Transferable Coupled Network for Zero-Shot Sketch-Based Image Retrieval

Hao Wang, Cheng Deng, *Senior Member, IEEE*, Tongliang Liu, and Dacheng Tao, *Fellow, IEEE*

Abstract—Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR) aims at searching corresponding natural images with the given free-hand sketches, under the more realistic and challenging scenario of Zero-Shot Learning (ZSL). Prior works concentrate much on aligning the sketch and image feature representations while ignoring the explicit learning of heterogeneous feature extractors to make themselves capable of aligning multi-modal features, with the expense of deteriorating the transferability from seen categories to unseen ones. To address this issue, we propose a novel Transferable Coupled Network (TCN) to effectively improve network transferability, with the constraint of soft weight-sharing among heterogeneous convolutional layers to capture similar geometric patterns, e.g., contours of sketches and images. Based on this, we further introduce and validate a general criterion to deal with multi-modal zero-shot learning, i.e., utilizing coupled modules for mining modality-common knowledge while independent modules for learning modality-specific information. Moreover, we elaborate a simple but effective semantic metric to integrate local metric learning and global semantic constraint into a unified formula to significantly boost the performance. Extensive experiments on three popular large-scale datasets show that our proposed approach outperforms state-of-the-art methods to a remarkable extent: by more than 12% on Sketchy, 2% on TU-Berlin and 6% on QuickDraw datasets in terms of retrieval accuracy. The project page is available at: https://haowang1992.github.io/publication/TCN.

Index Terms—Transferable coupled network, semantic metric, sketch-based image retrieval, zero-shot learning

1 Introduction

With the proliferation of mobile devices, sketches can be obtained effortlessly and massively by drawing on tablets, phones and even smart watches. Due to they illustrate target objects visually and concisely, sketch-oriented applications [1], [2], [3], [4], [5], [6], especially Sketch-Based Image Retrieval (SBIR) [7], [8], [9], [10], [11], have garnered considerable attention. However, conventional SBIR must obey that the categories are same across the training and testing stage, which is hard to guarantee in realistic scenarios. Hence, Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR) [12], [13], [14], [15] has emerged recently, which conducts SBIR under the setting of Zero-Shot Learning (ZSL) [16], [17], [18], [19], as illustrated in Fig. 1. This task is extremely challenging, since it requires the alignment learned at training stage between sketches and images can be effectively transferred at testing stage.

Prior works on ZS-SBIR [12], [13], [14], [15], [20], [21], [22], [23] can be roughly divided into two categories, depending on whether they project sketches and images into a common space [12], [14], [15], [20], [22], [23] or utilize generative model to synthesize image features from sketch ones [13], [21]. However, most of them generally follow the paradigm of aligning multi-modal feature representations with extra modules, e.g., discriminators or classifiers, without the explicit modeling of heterogeneous feature extractors to make themselves learn to align. We advocate that it will deteriorate knowledge transferability from seen categories to unseen ones, since their feature extractors are prone to overfit on seen categories to achieve the goal of multi-modal alignment. Specifically, previous works fine-tune pre-trained models on sketches and images either separately or with hard weight-sharing strategy (i.e., all corresponding weights are same). We hold that there are two essential drawbacks: (1) although using two-branch networks and fine-tuning them individually on sketches and images can learn alignment to some extent, it inevitably introduces redundant modality-specific parameters to over-fit the goal of multi-modal alignment, which obviously performs poor when meeting unseen categories at testing stage, (2) even though employing hard weight-sharing strategy between heterogeneous backbones can jointly model sketches and images to obtain certain alignment and eliminate modality-specific parameters, the whole model prefers to learning on images as the optimization on them is easier than on sketches, which produces much irrelevant parameters for sketches and deteriorates its knowledge transferability. In brief, previous works fail to design proper heterogeneous feature extractors

- *Hao Wang and Cheng Deng are with the School of Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, China. E-mail: {haowang.xidian, chdeng.xd}@gmail.com.*
- *Tongliang Liu is with the School of Computer Science, Faculty of Engineering, The University of Sydney, Darlington, NSW 2008, Australia. E-mail: tongliang.liu@sydney.edu.au.*
- *Dacheng Tao is with the JD Explore Academy, Beijing 101100, China. E-mail: dacheng.tao@gmail.com.*

Manuscript received 24 October 2020; revised 14 April 2021; accepted 11 October 2021. Date of publication 27 October 2021; date of current version 3 November 2022.
This work was supported in part by the National Natural Science Foundation of China under Grants 62132016, 62171343, and 62071361, in part by the Key Research and Development Program of Shaanxi under Grant 2021ZDLGY01-03, in part by the Fundamental Research Funds for the Central Universities under Grant ZDRC2102, and in part by the Australian Research Council Projects DE-190101473 and IC-190100031.
(Corresponding authors: Cheng Deng and Tongliang Liu.)
Recommended for acceptance by B. Caputo.
Digital Object Identifier no. 10.1109/TPAMI.2021.3123315

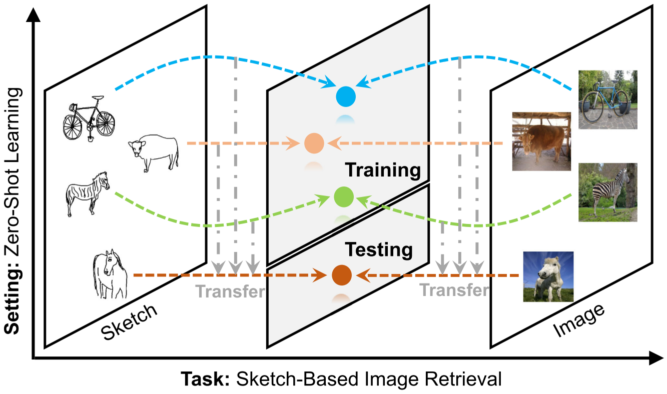

Setting: Zero-Shot Learning
Training
Testing
Transfer
Transfer
Sketch
Image
Task: Sketch-Based Image Retrieval

Fig. 1. Zero-shot sketch-based image retrieval aims at performing sketch-based image retrieval under the realistic scenario of zero-shot learning.

for ZS-SBIR. Besides, prior researches generally adopt local metric learning (e.g., triplet metric) and global semantic constraint (e.g., semantic regression) to improve feature discriminability and transferability, respectively. However, they are usually weighted with empirical coefficient to achieve the trade-off, which is complicated for tuning hyper-parameters at training stage.

Before devising the feature extractor of ZS-SBIR, we first present a brief review of main components, e.g., convolutional layer, pooling layer and batch normalization layer, in modern convolutional neural network. As we all known, convolutional layer is designed to mimic the primary visual cortex in human brain, leveraging the ideas of sparse interactions, parameter sharing and equivariant representations. With the aid of pooling layer, they can capture the edges, textures and parts of object progressively [24], which is able to discover the patterns of input data. For batch normalization layer, it is originally proposed to accelerate the training by solving internal covariate shift [25]. Nonetheless, it contains the calculation of batch-wise statistics, e.g., mean and variance, and the memorization of dataset-wise statistics, e.g., running mean and running variance, which can depict the uniqueness of input modality.

Based on the observation and discussion above, we then jointly consider the characteristics of network components and the nature of ZS-SBIR task to frame our architecture. Concretely, we need to learn modality-common information and modality-specific ones for cross-modal retrieval, which can be implemented with weight-shared convolutional layers and independent batch normalization layers, respectively. Furthermore, inspired by the intuition that human beings rely on geometric patterns (e.g., contours) of sketches and images to match them, we transform the weight-shared heterogeneous convolutional layers into coupled ones via soft weight-sharing strategy to effectively guide the learning of images with sketch patterns, yielding a novel Transferable Coupled Network (TCN) for ZS-SBIR. Based on this, we explicitly present a general criterion, i.e., adopting coupled modules for mining modality-common knowledge while independent modules for learning modality-specific information, to deal with multi-modal zero-shot learning. Moreover, to simplify the complicated training procedure of utilizing separate local metric learning and global semantic constraint, we unify them through integrating global semantic information into the anchor generation of local metric learning, producing a simple but effective semantic metric.

The main contributions of this work are as follows:

- We frame a novel transferable coupled network to force heterogeneous feature extractors learn to align with similar geometric patterns, which effectively improve knowledge transferability for ZS-SBIR.
- We elaborate a simple but effective semantic metric by integrating local metric learning and global semantic constraint, which significantly simplifies the training procedure and boosts the performance.
- To our best knowledge, we are the first to explicitly present a general criterion for multi-modal zero-shot learning, which will greatly help the researchers to tailor their own frameworks. Especially, it can be seamlessly and effortlessly employed in zero-shot and unsupervised domain adaptation.
- Extensive experiments on three popular large-scale datasets, Sketchy, TU-Berlin and QuickDraw, demonstrate that our approach outperforms the state-of-the-art methods by a large margin.

The rest of this paper is organized as follows. We first present a brief review in Section 2. Then we give a formal problem statement of ZS-SBIR and propose our approach in Section 3. Section 4 demonstrates the experiments, followed by the conclusion in Section 5.

2 Related Work

2.1 Sketch-Based Image Retrieval

With the release of two large-scale datasets, i.e., Sketchy [8] and TU-Berlin [26], SBIR has attracted ever-increasing attention among computer vision community. Early works employed hand-crafted features, e.g., gradient field HOG [27], Histogram of Edge Local Orientations (HELO) [28] and Learned KeyShapes (LKS) [7], to represent the sketches for subsequent cross-modal retrieval. Recently, deep learning has been introduced into this field and obtained appealing performance. Sketch-a-Net [3] was the first to successfully apply deep convolutional neural network for sketch recognition. To further improve the discriminability of multi-modal feature representation, siamese network [9] and triplet network [10] were utilized for SBIR. Besides, Semi3-Net [11] employed semi-heterogeneous architecture for feature mapping, with the aid of edgemap. euclidean Margin Softmax (EMS) [29] attempted to minimize both intra-class and inter-class distance. Visual Trait Descriptor (VTD) [30] built a universal manifold of prototypical visual sketch traits to parameterize the learning of a sketch or image representation for fine-grained SBIR. However, these methods are not specifically elaborated for ZS-SBIR, meaning that they neglect the learning of knowledge transfer from seen categories to unseen ones. In contrast, our proposed architecture is elaborated to deal with SBIR and ZSL simultaneously.

2.2 Zero-Shot Sketch-Based Image Retrieval

To investigate the transferability of learned cross-modal representations under ZSL scenario, more realistic ZS-SBIR [13], [14] works have been reported recently. Zero-shot

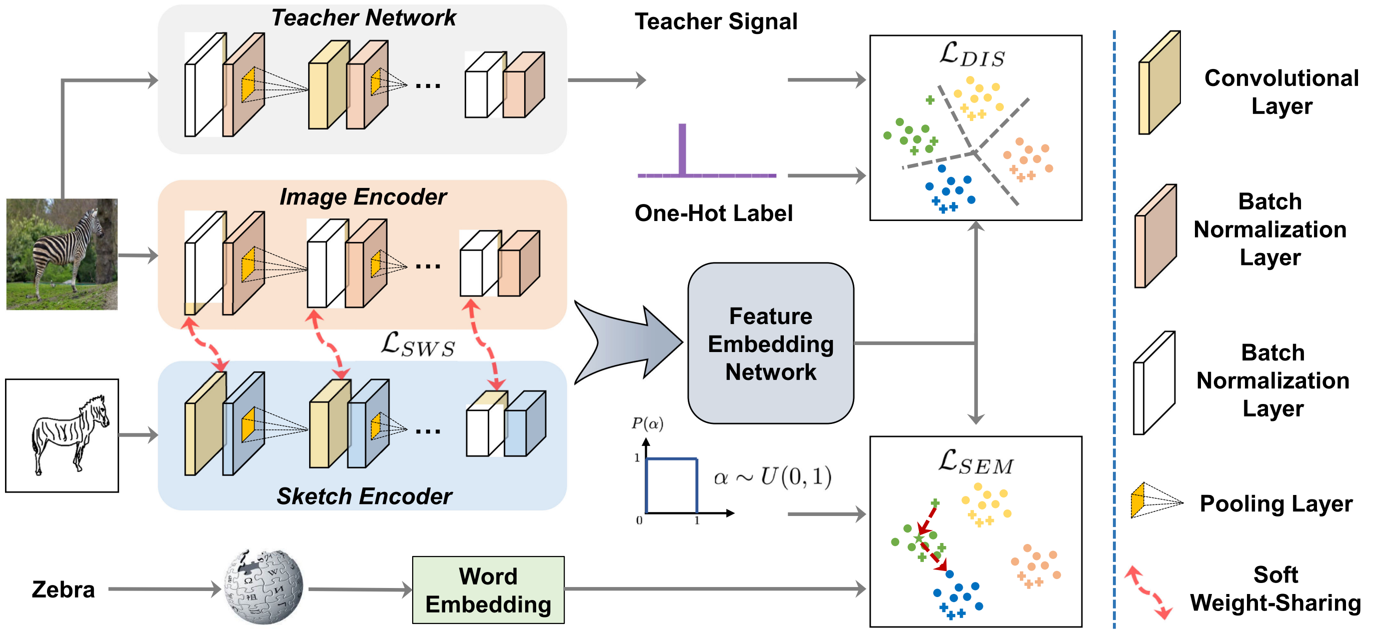


Fig. 2. Our framework consists of transferable coupled network (i.e., coupled image and sketch encoders), feature embedding network, discrimination module and semantic metric module. Specifically, the first one includes soft weight-shared convolutional layers (i.e., constrained with $\mathcal{L}_{SWS}$) and independent batch normalization layers. Retrieval features are obtained through feature embedding network, along with discrimination loss $\mathcal{L}_{DIS}$ and semantic metric loss $\mathcal{L}_{SEM}$. At training stage, the guiding signals produced from teacher network as well as benchmark one-hot labels are provided to calculate $\mathcal{L}_{DIS}$. Similarly, word vectors and uniform noise are offered to compute $\mathcal{L}_{SEM}$.

Sketch-Image Hashing (ZSIH) [12] leveraged two-branch encoders and semantic graph to tackle their introduced zero-shot hashing task of SBIR. Conditional Variational AutoEncoders (CVAE) [13] attempted to solve ZS-SBIR via generative model to synthesize natural image features from sketch features. SEM-PCYC [15] employed more complicated generative model and cycle consistency to learn better modality-common embedding. Content-Style Decomposition (CSD) [21] generated style-guided image features via decomposition and fusion technologies for retrieval. Semantic-Aware Knowledge prEservation (SAKE) [20] fine-tuned hard weight-sharing feature extractors while kept the knowledge acquired from ImageNet [31], which yields state-of-the-art performance while requires specialistic designing of data sampling to avoid severe imbalanced learning. However, all works either simply adopt the models pre-trained on ImageNet as feature extractors [13], fine-tune individual backbones to over-fit the goal of multi-modal alignment [15], or adopt hard weight-sharing strategy between various backbones [20], leading to unsatisfied zero-shot retrieval performance. Different from them, we learn to align sketch and image feature representations from the perspective of modeling heterogeneous feature extractors themselves, instead of aligning category-level heterogeneous features.

2.3 Zero-Shot Metric Learning

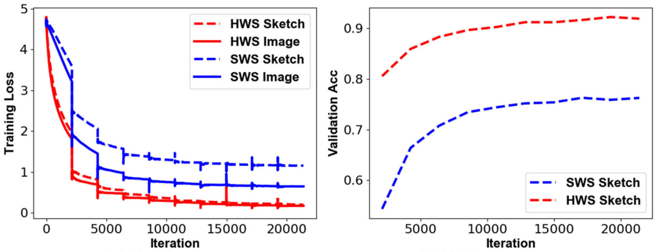
Metric learning has played a great role in several tasks. Recently, many works attempt to extend it into the field of ZSL. Early work [32] improved semantic embedding consistency during pairwise metric learning for zero-shot classification. Adaptive Metric Learning (AML) [33] treated the most similar seen category samples as substitution of unseen ones to regularize the compatibility metric function. Decoupled Metric Learning (DeML) [34] strengthened the generalization ability of model by decoupling unified representations into multiple attention-specific learners. Model-Agnostic Metric (MAM) [35] leveraged cosine metric to alleviate hubness problem. However, none of them considers local metric learning and global semantic constraint simultaneously. In contrast, our proposed semantic metric combines them into a unified formula, which significantly simplifies the training procedure.

3 Methodology

In this section, we first introduce the problem of ZS-SBIR. Then we detail our proposed framework, which consists of transferable coupled network (i.e., coupled image and sketch encoders), feature embedding network, discrimination module and semantic metric module, as illustrated in Fig. 2. Specifically, multi-modal representations are first obtained from the transferable coupled network. Then retrieval features are generated via feature embedding network, with the aid of discrimination learning and semantic metric learning.

3.1 Problem Statement

The dataset of ZS-SBIR consists of two disjoint parts, i.e., $D^{Tr} = \{I^{Tr}, S^{Tr}\}$ and $D^{Te} = \{I^{Te}, S^{Te}\}$, where I and S are the subsets of natural images and sketches, respectively. The superscripts Tr and Te stand for training and testing split, and ZSL setting indicates that $D^{Tr} \cap D^{Te} = \emptyset$. During training stage, in addition to D^{Tr}, the category-level one-hot vector Y^C (e.g., hard label) and semantic vector Y^S (e.g., word vector [36]) are provided to train the model. At testing stage, given a sketch query, the learned model is to retrieve corresponding natural images from testing image gallery. Therefore, the essence of ZS-SBIR is to make the learned multi-



(a) Training Loss (b) Validation Accuracy

Fig. 3. Training loss and validation accuracy versus the number of iterations with 64 dimensional features on Sketchy.

modal alignment transferred from seen categories to unseen ones. This produces extremely high requirements on designing of network architecture.

3.2 Transferable Coupled Network

Since sketches lack the detailed information, such as color and texture, the models pre-trained on ImageNet are incapable of extracting sketch features. Hence, prior works [12], [14], [15] mostly fine-tune the pre-trained models with two individual branches for feature extraction. However, they still perform poor when meeting unseen categories at testing stage, as they introduce modality-specific parameters to over-fit the goal of multi-modal alignment. Intuitively, we need to make heterogeneous feature extractors themselves capable of aligning sketch and image representations, instead of merely aligning the multi-modal features. Therefore, recent work, i.e., SAKE [20], attempts to jointly model sketches and images with hard weight-shared backbones. Unfortunately, the whole model is prone to optimize on image modality and generates much irrelevant parameters for the feature extraction of sketches, thus performing poor on knowledge transferability at testing stage.

Hence, we propose a novel transferable coupled network for ZS-SBIR, to force heterogeneous feature extractors to explicitly learn similar geometric information, via guiding the learning of images with sketch patterns. Fundamentally, given sketches and images, their feature representation can be formulated as

$$F^S = \mathcal{G}_S(S;\theta_S), \quad F^I = \mathcal{G}_I(I;\theta_I), \tag{1}$$

where $\mathcal{G}_S$ and $\mathcal{G}_I$ are sketch encoder parameterized with θ_S and image encoder parameterized with θ_I, respectively. F^S and F^I denote feature representations of sketches and images extracted after the final convolutional stage. Then we formally describe the core component of transferable coupled network as

$$\mathcal{L}_{SWS} = \sum_l \mathbb{1}[l \notin \mathrm{BN}] \cdot \|\theta_S^l - \theta_I^l\|_2^2, \tag{2}$$

where θ_S^l and θ_I^l are parameters of sketch encoder $\mathcal{G}_S$ and image encoder $\mathcal{G}_I$ at layer l, respectively. This fashion of soft weight-sharing, i.e., $\|\theta_S^l - \theta_I^l\|_2^2$, forces the image encoder to model geometric patterns of images through the guiding of sketch parameters. Besides, the indicator function $\mathbb{1}[l \notin \mathrm{BN}]$ equals to 1 if the layer l of sketch or image encoder is not the batch normalization layer, and 0 otherwise. Here, independent batch normalization layers are adopted to separate modality-common information from modality-specific one for cross-modal retrieval.

In what follows, we present a detailed comparison of existing two strategies, i.e., hard weight-sharing and soft weight-sharing, to verify the effectiveness of our proposed transferable coupled network, from the perspective of optimization and experimental results respectively.

Mathematically, given the sketch network g_s parameterized with θ_s and image network g_i parameterized with θ_i, the loss can be formulated as

$$\mathcal{L} = \mathcal{L}_s(\mathrm{g}_s(x_s;\theta_s), y_s) + \mathcal{L}_i(\mathrm{g}_i(x_i;\theta_i), y_i) + \lambda\|\theta_s - \theta_i\|_2^2, \tag{3}$$

where y_s and y_i are labels of input sketches x_s and images x_i, and λ is coefficient of soft sharing loss. For hard weight-sharing, i.e., $\theta = \theta_s = \theta_i$, the third loss is equal to 0. Thus, the gradient of θ is calculated as

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}_s(\mathrm{g}_s(x_s;\theta), y_s) + \mathcal{L}_i(\mathrm{g}_i(x_i;\theta), y_i)}{\partial \theta}, \tag{4}$$

which will be dominated by image modality as the image contains visual cues such as color and texture that makes it easier to be classified than the iconic and abstract sketch. This can be verified in Fig. 3a that the training loss (i.e., red lines) of image modality decreases faster than the one of sketch modality. Similarly, the high validation accuracy (i.e., red line in Fig. 3b) and low retrieval performance in Table 4 indicate the learning of hard weight-sharing is dominated by image modality, namely, the learning is imbalanced. For soft weight-sharing, the gradient of θ_s can be described as

$$\frac{\partial \mathcal{L}}{\partial \theta_s} = \frac{\partial \mathcal{L}_s(\mathrm{g}_s(x_s;\theta_s), y_s) + \lambda\|\theta_s - \theta_i\|_2^2}{\partial \theta_s}. \tag{5}$$

Similarly, the gradient of θ_i is

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = \frac{\partial \mathcal{L}_i(\mathrm{g}_i(x_i;\theta_i), y_i) + \lambda\|\theta_s - \theta_i\|_2^2}{\partial \theta_i}. \tag{6}$$

We can find that it can achieve better trade-off between sketches and images, which significantly alleviate the problem of imbalanced learning. Specifically, when the image model prefers to learn the parameters suitable for image modality during each optimization step, the sketch model has a degree of liberty to learn the parameters suitable for sketch modality. Hence, the soft weight-sharing loss produced by the difference of parameters will prevent the image model learning too much from image modality. It can be verified by the higher training losses (i.e., blue lines) in Fig. 3a. Moreover, the acceptable validation accuracy (i.e., blue line in Fig. 3b) and high retrieval performance imply the imbalanced learning of hard weight-sharing is effectively alleviated. On the other hand, it is also necessary to employ soft weight-sharing, since the modalities are related but not the same, especially at the low-level stages. Therefore, the learning of natural images can be guided with the weights of sketch encoder, which makes themselves capable of capturing similar geometric patterns.

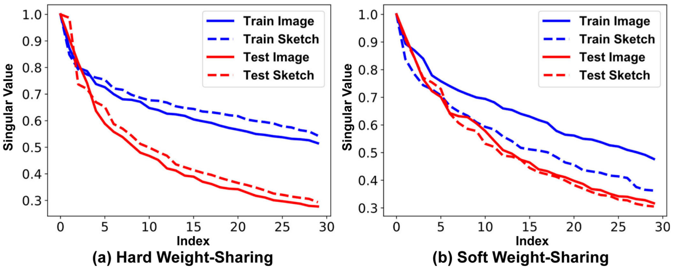


Fig. 4. Analysis of top-30 singular values (max-normalized in each modality) with 64 dimensional features on Sketchy.

Then we confirm the advantage of using soft weight-sharing strategy from Fig. 4 by analyzing Singular Value Decomposition (SVD) on sketch and image representations, respectively. Based on the spectral analysis, feature representations can be decomposed into eigenvectors with importance quantified by the corresponding singular values. Obviously, there are more large singulars in training sketch features than image ones when adopting hard weight-sharing strategy. However, the sketches are sparse and should need fewer large singulars and its eigenvectors to describe. From this perspective, we make sure the learning is imbalanced (as illustrated in Fig. 3) and over-fitting is occurred (as illustrated in Fig. 4) when adopting hard weight-sharing, as the feature backbone also learns more large singulars at training stage than at testing phase to describe sketches. In contrast, it is more reasonable that there are less large singulars in training sketch features when utilizing soft weight-sharing strategy. Besides, the curves of training and testing sketch features are not far apart, indicating the sketch encoder trained on seen categories can be effectively transferred on unseen ones.

Furthermore, we explicitly present a general criterion for multi-modal zero-shot learning, namely adopting coupled modules for exploring modality-common knowledge while independent ones for capturing modality-specific information. Experimental results show that it can be seamlessly employed in other related tasks, e.g., zero-shot and unsupervised domain adaptation.

3.3 Feature Embedding Network

The original dimension of sketch and image features after encoding is usually too large, e.g., 2048, to perform retrieval in realistic scenario. Hence, to reduce the dimension of extracted feature for subsequent cross-modal retrieval, a common network $\mathcal{G}_R$ parameterized with θ_R is introduced to generate retrieval features. Specifically, the retrieval feature R^S of sketch and the retrieval feature R^I of image can be defined as

$$R^S = \mathcal{G}_R(F^S;\theta_R), \quad R^I = \mathcal{G}_R(F^I;\theta_R). \tag{7}$$

Meantime, to improve feature discriminability, the standard classification and knowledge distillation [37] are employed by following the SAKE [20], which studies the connection between zero-shot learning and incremental learning, and introduces the distillation loss to effectively prevent the catastrophic forgetting problem in ZS-SBIR. Mathematically, each one can be implemented with single fully connected layer, which can be formulated as

$$O^C = \mathcal{G}_C(R;\theta_C), \quad O^T = \mathcal{G}_T(R;\theta_T), \tag{8}$$

where O^C and O^T stand for the outputs of classification branch ($\mathcal{G}_C$ parameterized with θ_C) and knowledge distillation branch ($\mathcal{G}_T$ parameterized with θ_T), respectively. For simplicity, the input R can be either sketch retrieval feature R^S or image retrieval feature R^I.

Based on the outputs of classification and knowledge distillation branch, we compute the discrimination loss, which includes classification loss and distillation loss, as

$$\begin{gathered}\mathcal{L}_{DIS} = \frac{1}{N}\sum_{i=1}^{N} -Y_i^C \log P_i^C + \frac{1}{N}\sum_{i=1}^{N} -Y_i^T \log P_i^T, \\ P_i^C = \mathrm{Softmax}(O_i^C), \quad P_i^T = \mathrm{Softmax}(O_i^T),\end{gathered} \tag{9}$$

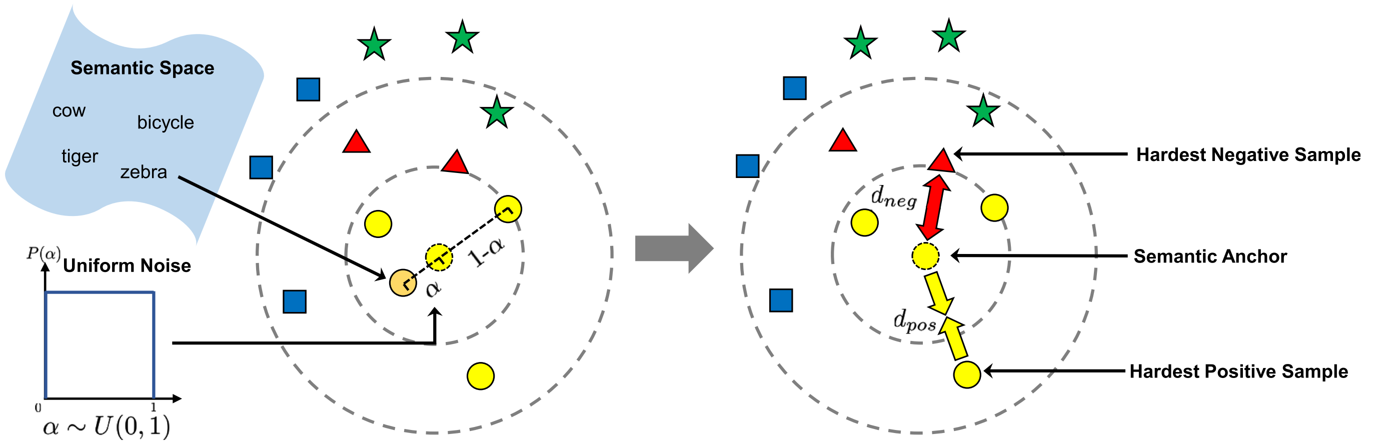
where Y_i^C and P_i^C are sample-level one-hot label and normalized probability of classification branch, respectively. Similarly, Y_i^T and P_i^T are sample-level supervised signal from teacher network, and normalized probability of knowledge distillation branch.

The teacher network utilized for knowledge distillation can be any models pre-trained on ImageNet. Following SAKE [20], we adopt the same model of image encoder as teacher network to regularize the feature embedding generated from student network, i.e., image encoder. Here, the teacher network is employed to address the catastrophic forgetting, which means eliminating most previously acquired knowledge with fresh ones, when fine-tuning the pre-trained models on target image dataset. Hence, to keep their knowledge learned from ImageNet, which maybe useful for unseen categories, we force the image encoder to behave like the teacher network. Experimental results validates its effectiveness even if some categories are not present in the 1,000 classes of ImageNet.

3.4 Semantic Metric Learning

To improve feature discriminability for cross-modal retrieval, prior works [10], [14] adopt local metric learning, e.g., triplet metric. Besides, they also tend to utilize global semantic constraint, e.g., semantic regression, to improve feature transferability. However, most works attempt to optimize these two separate losses with empirical coefficient, which is complicated for training. In following, we first give an in-depth analysis of these two losses and then propose a concise yet effective semantic metric to simultaneously incorporate local metric learning and global semantic constraint into a unified formula, as illustrated in Fig. 5.

The goal of metric learning, e.g., triplet metric learning [38], [39] in this paper, is to force the semantically similar samples close in embedding space while the dissimilar ones far away. Furthermore, in order to efficiently take advantage of triplets, for each anchor sample, we select the hardest positive (the farthest positive) and hardest negative (the nearest negative) samples within each batch. As for semantic regression, previous methods usually drive the embedding of retrieval features close to their category-level semantic vectors. Based on the analysis above, we observe that the uniqueness of semantic vector in each category makes it greatly suitable to be a global anchor. Hence, we propose a unified formula and mathematically define it as



(a) Semantic Anchor Generation

(b) Semantic Metric Learning

Fig. 5. The illustration of our proposed semantic metric within each batch. First, the semantic anchor (e.g., the yellow point with dotted circle) is generated by taking semantic representation and uniform noise as inputs. Then the hardest negative and positive sample are determined based on the new anchor. We believe that new anchor will be more central in its class center than other samples. Finally, we optimize the embedding by minimizing the positive distance (e.g., d_{pos}) and maximizing the negative one (e.g., d_{neg}).

$$\begin{aligned} R_j^{SA} &= \mathcal{G}_A(Y_j^S;\theta_A), \\ \tilde{R}_j^A &= \alpha R_j^{SA} + (1-\alpha)R_j^A, \alpha \sim U(0,1), \\ R_j^P &= \mathop{\arg\max}_{k, Y_k^C = Y_j^C} \mathrm{Dist}(\tilde{R}_j^A, R_k), \\ R_j^N &= \mathop{\arg\min}_{k, Y_k^C \neq Y_j^C} \mathrm{Dist}(\tilde{R}_j^A, R_k), \end{aligned} \tag{10}$$

where $\mathcal{G}_A$ is the single fully connected layer parameterized with θ_A to translate semantic vector Y_j^S into feature embedding R_j^{SA}. Then the new semantic-based anchor $\tilde{R}_j^A$ is generated via uniform interpolation. Based on it, R_j^P and R_j^N are selected as the hardest positive sample and the hardest negative one, respectively. Then the semantic metric loss can be described as

$$\mathcal{L}_{SEM} = \sum_{j=1}^{B} \delta(\mathrm{Dist}(\tilde{R}_j^A, R_j^P) - \mathrm{Dist}(\tilde{R}_j^A, R_j^N)). \tag{11}$$

We adopt standard euclidean distance as the Dist function, and Softplus activation as the δ function. Here, we advocate the generation of anchor $\tilde{R}_j^A$ has two advantages: (1) it naturally incorporates the global semantic constraint into local metric learning, overcoming the instability of local metric learning and meantime improving the transferability of features, (2) the uniform interpolation acts like data augmentation, which will help the training of whole model.

3.5 Training and Inference

After the description of our framework and losses, the full objective of our approach can be formulated as

$$\mathcal{L} = \lambda_{SWS}\mathcal{L}_{SWS} + \lambda_{DIS}\mathcal{L}_{DIS} + \lambda_{SEM}\mathcal{L}_{SEM}, \tag{12}$$

where λ_{SWS}, λ_{DIS} and λ_{SEM} are coefficients to balance the overall performance, and all losses are trained end-to-end simultaneously. Here, after obtaining retrieval features, the discrimination loss as well as the semantic metric loss are employed atop of them, without distinguishing between sketches and images. The overall optimization can be summarized as Algorithm 1.

At testing stage, given a sketch query from testing set, we retrieve corresponding images from the retrieval gallery and rank the results based on their cosine distances to the sketch query. To make fair comparison to prior works [12], [20], the performance of binary hashing, which is encoded as binary code from real-valued feature to accelerate retrieval speed, is also evaluated. Specifically, we generate the hash codes by applying ITerative Quantization (ITQ) [40] algorithm on real-valued retrieval features.

Algorithm 1. Optimization Algorithm.

```
Input: Training dataset $D^{Tr} = \{I^{Tr}, S^{Tr}, Y^C, Y^S\}$, maximum training epochs $N_E$, batch size $N_B$, $\lambda_{SWS}$=1,000, $\lambda_{DIS} = 1$, $\lambda_{SEM}$=1
Output: Learned model parameters $\theta_S, \theta_I, \theta_R, \theta_C, \theta_T, \theta_A$
Initialize parameters $\theta_S, \theta_I, \theta_R, \theta_C, \theta_T, \theta_A$
repeat
    Sample mini-batch data $\{I_i, S_i, Y_i^C, Y_i^S\}_{i=1}^{N_B}$
    Forward model to generate $R_i^S, R_i^I, O_i^C, O_i^T, R_i^{SA}$
    Calculate $\mathcal{L}^{SWS}, \mathcal{L}^{DIS}, \mathcal{L}^{SEM}$
    $\mathcal{L} \leftarrow \lambda_{SWS}\mathcal{L}_{SWS} + \lambda_{DIS}\mathcal{L}_{DIS} + \lambda_{SEM}\mathcal{L}_{SEM}$
    Update $\theta_S \xleftarrow{+} -\nabla_{\theta_S}(\mathcal{L})$
    Update $\theta_I \xleftarrow{+} -\nabla_{\theta_I}(\mathcal{L})$
    Update $\theta_R \xleftarrow{+} -\nabla_{\theta_R}(\mathcal{L})$
    Update $\theta_C \xleftarrow{+} -\nabla_{\theta_C}(\mathcal{L})$
    Update $\theta_T \xleftarrow{+} -\nabla_{\theta_T}(\mathcal{L})$
    Update $\theta_A \xleftarrow{+} -\nabla_{\theta_A}(\mathcal{L})$
until max training epochs $N_E$ is reached;
```

4 Experiments

4.1 Datasets and Settings

In this paper, three popular large-scale benchmarks, i.e., Sketchy [8], TU-Berlin [26] and QuickDraw [14], are adopted to evaluate our proposed approach. The overall statistics of them are reported in Table 1 and the qualitative comparison is illustrated in Fig. 6.

Sketchy [8] originally consists of 75,471 sketches and 12,500 images from 125 categories. In [41], the extended version is released by collecting extra 60,502 images from

TABLE 1
The Statistics of Three Datasets

	Sketchy	TU-Berlin	QuickDraw
#Sketches	75,471	20,000	330,000
#Sketches per Class	~ 500	80	3,000
#Images	73,002	204,489	204,000
#Images per Class	600-700	~ 764	$\sim 1,854$
#Classes	125	250	110
#Training Classes	100/104	220	80
#Testing Classes	25/21	30	30

ImageNet, yielding a total of 73,002 natural images. Following [12], [20], we randomly select 25 categories for testing and the remaining 100 categories for training, which we refer as split1. Besides, if the selected testing categories are also present in the ImageNet dataset, it will violate the assumption of ZSL. Therefore, a more careful and challenging split, i.e., split2, is utilized in [13], which consists of 21 testing categories that are not present in the ImageNet dataset and 104 training categories.

TU-Berlin [26] only consists of 20,000 free-hand sketches evenly distributed over 250 categories for sketch classification and recognition. To perform cross-modal retrieval, the extended version containing 204,489 natural images is adopted in [42]. Following [12], [20], we randomly select 30 categories for testing and the rest 220 categories for training. Here, each testing category is required to have at least 400 natural images to guarantee the retrieval.

QuickDraw [14] is recently released by taking the practical factors, e.g., large domain gap and larger scale, into consideration. It consists of 330,000 sketches and 204,000 images from 110 categories. Following [14], we adopt 30 categories for testing and the remaining 80 categories for training. There are 3,000 sketches and about 1,854 images in each category, which will help to better understand the problem of ZS-SBIR in realistic scenario.

For evaluation criteria, mean Average Precision (mAP) and Precision (Prec) are employed to compare the performance of all methods.

4.2 Implementation Details

All experiments are implemented with PyTorch [44] package on two TITAN XP GPUs. We adopt the ResNet-50 [45] pre-trained on ImageNet as the backbone. It is also used as teacher network for knowledge distillation. Besides, to evaluate the effect of using different feature backbones, we also conduct experiments with VGG networks, SE-ResNet-50 [46] and CSE-ResNet-50 [29] by following [14], [20]. During training stage, paired sketches and images are loaded simultaneously and processed for subsequent classification, knowledge distillation and metric learning. We extract word vectors [36] as category-level semantic information via the text model pre-trained on Googel News Dataset [47].

We train the model with Adam [48] optimizer and set its weight decay as 5×10^{-4}. The batch size and maximum number of training epochs are 32 and 10, respectively. The learning rate starts at 1×10^{-4} and exponentially decays to 3×10^{-6} during training. λ_{DIS} and λ_{SEM} are set to 1 while λ_{SWS} is set to 1,000, which is empirically determined by the retrieval performance. The best model is selected through sketch recognition on validation set, which is a partial of training sketches. It should be noted this validation is more reasonable and fair than performing retrieval on the half of testing samples in [15].

4.3 Comparison With Existing Methods

Quantitative Comparison. To make comprehensive comparison of our proposed approach and existing methods, the performance of three prior works on SBIR, i.e., GN-Triplet [8], DSH [41] and EMS [29], two zero-shot methods, i.e., ZSH [43] and SAE [18] and five prior works on ZS-SBIR, i.e., ZSIH [12], CAAE and CVAE [13], Doodle [14], SEM-PCYC [15] and SAKE [20] are demonstrated in Table 2. We can observe that our approach consistently outperforms all these methods by a large margin, even on the challenging TU-Berlin and QuickDraw datasets. Specifically, it outperforms state-of-the-art methods by more than 12% on Sketchy dataset and 2% on TU-Berlin dataset using 64-bit binary hashing codes. It also outperforms state-of-the-art methods on more practical QuickDraw dataset, which yields more than 6% improvement in terms of retrieval performance. The way of sketch creation plays a pivotal role in retrieval performance, as illustrated in Fig. 6. Specifically, when creating sketches of Sketchy dataset, the crowd workers have corresponding photos as reference at drawing time, which produces highly detailed or less abstract sketches. For TU-Berlin and QuickDraw dataset, workers are asked to draw sketches giving them only the name of category, thereby generating more abstract and diverse sketches. Hence, the quality of sketches, e.g., the degree of abstract and diverse, severely affects the retrieval performance.

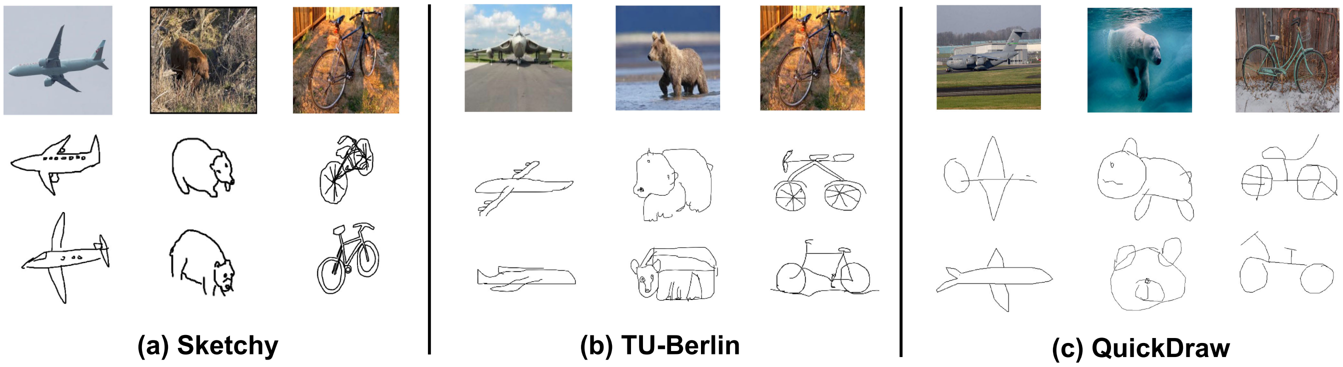


Fig. 6. The Qualitative comparison of Sketchy, TU-Berlin and QuickDraw datasets.

TABLE 2
The Comparison of ZS-SBIR Performance Between TCN and Existing methods

	Method	Dimension	TU-Berlin		Sketchy Split1		Sketchy Split2		QuickDraw	
			mAP@all	Prec@100	mAP@all	Prec@100	mAP@200	Prec@200	mAP@all	Prec@100
SBIR	GN-Triplet [8]	1024	0.189	0.241	0.211	0.310	0.083	0.169	-	-
	DSH [41]	64†	0.122	0.198	0.164	0.227	0.059	0.153	-	-
	EMS [29]	512	0.259	0.369	-	-	-	-	-	-
		64†	0.165	0.252	-	-	-	-	-	-
ZSL	SAE [18]	300	0.161	0.210	0.210	0.302	0.136	0.238	-	-
	ZSH [43]	64†	0.139	0.174	0.165	0.217	-	-	-	-
ZS-SBIR	ZSIH [12]	64†	0.220	0.291	0.254	0.340	-	-	-	-
	CAAE [13]	4096	-	-	0.196	0.284	0.156	0.260	-	-
	CVAE [13]	4096	-	-	-	-	0.225	0.333	0.003	-
	Doodle [14]	256	0.109	-	-	-	0.460	0.370	0.075	-
	SEM-PCYC [15]	64	0.297	0.426	0.349	0.463	-	-	-	-
		64†	0.293	0.392	0.344	0.399	-	-	-	-
	SAKE [20]	512	0.475	0.599	0.547	0.692	0.497	0.598	-	-
		64†	0.359	0.481	0.364	0.487	0.356	0.477	-	-
ZS-SBIR	**TCN**	512	**0.495**	**0.616**	**0.616**	**0.763**	**0.516**	**0.608**	**0.140**	**0.231**
		64†	**0.381**	**0.506**	**0.488**	**0.644**	**0.401**	**0.514**	**0.110**	**0.150**

Here, "†" denotes experiments using binary hashing codes while the remaining use real-valued features for retrieval. "-" indicates the results are not presented by the authors.

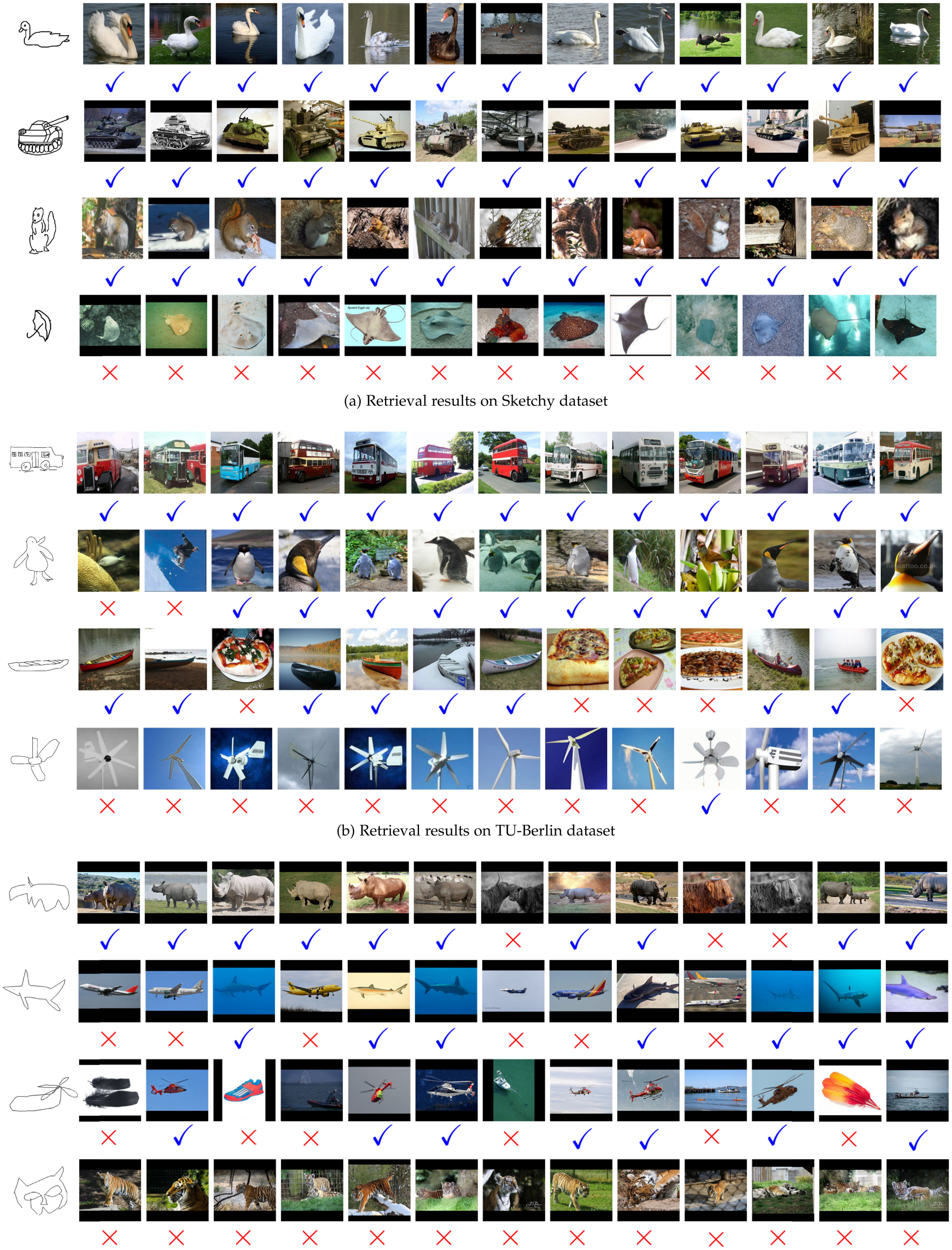
To evaluate the effectiveness of TCN, we also conduct experiments on the more challenging Sketchy split2, which strictly selects the categories out of the ImageNet as testing set. The results reported in Table 2 show that our proposed TCN significantly beats the existing state-of-the-art methods, which proves our method can learn better transferable cross-modal representation for ZS-SBIR.

Qualitative Comparison. The retrieval results on three datasets are demonstrated in Fig. 7. We observe that most retrieved candidates belong to the same category of their queries. However, the fourth sample on Sketchy dataset, *umbrella*, has too similar shape with the *ray* (a kind of fish) to obtain wrong retrievals. In the fourth row of TU-Berlin dataset, the *fan* fails to retrieve correct images as its shape is also similar to *windmill*. Besides, the *canoe*, the third row of TU-Berlin dataset, has similar shape with the piece of *pizza* or similar warping edge with whole *pizza*, thereby leading to wrong retrieval results. The main reason maybe lies in that the model confuses about the examples with similar shapes as the sketches lack visual cues such as color, texture and background. On QuickDraw dataset, the sketches are rough conceptual abstractions of images, e.g., the *raccoon* in the fourth row, which cannot even be identified by the human. Similarly, the candidates are mostly determined by their shape, leading to unsatisfied performance when the query and candidates have similar shape, e.g., the *shark* and *airplane* in the second row. On the other hand, it also confirms that TCN effectively forces the image encoder to capture geometric information by the guiding of sketch branch. We also provide the retrieval results of the selected 8 categories (i.e., cup, swan, harp, squirrel, snail, ray, pineapple, volcano) on Sketchy in Fig. 8 to help to understand how the model works.

4.4 Results Analysis

Feature Backbone. In order to make fair comparison with previous works [13], [14], [15], we also adopt VGG network as the feature backbone as well as teacher network. At the meantime, we keep everything else the same. Specifically, we employ VGG16 and VGG19 network with batch normalization layers, which can be easily obtained from the PyTorch repository. The effect of utilizing spatial attention module [10], [14] is also evaluated in the backbone. Moreover, following prior works [20], [29], SE-ResNet-50 and CSE-ResNet-50 are also evaluated for fair comparison. First, we observe that ResNet-50 performs better than other architectures in our proposed TCN, as it has more advanced network structures than VGG networks while contains fewer parameters, which are irrelevant with convolutional operation for modeling patterns, than its modified versions. Second, our approach still beats the method, SEM-PCYC [15], via adopting the same VGG backbone. Finally, using spatial attention module in VGG networks will slightly increase or achieve comparable performance. The comparison is summarized in Table 3.

Weight-Sharing Strategy. To verify the effectiveness of each component, we conduct detailed ablation studies in Table 4. First, the performance is poor when adopting no weight-sharing strategies. We argue that the model is over-fitting, since its validation results of sketch recognition we observed are pretty high. Then the model using hard weight-sharing strategy obtains slightly better performance as it forces the model using same parameters except modality-specific batch normalization layers to jointly modeling sketches and images. As we all known, the discrimination loss atop of the feature encoders can be more easily optimized on image modality than the sketch, which makes these parameters more suitable for the natural images and ignores the learning of sketches. To balance the joint modeling, we simply attempt to weight the losses of different modalities. However, the performance is still poor whether the weight coefficient increases or decreases. Our proposed soft weight-sharing strategy outperforms the former variants by a large margin via guiding the learning of natural images with sketch patterns, which strongly verifies its effectiveness as well as superiority. We further confirm that our approach can extract heterogeneous features appropriately through



(a) Retrieval results on Sketchy dataset

(b) Retrieval results on TU-Berlin dataset

(c) Retrieval results on QuickDraw dataset

Fig. 7. Top-13 retrieval results of testing samples on three large-scale datasets by using 64 dimensional real-valued features. The blue ticks stand for correctly retrieved candidates while the red crosses denote wrong retrievals.

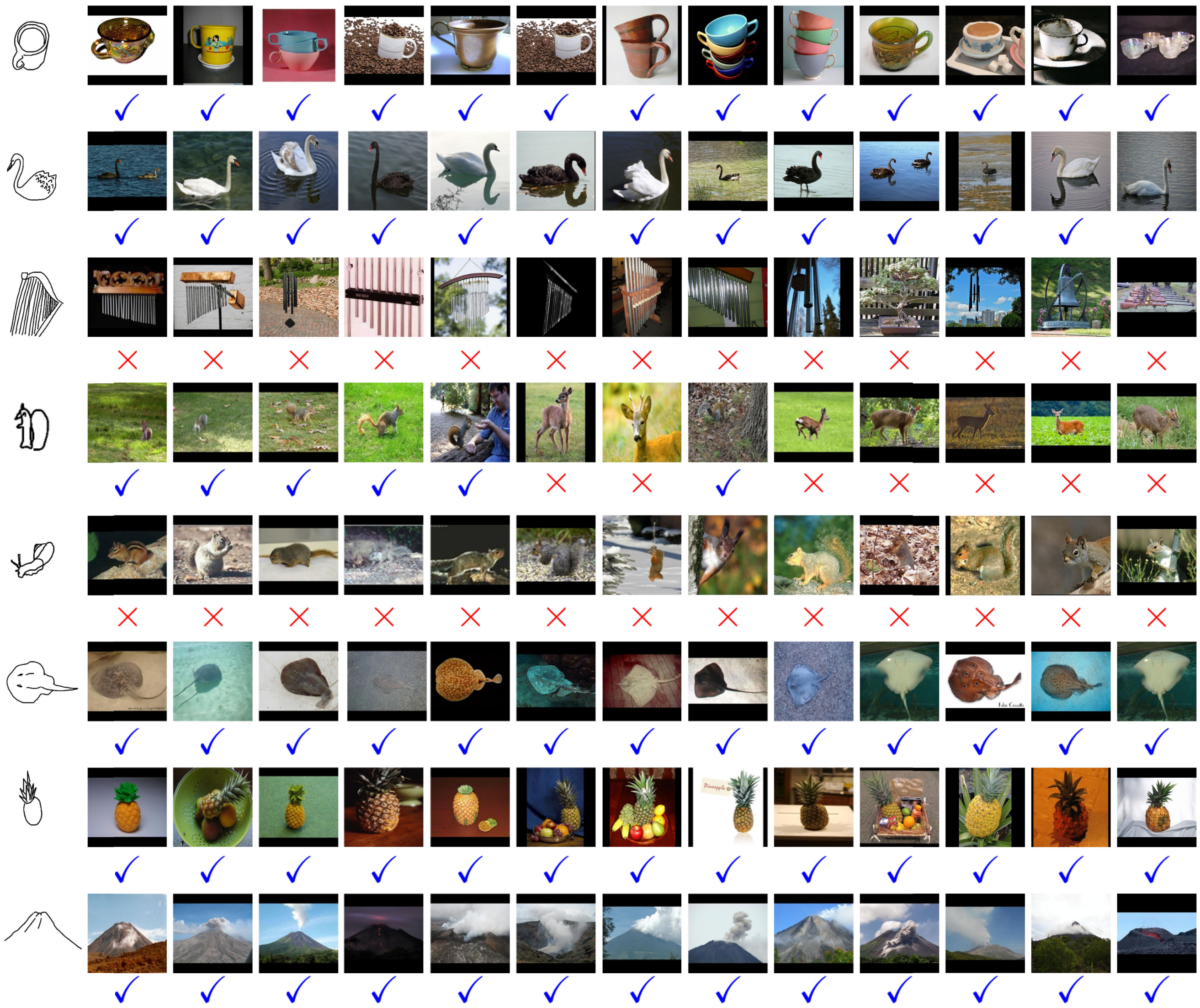


Fig. 8. Retrieval results of the random selected 8 categories on Sketchy dataset.

the analysis of singular values of each modality in Fig. 9. This character maybe be helpful when exploring what is the best architecture of ZS-SBIR. The t-SNE visualization also shows the effectiveness of our proposed weight-sharing strategy in Fig. 10. Compared to the counterparts without using weight-sharing strategy, our method with weight-sharing strategy significantly improves the retrieval performance on most categories, as illustrated in Table 5. It should be noted that the *wine bottle* and *butterfly* have the biggest (i.e., 316.5%) and smallest (i.e., -5.0%) relative increase. In addition, we conduct experiments to track the training losses on sketch and image modality as well as the validation accuracy of sketch that varies with the number of iterations, as illustrated in Fig. 3. For both hard and soft weight-sharing, the loss on image modality decreases faster than on sketch modality, indicating the model prefers to learn parameters suitable for image modality during each optimization step. For hard weight-sharing, the training losses are very small and the validation accuracy of sketch is pretty high. However, the retrieval performance is poor, which means over-fitting is occurred at training stage. In contrast, using soft weight-sharing achieves much better retrieval performance at the expense of large training losses on sketch and image.

TABLE 3
Performance Comparison of Using Different Backbones in TCN with 64 Dimensional features

Backbone	Attention	Sketchy Split1	
		mAP@all	Prec@100
VGG16		0.408	0.524
VGG16	✓	0.414	0.523
VGG19		0.435	0.543
VGG19	✓	0.442	0.549
ResNet-50		**0.582**	**0.711**
SE-ResNet-50		0.499	0.627
CSE-ResNet-50		0.509	0.625

Here, Attention denotes spatial attention module [10], which aims to focus on fine-grained details.

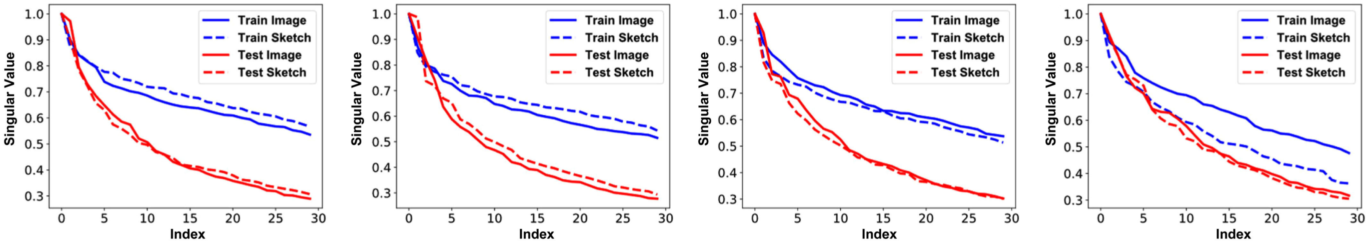
Semantic Metric. From Table 4, we can observe that using original batch hard triplet metric learning yields no significant improvement of performance. We advocate that local feature metric learning helps SBIR but has difficulty in addressing ZS-SBIR, since it lacks the global guidance to

TABLE 4
Ablation Studies of Individual Component in TCN With 64 Dimensional Features

Component	Variant	Sketchy Split1		TU-Berlin	
		mAP	Prec	mAP	Prec
		@all	@100	@all	@100
Weight-Sharing	w/o	0.368	0.449	0.318	0.379
	Hard (0.1 : 1)	0.403	0.508	0.337	0.430
	Hard (1 : 1)	0.399	0.520	0.335	0.438
	Hard (10 : 1)	0.367	0.488	0.333	0.445
	Soft	**0.582**	**0.711**	**0.470**	**0.569**
Semantic Metric	w/o	0.557	0.691	0.465	0.563
	BHT	0.567	0.702	0.466	0.562
	BHT+SR (1 : 0.1)	0.571	0.703	0.470	0.566
	BHT+SR (1 : 1)	0.570	0.702	0.470	0.565
	BHT+SR (1 : 10)	0.569	0.700	0.469	0.565
	SBHT	**0.582**	**0.711**	**0.470**	**0.569**
Teacher Network	w/o	0.480	0.649	0.457	0.562
	With	**0.582**	**0.711**	**0.470**	**0.569**

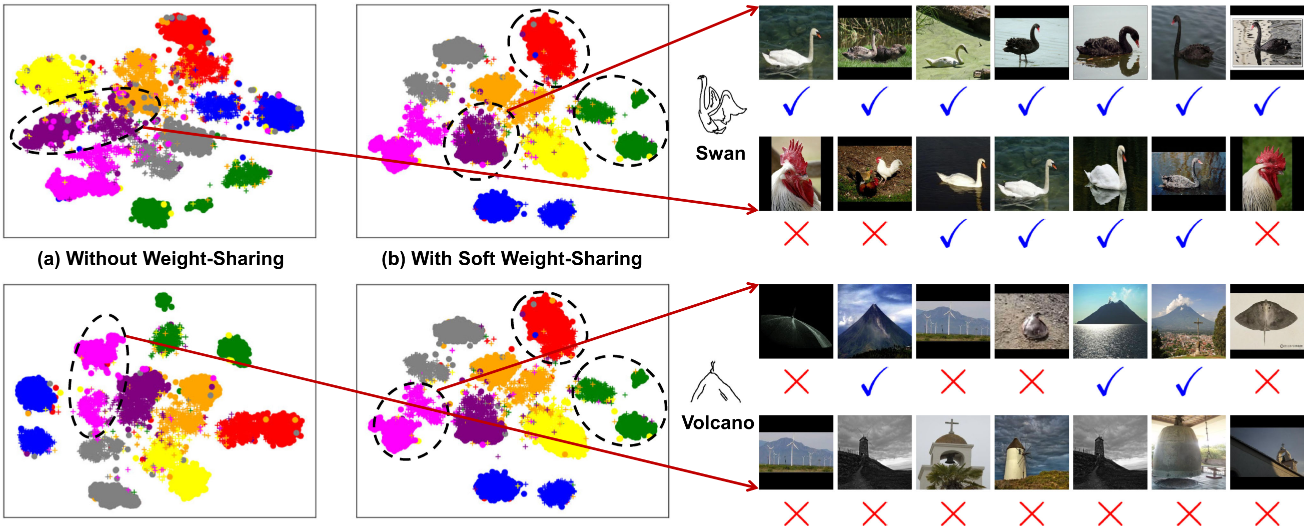
Here, Hard and Soft denote hard weight-sharing and soft weight-sharing strategies, respectively. Besides, BHT, SBHT and SR stand for traditional batch hard triplet metric, our proposed semantic batch hard triplet metric and semantic regression.

ensure the transferablity of whole model. Then we implement experiments through combining separate batch hard triplet and semantic regression. It slightly outperforms the method of using local metric learning, i.e., batch hard triplet, no matter the weight coefficient increases or decreases. Furthermore, when we integrate the semantic information into traditional metric learning, it remarkably boosts the performance. This comparison greatly demonstrate the superiority of our proposed semantic metric, which is extremely concise yet effective. Besides, more clustered



(a) Without Weight-Sharing (b) Hard Weight-Sharing (c) λ_{SWS}=10.0 (d) λ_{SWS}=1000.0

Fig. 9. The results of top-30 singular values (max-normalized in each modality) with 64 dimensional features on Sketchy.



(a) Without Weight-Sharing (b) With Soft Weight-Sharing

(c) Without Semantic Metric (d) With Semantic Metric

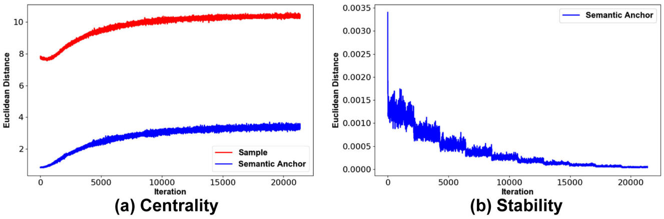
Fig. 10. Left: t-SNE results of using 64 dimensional features on the random selected 8 testing categories of Sketchy. The dot and cross stand for natural image and sketch, respectively. Right: corresponding retrieval examples.

TABLE 5
The Performance Improvement (Relative Increase) of the Random Selected 8 Categories at Testing Stage With 64 Dimensional Features on Sketchy

Comparison	Cup	Swan	Harp	Squirrel	Snail	Ray	Pineapple	Volcano	Rifle	Scissors
WS	32.9%	8.4%	189.3%	14.7%	57.8%	106.3%	48.2%	59.6%	23.7%	52.4%
SM	1.3%	11.5%	10.6%	-5.7%	-1.4%	**-9.7%**	0.4%	10.0%	3.6%	1.7%
Comparison	**Parrot**	**Windmill**	**Teddy Bear**	**Tree**	**Wine Bottle**	**Deer**	**Chicken**	**Airplane**	**Wheelchair**	**Tank**
WS	29.7%	84.8%	279.3%	141.7%	**316.5%**	36.2%	77.9%	34.3%	30.1%	50.7%
SM	14.9%	-0.7%	4.7%	9.5%	7.2%	1.4%	**17.5%**	11.1%	8.7%	8.5%
Comparison	**Umbrella**	**Butterfly**	**Camel**	**Horse**	**Bell**					
WS	264.1%	**-5.0%**	68.2%	64.0%	239.7%					
SM	-7.8%	8.4%	-1.4%	10.9%	2.1%					

WS stands for the comparison of soft weight-sharing versus without weigh-sharing and SM represents the comparison of semantic metric versus without semantic metric.

map can be observed in Fig. 10, which also confirms its effectiveness. In contrast to the counterparts without using semantic metric, the utilization of proposed semantic metric improves the retrieval performance on most categories, as illustrated in Table 5. It is worth nothing that the *chicken* and *ray* have the biggest (i.e., 17.5%) and smallest (i.e., -9.7%) relative increase. We also conduct experiments to verify that semantic anchor is more central and stable than sketch or image sample in each batch, as illustrated in Fig. 11. Specifically, we observe that semantic anchors are much closer to the corresponding class centers than batch samples, which means they are more central than other samples, i.e., images and sketches in each batch, as illustrated in Fig. 11a. The euclidean distance between semantic anchors of adjacent iteration is very small (i.e., less than 0.0035 in Fig. 11b), which indicates that the semantic anchor changes slightly during training stage.



(a) Centrality

(b) Stability

Fig. 11. The euclidean distance versus the number of iterations with 64 dimensional features on Sketchy.

Teacher Network. The extra prior knowledge contained in teacher network provides strong regularization for feature embedding, which is proved to be effective among plentiful tasks, including ZS-SBIR. It should be noted that our approach still outperforms most methods even without adopting teacher network as illustrated in Table 4, reflecting the superiority of our proposed TCN.

Length of Hashing Code. In Fig. 12, we compare our approach with existing zero-shot hashing methods, e.g., ZSH [43], ZSIH [12] and SAKE [20], on Sketchy dataset. Specifically, the TCN outperforms state-of-the-art methods by more than 10.8%, 12.4% and 13.8% with 32-bit, 64-bit and 128-bit hashing codes. As expected, the performance gap increases along with the length of hashing codes, which strongly proves the effectiveness of our proposed approach.

Effect of Coefficient. In Fig. 13, we analyze the effect of hyper-parameters λ_{SWS}. For simplicity, we fix the coefficients λ_{DIS} and λ_{SEM}, which means $\lambda_{DIS} = \lambda_{SEM} = 1$ for all experiments. When $\lambda_{SWS} = 0$, the model adopts two individual base feature extractors to extract features, which gets poor performance due to the over-fitting. The performance increases along with the λ_{SWS} and reaches the peak value at 1,000.

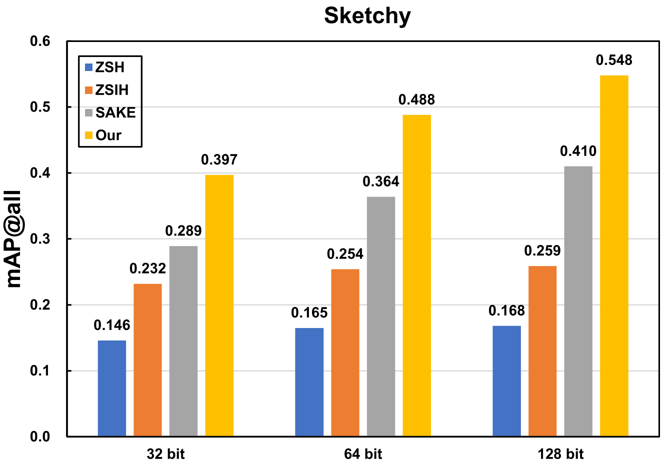


Fig. 12. The results of mAP@all between TCN and existing zero-shot hashing methods. 32 bit, 64 bit, and 128 bit denote the different length of binary hashing codes.

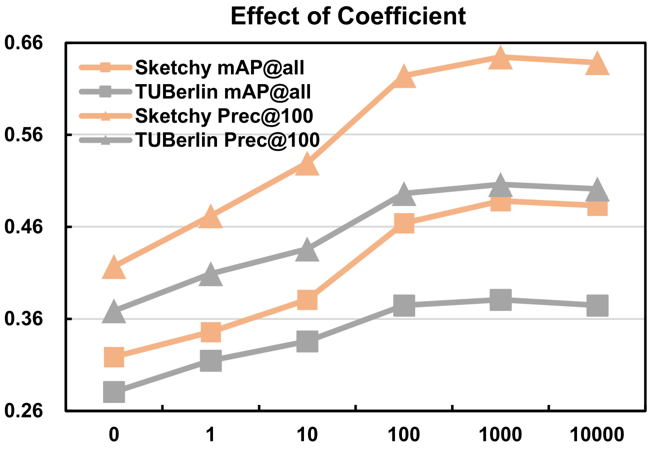


Fig. 13. The effect of using different λ_{SWS}. Here, the results of mAP@all are compared with 64-bit binary hashing codes.

TABLE 6
Classification Accuracy of Digit Images

Method	MNIST (S) → MNIST-M (T)
Source Only	52.2%
DANN*	92.4%
DANN*+TCN	**93.9%**

*Here, S and T stand for source and target domain, respectively. And the * denotes it is implemented by the public code.*

Extension of TCN. We believe that our proposed approach can be extended into many tasks, such as zero-shot domain adaptation [49], [50] and unsupervised domain adaptation [51], [52]. Since there are no accessible code for the former task, we conduct experiments on the later one to evaluate the expandability of our proposed strategy. We choose the classic method, DANN [51], as competitors to perform the adaptation from source MNIST to MNIST-M. Here, the target dataset is generated by blending digits from the MNIST over patches randomly extracted from color photos of BSDS500 dataset [53]. From Table 6, we observe that partial soft weight-sharing strategy can significantly boost the classification performance, which shows the great potential of our approach among related tasks.

5 Conclusion

In this paper, we propose a novel transferable coupled network to handle the task of ZS-SBIR, which is able to learn a better transferable cross-modal representation than state-of-the-art methods. Furthermore, we explicitly introduce a general criterion for multi-modal zero-shot learning, which can be seamlessly and effortlessly employed in other related tasks. Moreover, a simple but effective semantic metric is adopted to significantly improve feature discriminability and trasferability.

In the future, we should first dig deep into the relationship between singular values and feature modeling. Next we should devote more efforts on the feature extraction of sketches to address the problems of deformation and visual sparsity. The latter one could be addressed by exploring augmented data, such as the retrieved natural images on Internet. Besides, the way of utilizing auxiliary information, e.g., word vectors or attributes, should also be emphasized to improve the transferability of whole model. Finally, we will also explore the semantic metric in other zero-shot tasks in the future, such as zero-shot classification and zero-shot semantic segmentation.

References

[1] H. Chen and S.-C. Zhu, "A generative sketch model for human hair analysis and synthesis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1025–1040, Jul. 2006.

[2] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Nov. 2009.

[3] 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, 2017.

[4] Y. Li, Y.-Z. Song, T. M. Hospedales, and S. Gong, "Free-hand sketch synthesis with deformable stroke models," *Int. J. Comput. Vis.*, vol. 122, no. 1, pp. 169–190, 2017.

[5] W. Chen and J. Hays, "SketchyGAN: Towards diverse and realistic sketch to image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9416–9425.

[6] L. Tran, J. Kossaifi, Y. Panagakis, and M. Pantic, "Disentangling geometry and appearance with regularised geometry-aware generative adversarial networks," *Int. J. Comput. Vis.*, vol. 127, no. 6–7, pp. 824–844, 2019.

[7] J. M. Saavedra, J. M. Barrios, and S. Orand, "Sketch-based image retrieval using learned keyshapes (LKS)," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 1–11.

[8] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: Learning to retrieve badly drawn bunnies," *ACM Trans. Graphics*, vol. 35, no. 4, pp. 1–12, 2016.

[9] 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.*, 2016, pp. 2460–2464.

[10] 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. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5551–5560.

[11] J. Lei, Y. Song, B. Peng, Z. Ma, L. Shao, and Y.-Z. Song, "Semi-heterogeneous three-way joint embedding network for sketch-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 3226–3237, Sep. 2020.

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

[13] S. K. Yelamarthi, S. K. Reddy, A. Mishra, and A. Mittal, "A zero-shot framework for sketch based image retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 316–333.

[14] S. Dey, P. Riba, A. Dutta, J. Llados, and Y.-Z. Song, "Doodle to search: Practical zero-shot sketch-based image retrieval," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2179–2188.

[15] 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.*, 2019, pp. 5089–5098.

[16] R. Socher, M. Ganjoo, C. D. Manning, and A. Y. Ng, "Zero-shot learning through cross-modal transfer," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 935–943.

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

[18] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4447–4456.

[19] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2251–2265, Sep. 2019.

[20] Q. Liu, L. Xie, H. Wang, and A. L. Yuille, "Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3661–3670.

[21] T. Dutta and S. Biswas, "Style-guided zero-shot sketch-based image retrieval," in *Proc. Brit. Mach. Vis. Conf.*, 2019, pp. 1–12.

[22] X. Xu, M. Yang, Y. Yang, and H. Wang, "Progressive domain-independent feature decomposition network for zero-shot sketch-based image retrieval," in *Proc. 29th Int. Joint Conf. Artif. Intell. 17th Pacific Rim Int. Conf. Artif. Intell.*, 2020, pp. 984–990.

[23] C. Deng, X. Xu, H. Wang, M. Yang, and D. Tao, "Progressive cross-modal semantic network for zero-shot sketch-based image retrieval," *IEEE Trans. Image Process.*, vol. 29, no. 9, pp. 8892–8902, Sep. 2020.

[24] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[25] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[26] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–10, 2012.

[27] R. Hu and J. Collomosse, "A performance evaluation of gradient field HOG descriptor for sketch based image retrieval," *Comput. Vis. Image Understand.*, vol. 117, pp. 790–806, 2013.

[28] 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.*, 2014, pp. 2998–3002.

[29] P. Lu, G. Huang, Y. Fu, G. Guo, and H. Lin, “Learning large euclidean margin for sketch-based image retrieval,” 2018, *arXiv:1812.04275*.

[30] K. Pang *et al.*, “Generalising fine-grained sketch-based image retrieval,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 677–686.

[31] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[32] M. Bucher, S. Herbin, and F. Jurie, “Improving semantic embedding consistency by metric learning for zero-shot classiffication,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 730–746.

[33] H. Jiang, R. Wang, S. Shan, and X. Chen, “Adaptive metric learning for zero-shot recognition,” *IEEE Signal Process. Lett.*, vol. 26, no. 9, pp. 1270–1274, Sep. 2019.

[34] B. Chen and W. Deng, “Hybrid-attention based decoupled metric learning for zero-shot image retrieval,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2745–2754.

[35] J. Shen, H. Wang, A. Zhang, Q. Qiu, X. Zhen, and X. Cao, “Model-agnostic metric for zero-shot learning,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 786–795.

[36] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” 2013, *arXiv:1301.3781*.

[37] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” 2015, *arXiv:1503.02531*.

[38] A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” 2017, *arXiv:1703.07737*.

[39] 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.

[40] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, “Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2916–2929, Dec. 2013.

[41] 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.*, 2017, pp. 2298–2307.

[42] H. Zhang, S. Liu, C. Zhang, W. Ren, R. Wang, and X. Cao, “SketchNet: Sketch classification with web images,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1105–1113.

[43] Y. Yang, Y. Luo, W. Chen, F. Shen, J. Shao, and H. T. Shen, “Zero-shot hashing via transferring supervised knowledge,” in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 1286–1295.

[44] A. Paszke *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” in *Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.

[45] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[46] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[47] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Proc. NeurIPS*, 2013, pp. 3111–3119.

[48] D. P. Kingma and J. Ba, “Adam: a method for stochastic optimization,” *arXiv:1412.6980*, 2014.

[49] K.-C. Peng, Z. Wu, and J. Ernst, “Zero-shot deep domain adaptation,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 764–781.

[50] J. Wang and J. Jiang, “Conditional coupled generative adversarial networks for zero-shot domain adaptation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3375–3384.

[51] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[52] Y. Ganin *et al.*, “Domain-adversarial training of neural networks,” *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.

[53] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, “Contour detection and hierarchical image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2010.

Hao Wang received the BE degree in electronic and information engineering from Hangzhou Dianzi University, Hangzhou, China, and the PhD degree in circuits and systems from Xidian University, Xi’an, China. His main research interests include human action recognition, video and language understanding, and zero-shot learning.

Cheng Deng (Senior Member, IEEE) received the BE, MS, and PhD 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. He is the author and coauthor of more than 100 scientific articles at top venues, including *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Cybernetics*, *IEEE Transactions on Multimedia*, *IEEE Transactions on Systems, Man, and Cybernetics*, ICCV, CVPR, ICML, NIPS, IJCAI, and AAAI. His research interests include computer vision, pattern recognition, and information hiding.

Tongliang Liu (Senior Member, IEEE) is currently a lecturer with the School of Computer Science, University of Sydney. He is heading the Trustworthy Machine Learning Laboratory and is also a visiting scientist with RIKEN AIP. He has authored and coauthored more than 80 research articles including ICML, NeurIPS, ICLR, CVPR, ICCV, ECCV, AAAI, IJCAI, KDD, ICME, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Neural Networks and Learning Systems*, and *IEEE Transactions on Image Processing*. He is and was a (senior-) meta reviewer for many conferences, including NeurIPS, ICLR, AAAI, and IJCAI. His research interests include trustworthy machine learning and its interdisciplinary applications, with a particular emphasis on learning with noisy labels, adversarial learning, transfer learning, unsupervised learning, and statistical deep learning theory. He was the recipient of the Discovery Early Career Researcher Award (DECRA) from Australian Research Council (ARC), the Cardiovascular Initiative Catalyst Award by the Cardiovascular Initiative, best paper awards including the 2019 ICME Best Paper Award and the PacificVis 2021 Best VisNotes Paper Award and was named in the Early Achievers Leaderboard of Engineering and Computer Science by The Australian in 2020.

Dacheng Tao (Fellow, IEEE) is currently the president of the JD Explore Academy and a senior vice president of JD.com. He is also an advisor and chief scientist with the Digital Science Institute, University of Sydney. He has authored or coauthored more than 200 publications in prestigious journals and proceedings at leading conferences, and one monograph in his research fields which include, statistics and mathematics to artificial intelligence and data science. He was the recipient of the 2015 Australian Scopus-Eureka Prize, the 2018 IEEE ICDM Research Contributions Award, and the 2021 IEEE Computer Society McCluskey Technical Achievement Award. He is a fellow of the Australian Academy of Science, AAAS, and ACM.