

Output: $\theta_{ske}, \theta_{img}, \theta_{\hat{D}}$

Initialize parameters $\theta_D, \theta_{ske}, \theta_{img}, \theta'_{ske}, \theta'_{img}, \theta_{\hat{D}}$
for i = 1 to M **do**
  Forward model to generate x^{sem}_{ske}, x^{sem}_{img}, x^{ret}_{ske} and x^{ret}_{img};
  Calculate adversarial loss $\mathcal{L}_{adv}$ with Eq. (1);
  Calculate cross-reconstruction loss $\mathcal{L}_{rec}$ with Eq. (2), Eq. (3), Eq. (4), Eq. (5), Eq. (6), Eq. (7) and Eq. (8);
  Calculate classification loss $\mathcal{L}_{cls}$ and multi-modal Euclidean loss $\mathcal{L}_{cmt}$ with Eq. (9) and Eq. (12);
  Update $\theta_D \overset{+}{\leftarrow} -\nabla_{\theta_D}(\mathcal{L}_{dis})$;
  Sum $\mathcal{L}_{gen}, \mathcal{L}_{rec}, \mathcal{L}_{cls}, \mathcal{L}_{cmt}$ as $\mathcal{L}_s$.
  Update $\theta_{ske} \overset{+}{\leftarrow} -\nabla_{\theta_{ske}}(\mathcal{L}_s)$;
  Update $\theta_{img} \overset{+}{\leftarrow} -\nabla_{\theta_{img}}(\mathcal{L}_s)$;
  Update $\theta'_{ske} \overset{+}{\leftarrow} -\nabla_{\theta'_{ske}}(\mathcal{L}_s)$;
  Update $\theta'_{img} \overset{+}{\leftarrow} -\nabla_{\theta'_{img}}(\mathcal{L}_s)$;
  Update $\theta_{\hat{D}} \overset{+}{\leftarrow} -\nabla_{\theta_{\hat{D}}}(\mathcal{L}_s)$;
end for
return $\theta_{ske}, \theta_{img}, \theta_{\hat{D}}$

IV. Experiments

A. Datasets and Settings

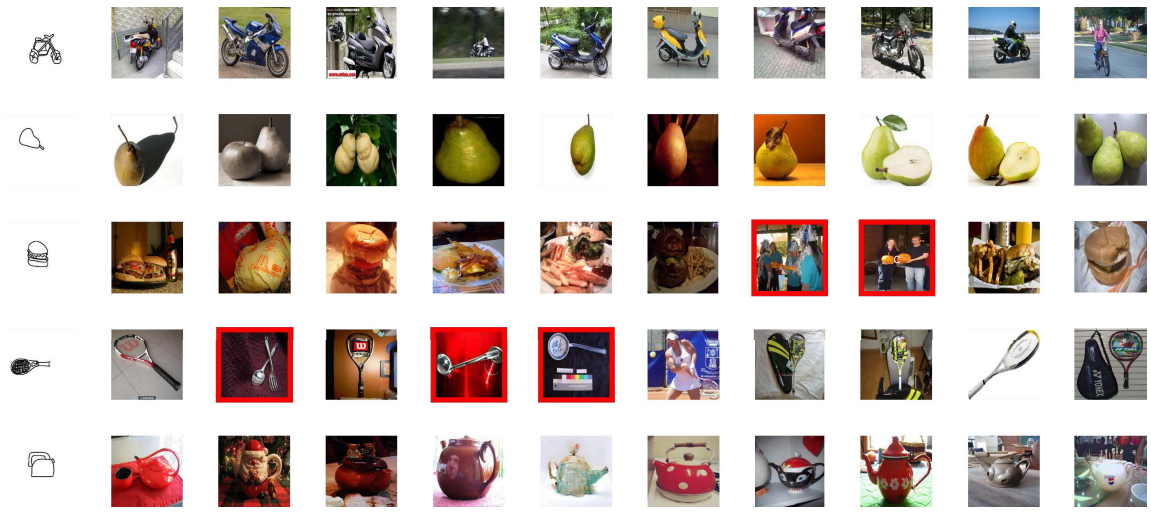
There are two large-scale sketch datasets, *i.e.* Sketchy [11] and TU-Berlin [12], that are widely used for ZS-SBIR. Therefore, we opt to conduct our experiments on these two datasets.

Sketchy is a large-scale sketch dataset that originally consisted of 75,479 sketches and 12,500 images from 125 categories. Liu *et al.* [44] subsequently extended the image gallery by collecting an extra 60,502 images from ImageNet [40], such that the total number of images in the extended dataset is now 73,002. Following the zero-shot data partitioning method outlined in SEM-PCYC [36], we select the same 25 categories as the unseen test set for ZS-SBIR, while the remaining 100 seen classes are used for training.

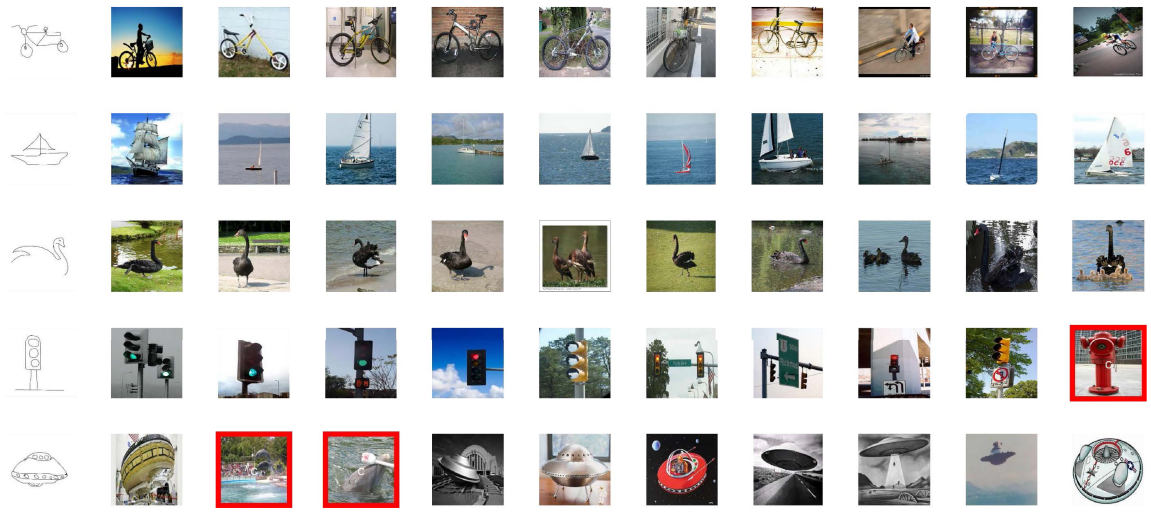
TU-Berlin consists of 20,000 unique free-hand sketches distributed evenly over 250 object categories. Unlike Sketchy [11], TU-Berlin [12] has only category-level matches rather than instance-level matches. In line with the settings outlined in SEM-PCYC [36], the same 30 categories are selected as a query set in the retrieval gallery, while the remaining 220 classes are utilized for training.

To evaluate the performance of the proposed approach, we follow the sketch-based image retrieval evaluation criterion utilized in most previous works [34], [36] in terms of mean average precision (mAP@all) and precision considering the top 100 (Precision@100). Given a query sketch and a list of K ranked retrieval results, the AP for this query is defined as follows:

$$AP(K) = \frac{1}{K}\sum_{n=1}^{K}\delta(r), \tag{14}$$



(a) Results on Sketchy.



(b) Results on TU-Berlin.

Fig. 3. The top 10 images retrieved by our model on the two datasets. The red borders indicate that the retrieved images do not belong to the correct class.

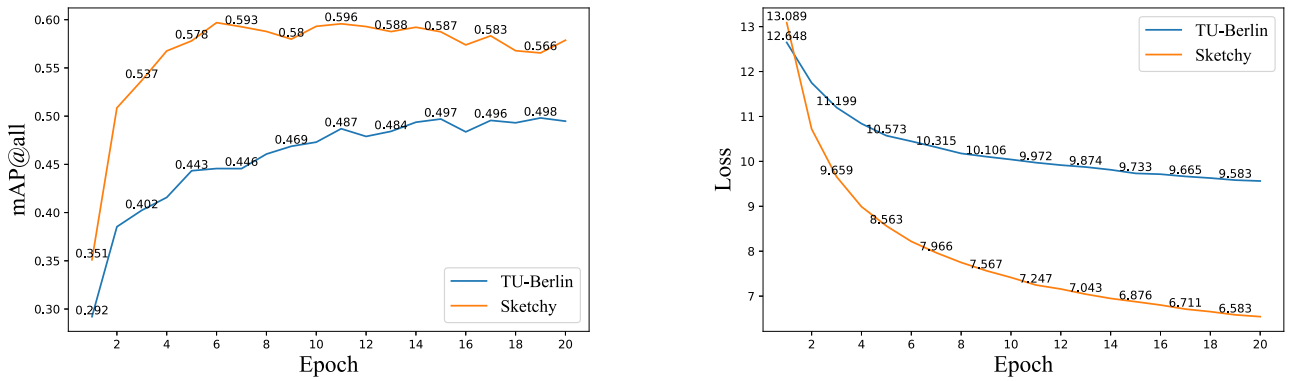
where $\delta(r) = 1$ when the r-th retrieved candidate corresponds to the query; otherwise, $\delta(r) = 0$. The mAP for this query takes ranking information into consideration and can be formulated as follows:

$$mAP(K) = \frac{1}{K}\sum_{n=1}^{K} AP(r). \tag{15}$$

B. Implementation Details

Our model is trained by using the Adam [41] optimizer on PyTorch with an initial learning rate $lr = 0.0001$, $\beta_1 = 0.5$, $\beta_2 = 0.99$. The input size of the image is 224×224. We use the grid search method to select coefficients and determine the best coefficients based on their performance on the validation set. Specifically, the value set of λ_{adv} is $\{0.5, 1.0\}$, while the value set of the remaining coefficients is $\{0.1, 0.2, 0.4, 0.5, 1.0\}$. After searching, we obtain the best coefficients of each loss, which are $\lambda_{adv} = 1.0$, $\lambda_{rec} = 1.0$, $\lambda_{cls} = 0.4$, $\lambda_{cmt} = 0.4$ on Sketch [11] and $\lambda_{adv} = 1.0$, $\lambda_{rec} = 0.5$, $\lambda_{cls} = 0.1$, $\lambda_{cmt} = 0.4$ on TU-Berlin [12].

For feature extraction, we adopt the VGG16 [39] model pre-trained on the ImageNet [40] dataset as a feature extractor for both sketches and images. Moreover, the text-based model [32] trained on Wikipedia is adopted to extract word vectors with dimension of 300. Furthermore, under the zero-shot setting, only the seen classes ought to be considered when constructing the hierarchy for obtaining the class embedding. Therefore, the hierarchical model [33] contains 354 and 664 nodes for Sketchy [11] and TU-Berlin [12], respectively. Firstly, the word vectors from the text-based model and the hierarchical model are concatenated and used as knowledge supervision for the learning of the semantic features. The semantic features are obtained in a generative adversarial fashion, and the decoder is introduced to obtain the 64-dimensional retrieval features. During the testing stage, all test sketches and images are inputted into the network to obtain the retrieval features. Subsequently, the distance between the retrieval features of the sketch and all the natural image retrieval features in the test image dataset are calculated. Finally, we select the top 100 most similar images and compute the accuracy. Although our method does not produce binary hash code as a final representation for matching the sketch and image, the iterative quantization (ITQ) [50] algorithm can be used directly to generate the binary codes.



(a) The map@all value of each epoch

(b) The loss value of each epoch

Fig. 4. As the number of training epochs increases, the loss of our proposed model decreases and the mAP@all on validation set increases. The best model is selected according to the performance on validation set.

Moreover, our model adopts the same validation set used in SEM-PCYC [36] to test the performance after each training epoch. The performance on the validation set and the total loss of our proposed model are shown in Figure 4; each training epoch takes about 70 seconds. Since the training loss is calculated under the zero-shot scenario and the samples of unseen classes cannot be used during training, the optimal parameters for our proposed model are based on the performance on the validation set.

C. Comparison With Peer Methods

Existing relevant SBIR and ZSL approaches are also adopted for retrieval performance evaluation, since the ZS-SBIR task can be considered as a combination of SBIR and ZSL. To facilitate fair comparison, the same seen-unseen splits of categories are used for all relevant experiments.

The performances of all comparison methods under the same zero-shot setting on two datasets are presented in Table I. As can be seen from Table I, most of the ZS-SBIR methods achieve better performance than the SBIR and ZSL methods, while GN Triplet [11] and SAE [9] achieve the best performance in SBIR and ZSL, respectively. The SBIR methods mainly solve the problem of cross-modal retrieval, while the ZSL methods migrate the semantic knowledge from seen to unseen classes. However, because these two methods each only solve one aspect of the problem, it is difficult for them to solve the combined problem of ZS-SBIR. The ZS-SBIR methods therefore achieve better performance, as they possess both the ability to reduce the domain gap and the ability to transfer the semantic knowledge. Moreover, all methods performed worse on TU-Berlin [12]; this may be caused by the large number of classes in this dataset. Furthermore, under the zero-shot setting, our model significantly outperforms the best competitor [37] by more than 3% on Sketchy and 6% on TU-Berlin. This demonstrates that the effectiveness of our proposed model is derived from the progressive strategy employed, which can maintain the integrity of the knowledge and enhance the ability to reduce the domain gap and intra-class variances. Moreover, the cross-reconstruction loss further enables the semantic features to acquire more semantic knowledge, which is beneficial to the transfer of knowledge from seen to unseen classes. Finally, the classification loss ensures that the retrieval features are discriminative, while the multi-modal Euclidean loss ensures increased similarity between the retrieval features of the same class from different modalities; both of them improve the performance of our model in retrieval tasks.

Furthermore, the iterative quantization (ITQ) [50] algorithm is utilized to obtain the binary codes for the retrieval features. As can be seen from the experimental results, the retrieval performance of binary codes decreases to some extent; however, binary retrieval only requires XOR operations, which is faster than utilizing the original features. In order to better demonstrate the performance of binary codes, we count the retrieval time of our model, the results of which can be seen in Table II. From these results, we can conclude that using binary code takes an order of magnitude less time than using real features for retrieval.

The images retrieved using our model are shown in Figure 3. Red borders indicate wrongly returned images. From these results, we can observe that our proposed model is able to maintain semantic consistency well, and that this performance is not affected by size, pose, or background. For example, the retrieved 'sailboat' images differ in size, while the retrieved 'bicycle' images differ in terms of their background. However, the retrieved images closely match the outline of the searching sketch, which can result in the retrieval of images with similar shapes but different categories. For example, the retrieved results for 'signal-light' contain an image of hydrant.

D. Effect of Side Information

Side information is important in zero-shot learning, as it can provide semantic similarity between categories. Moreover, since different types of semantic embeddings have different impacts on performance, we analyze the effects of different semantic embeddings (as well as different combinations of such embeddings) on retrieval performance. Table III presents the quantitative results on both Sketchy and TU-Berlin with different side information mentioned and their combinations. As the results show, the combination of Word2Vec [32] and Jiang-Conrath's hierarchical similarity [51] reaches the highest mAP@all of 52.3% on Sketchy, while on the TU-Berlin dataset, the combination of Word2Vec [32] and path similarity reaches the highest mAP@all of 42.4%. Furthermore, the results also show the same conclusions as presented in [36]: that is, for ZS-SBIR, Word2Vec is

TABLE I
PERFORMANCE COMPARISONS WITH EXISTING SBIR, ZSL, AND ZS-SBIR APPROACHES

	Methods	Sketchy		TU-Berlin	
		mAP@ all	Precision@ 100	mAP@ all	Precision@ 100
SBIR	Softmax Baseline	0.114	0.172	0.050	0.031
	Siamese CNN [42]	0.132	0.175	0.109	0.141
	SaN [19]	0.115	0.125	0.089	0.108
	GN Triplett [11]	0.204	0.296	0.175	0.253
	3D Shape [43]	0.067	0.078	0.054	0.067
	DSH (binary) [44]	0.171	0.231	0.129	0.189
	GDH (binary) [45]	0.187	0.259	0.135	0.212
ZSL	CMT [25]	0.087	0.102	0.062	0.078
	DeViSE [8]	0.067	0.077	0.059	0.071
	SSE [46]	0.116	0.161	0.089	0.121
	JLSE [47]	0.131	0.185	0.109	0.155
	SAE [9]	0.216	0.293	0.167	0.221
	FRWGAN [48]	0.127	0.169	0.110	0.157
	ZSH (binary) [49]	0.159	0.214	0.141	0.171
ZS-SBIR	ZSIH (binary) [34]	0.258	0.342	0.223	0.294
	CVAE [35]	0.196	0.284	0.005	0.001
	SEM-PCYC [36]	0.349	0.463	0.297	0.426
	SEM-PCYC (binary) [36]	0.344	0.399	0.293	0.392
	CSDB [37]	0.484	0.375	0.355	0.254
	Ours	**0.523**	**0.616**	**0.424**	**0.517**
	Ours (binary)	**0.506**	**0.615**	**0.355**	**0.452**

TABLE II
RETRIEVAL TIME USING BINARY AND REAL FEATURES ON TWO DATASETS

	Sketchy	TU-Berlin
Real features	$8.786 \times 10^{-3}s$	$1.015 \times 10^{-3}s$
Binary	$4.080 \times 10^{-4}s$	$7.733 \times 10^{-4}s$

TABLE III
MAP@ALL OF ZS-SBIR BY USING DIFFERENT SEMANTIC EMBEDDINGS AND THEIR COMBINATIONS ON TWO DATASETS

Text Model		Hierarchical Model		Sketchy	TU-Berlin
Glove	Word2Vector	Path	Ji-Cn [51]		
✓				0.459	0.367
	✓			0.465	0.371
		✓		0.489	0.384
			✓	0.502	0.385
✓		✓		0.509	0.401
✓			✓	0.518	0.408
	✓	✓		0.508	**0.424**
	✓		✓	**0.523**	0.411

better than GloVe at capturing semantic similarity between words. Moreover, the text-based model and the hierarchical model complementarily work together to represent semantic information.

E. Ablation Studies

In this section, some ablation studies are presented to verify the effectiveness of our proposed model. The results are exhibited in Table IV. In order to verify the validity of generating retrieval features in a progressive way, we first train a model that projects the image and sketch features directly into the low-dimensional retrieval space, and meantime reduces the dimension of word vectors in order to obtain 64-dimensional features that are used to constrain the relationship between the projected features. Second, we train a baseline that generates the retrieval features in a progressive way. In order to obtain the semantic features of the same dimension as word vectors, the baseline first aligns the sketch and image features to the word vectors, then takes the semantic features as input of the decoders to generate 64-dimensional retrieval features. The decoders used to generate the retrieval features do not share parameters in this baseline. Next, we make the decoders that generate the retrieval features share parameters to prove that this strategy can alleviate over-fitting. Moreover, to demonstrate the effectiveness of the multi-modal Euclidean loss and cross-reconstruction loss, we conduct experiments by alternatively ablating $\mathcal{L}_{cmt}$ and $\mathcal{L}_{rec}$ in Eq. (13). Finally, in order to prove the effectiveness of the cross-reconstruction loss, we also conduct experiments with single-reconstruction loss, such that the learned semantic features only reconstruct the original modality of the current branch. The single reconstruction loss can be formulated as follows:

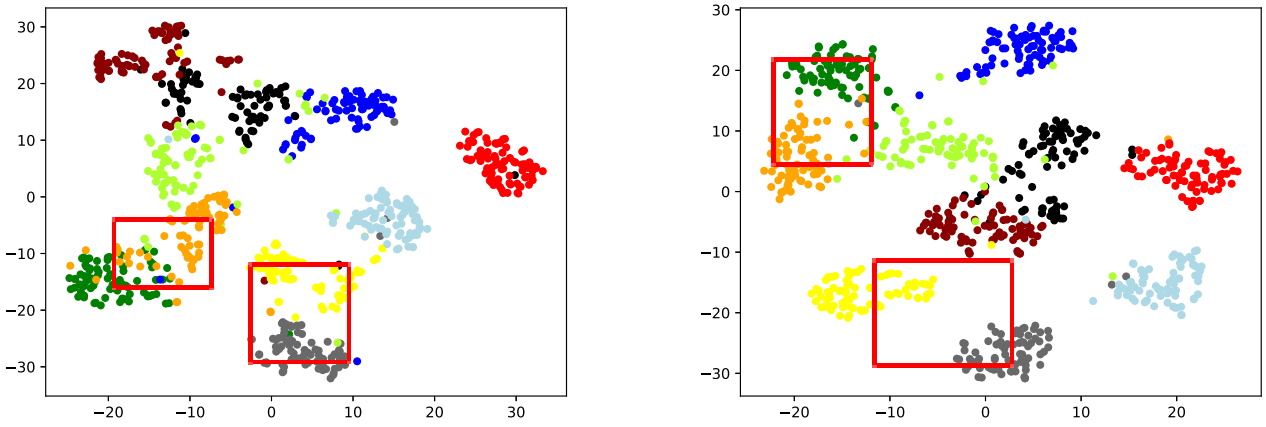
$$\mathcal{L}_{sin\_rec} = ||\tilde{x}_{ske}^{ske} - x^{ske}||_2^2 + ||\tilde{x}_{img}^{img} - x^{img}||_2^2. \quad (16)$$

The mAP@all values obtained by the baselines described above are presented in Table IV. As the results indicate, these baselines achieve lower performance than the complete model. The baseline, which generates the retrieval features in a progressive way, performs better than the model that directly projects the sketch and image features into a low-dimensional common space. Given that ZS-SBIR is a combination of

TABLE IV

ABLATION STUDIES ON OUR MODEL MAP@ALL RESULTS OF SEVERAL BASELINES

Description	Sketchy	TU-Berlin
Without progressive way	0.327	0.244
Baseline (The decoders of retrieval features do not share parameters)	0.396	0.299
Baseline + The decoders of retrieval features share parameters	0.416	0.318
Baseline + The decoders of retrieval features share parameters + $\mathcal{L}_{cmt}$	0.455	0.364
Baseline + The decoders of retrieval features share parameters + $\mathcal{L}_{rec}$	0.443	0.342
Baseline + The decoders of retrieval features share parameters + $\mathcal{L}_{sin\_rec}$	0.424	0.331
Baseline + The decoders of retrieval features share parameters + $\mathcal{L}_{cmt}$ + $\mathcal{L}_{sin\_rec}$	0.482	0.391
Baseline + The decoders of retrieval features share parameters + $\mathcal{L}_{cmt}$ + $\mathcal{L}_{rec}$	**0.523**	**0.424**



(a) Without progressive method (b) Progressive method

Fig. 5. Visualization of the retrieval features learned from sketch modality using two different methods on the TU-Berlin dataset. We sample 10 classes of from test categories and visualize the distribution of the features. Each color represents a particular class.

two tasks, the progressive way is reasonable; this is because the first step synthesizes the semantic features of the same dimension as the word vectors by explicit alignment, which maintains the same empirical distribution as the word vectors [52], while the second step projects the aligned features to a low-dimensional common space, which is significant for retrieval. We further visualize the distributions of the retrieval features with t-SNE [53]. From Figure 5, it can be seen that the distribution of retrieval features learned using the progressive method is more discriminative than the alternative. The results demonstrate that this strategy can improve performance by more than 6% on Sketchy and 5% on TU-Berlin. We also calculate the discrimination ratio [54] of the retrieval features, which are learned both with and without the progressive method. The discrimination ratio is measured by the ratio of between-class scatter to within-class scatter; the greater this ratio is, the more discriminative the retrieval features are. In order to measure the discrimination of retrieval features of a specific category, the between-class scatter is the average of the distances between the centroids of each unseen class and its nearest class. Moreover, the formula of within-class scatter is as follows:

$$S^2 = \mathbb{E}(||x_{ske}^{ret} - \mathbb{E}(x_{ske}^{ret})||_2^2). \tag{17}$$

We select ten unseen categories in TU-Berlin to compute their discrimination ratio; each of these categories contains 80 sketches. From Figure 6, we can see that the discrimination ratio of retrieval features generated in the progressive way

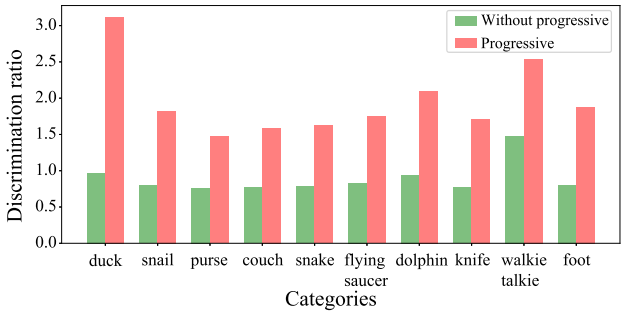


Fig. 6. The discrimination is measured by the ratio of between-class scatter to within-class scatter. The greater this ratio is, the more discriminative the categories are.

is greater than those generated without the progressive way, which demonstrates that the semantic information is better retained when the progressive method is used. Moreover, in the retrieval feature decoding stage, two decoders are trained respectively in the baseline, which can lead to over-fitting. Therefore, parameters sharing is beneficial to preventing both branches from over-fitting their own modalities, and also outperforms the baseline by around 2% both on Sketchy and TU-Berlin. Furthermore, the imposition of multi-modal Euclidean loss guarantees the similarity of retrieval features from different modalities, which is conducive to the metrics of distance space for the retrieval task. However, the semantic features learned by adversarial training only maps the sketch and image features to a semantic space, which

cannot guarantee that the sketch-image pairs of the same category will be matched. Therefore, the imposition of the reconstruction constraint ensures the correspondence of the sketch-image categories. The single-reconstruction loss only considers the translation back to the current modality, while the cross-reconstruction loss considers the corresponding modality, which facilitates the transfer of knowledge from both two modalities. Finally, the full model reaches the highest mAP@all of 52.3% on Sketchy and 42.4% on TU-Berlin.

V. Conclusion

In this article, we have presented a novel network designed to address the problem of ZS-SBIR more effectively in a progressive way. The progressive generation of retrieval features solves the problem of knowledge loss that occurs due to the lack of explicit alignment, which is conducive to the migration of knowledge from seen to unseen classes. Moreover, the cross-reconstruction loss guarantees the increased sufficiency of the semantic features, which is crucial to transferring knowledge from both modalities and reducing the domain gap. Subsequently, decoders with shared parameters are utilized to generate retrieval features under the constraint of classification loss, which can alleviate over-fitting. In addition, the use of multi-modal Euclidean loss enhances the similarity of retrieval features that are in the same class but from different modalities, which improves the retrieval performance by reducing the domain gap. Experiments on two large-scale datasets verified that our proposed model significantly outperforms existing methods in the ZS-SBIR task.

References

[1] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3893–3903, Aug. 2018.

[2] L. Jin, K. Li, H. Hu, G.-J. Qi, and J. Tang, "Semantic neighbor graph hashing for multimodal retrieval," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1405–1417, Mar. 2018.

[3] C. Deng, E. Yang, T. Liu, J. Li, W. Liu, and D. Tao, "Unsupervised semantic-preserving adversarial hashing for image search," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4032–4044, Aug. 2019.

[4] G.-J. Qi, X.-S. Hua, and H.-J. Zhang, "Learning semantic distance from community-tagged media collection," in *Proc. 17th ACM Int. Conf. Multimedia (MM)*, 2009, pp. 243–252.

[5] S. Chang, G.-J. Qi, C. C. Aggarwal, J. Zhou, M. Wang, and T. S. Huang, "Factorized similarity learning in networks," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 60–69.

[6] Z. Peng, Z. Li, J. Zhang, Y. Li, G.-J. Qi, and J. Tang, "Few-shot image recognition with knowledge transfer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 441–449.

[7] G.-J. Qi, W. Liu, C. Aggarwal, and T. Huang, "Joint intermodal and intramodal label transfers for extremely rare or unseen classes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1360–1373, Jul. 2017.

[8] A. Frome *et al.*, "Devise: A deep visual-semantic embedding model," in *Proc. NIPS*, 2013, pp. 2121–2129.

[9] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3174–3183.

[10] L. Chen, H. Zhang, J. Xiao, W. Liu, and S.-F. Chang, "Zero-shot visual recognition using semantics-preserving adversarial embedding networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1043–1052.

[11] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: Learning to retrieve badly drawn bunnies," *ACM Trans. Graph.*, vol. 35, no. 4, p. 119, Jul. 2016.

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

[13] R. Hu and J. Collomosse, "A performance evaluation of gradient field HOG descriptor for sketch based image retrieval," *Comput. Vis. Image Understand.*, vol. 117, no. 7, pp. 790–806, Jul. 2013.

[14] J. M. Saavedra, "Sketch based image retrieval using a soft computation of the histogram of edge local orientations (S-HELO)," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 2998–3002.

[15] J. M. Saavedra and J. M. Barrios, "Sketch based image retrieval using learned KeyShapes (LKS)," in *Proc.Brit. Mach. Vis. Conf.*, vol. 1, no. 2, 2015, p. 7.

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

[17] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 539–546.

[18] D. Xu, X. Alameda-Pineda, J. Song, E. Ricci, and N. Sebe, "Cross-paced representation learning with partial curricula for sketch-based image retrieval," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4410–4421, May 2018.

[19] Q. Yu, Y. Yang, F. Liu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Sketch-a-net: A deep neural network that beats humans," *Int. J. Comput. Vis.*, vol. 122, no. 3, pp. 411–425, May 2017.

[20] X. Li, S. Liao, W. Lan, X. Du, and G. Yang, "Zero-shot image tagging by hierarchical semantic embedding," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2015, pp. 879–882.

[21] T. Dutta and S. Biswas, "Generalized zero-shot cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5953–5962, Dec. 2019.

[22] J. Qin *et al.*, "Zero-shot action recognition with error-correcting output codes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2833–2842.

[23] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 7, pp. 1425–1438, Jul. 2016.

[24] S. Changpinyo, W.-L. Chao, and F. Sha, "Predicting visual exemplars of unseen classes for zero-shot learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3476–3485.

[25] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Proc. NeurIPS*, 2013, pp. 935–943.

[26] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 69–77.

[27] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2021–2030.

[28] F. Jurie, M. Bucher, and S. Herbin, "Generating visual representations for zero-shot classification," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2666–2673.

[29] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, Mar. 2014.

[30] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2927–2936.

[31] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 49–58.

[32] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: http://arxiv.org/abs/1301.3781

[33] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[34] Y. Shen, L. Liu, F. Shen, and L. Shao, "Zero-shot sketch-image hashing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3598–3607.

[35] S. Kiran Yelamarthi, S. Krishna Reddy, A. Mishra, and A. Mittal, "A zero-shot framework for sketch based image retrieval," in *Proc. ECCV*, 2018, pp. 300–317.

[36] A. Dutta and Z. Akata, "Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5089–5098.

[37] T. Dutta and S. Biswas, "Style-guided zero-shot sketch-based image retrieval," in *Proc. BMVC*, 2019, p. 209.

[38] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[42] Y. Qi, Y.-Z. Song, H. Zhang, and J. Liu, "Sketch-based image retrieval via siamese convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2460–2464.
[43] M. Wang, C. Wang, J. X. Yu, and J. Zhang, "Community detection in social networks: An in-depth benchmarking study with a procedure-oriented framework," *VLDB*, vol. 8, no. 10, pp. 998–1009, 2015.
[44] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao, "Deep sketch hashing: Fast free-hand sketch-based image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2862–2871.
[45] J. Zhang *et al.*, "Generative domain-migration hashing for sketch-to-image retrieval," in *Proc. ECCV*, vol. 2018, pp. 297–314.
[46] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4766–4779, Dec. 2015.
[47] Z. Zhang and V. Saligrama, "Zero-shot learning via joint latent similarity embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 6034–6042.
[48] R. Felix, V. B. Kumar, I. Reid, and G. Carneiro, "Multi-modal cycle-consistent generalized zero-shot learning," in *Proc. ECCV*, 2018, pp. 21–37.
[49] Y. Yang, Y. Luo, W. Chen, F. Shen, J. Shao, and H. T. Shen, "Zero-shot hashing via transferring supervised knowledge," in *Proc. ACM Multimedia Conf. (MM)*, 2016, pp. 1286–1295.
[50] L. Liu, M. Yu, and L. Shao, "Learning short binary codes for large-scale image retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1289–1299, Mar. 2017.
[51] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," 1997, *arXiv:cmp-lg/9709008*. [Online]. Available: https://arxiv.org/abs/cmp-lg/9709008
[52] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.
[53] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[54] B. Tong, C. Wang, M. Klinkigt, Y. Kobayashi, and Y. Nonaka, "Hierarchical disentanglement of discriminative latent features for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11467–11476.

Cheng Deng (Senior Member, IEEE) received the B.E., M.S., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China. He is currently a Full Professor with the School of Electronic Engineering, Xidian University. His research interests include computer vision, pattern recognition, and information hiding. He is the author and coauthor of more than 100 scientific articles at top venues, including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS), the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), the IEEE TRANSACTIONS ON CYBERNETICS (TCYB), the TRANSACTIONS ON MULTIMEDIA (TMM), the TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS (TSMC), ICCV, CVPR, ICML, NIPS, IJCAI, and AAAI.

Xinxun Xu received the B.Sc. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2017, and the M.S. degree in electronics and communications engineering from Xidian University, Xi'an, China, in 2020. His research interests include zero-shot learning, sketch-based image retrieval, and deep learning.

Hao Wang received the B.E. degree in electronic and information engineering from Hangzhou Dianzi University, China, in 2015. He is currently pursuing the Ph.D. degree with the School of Electronic Engineering, Xidian University, Xi'an, China. His main research interests include human action recognition, video understanding, and zero shot learning.

Muli Yang received the B.E. degree in electronic science and technology from Xidian University, Xi'an, China, where he is currently pursuing the Ph.D. degree with the School of Electronic Engineering. His research interests include computer vision, visual reasoning, and machine learning.

Dacheng Tao (Fellow, IEEE) is currently a Professor of computer science and an ARC Laureate Fellow with the School of Computer Science and the Faculty of Engineering, The University of Sydney. His research results in artificial intelligence have expounded in one monograph and more than 200 publications at prestigious journals and prominent conferences, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), IJCV, JMLR, AIJ, AAAI, IJCAI, NeurIPS, ICML, CVPR, ICCV, ECCV, ICDM, and KDD, with several best paper awards. He received the 2018 IEEE ICDM Research Contributions Award and the 2015 Australian Museum Scopus-Eureka Prize. He is a Fellow of AAAS, ACM, and the Australian Academy of Science.