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 Figure 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 over-fit 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 hand 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.

• H. Wang and C. Deng are with the School of Electronic Engineering, Xidian University, Xi’an, Shaanxi, China (e-mail: haowang.xidian, chdeng.xd@gmail.com).
• T. Liu is with the School of Computer Science, Faculty of Engineering, The University of Sydney, Darlington, NSW 2008, Australia (tongliang.liu@sydney.edu.au).
• D. Tao is with the JD Explore Academy in JD.com, China (dacheng.tao@gmail.com).
• C. Deng and T. Liu are the corresponding authors.
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 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 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, Semi3Net [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 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 specialist 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. Specifically, the first one includes soft weight-sharing convolutional layers (i.e., constrained with $L_{SW S}$) and independent batch normalization layers. Retrieval features are obtained through feature embedding network, along with discrimination loss $L_{DIS}$ and semantic metric loss $L_{SEM}$. At training stage, the guiding signals produced from teacher network as well as benchmark one-hot labels are provided to calculate $L_{DIS}$. Similarly, word vectors and uniform noise are offered to compute $L_{SEM}$.

**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-sharing convolutional layers (i.e., constrained with $L_{SW S}$) and independent batch normalization layers. Retrieval features are obtained through feature embedding network, along with discrimination loss $L_{DIS}$ and semantic metric loss $L_{SEM}$. At training stage, the guiding signals produced from teacher network as well as benchmark one-hot labels are provided to calculate $L_{DIS}$. Similarly, word vectors and uniform noise are offered to compute $L_{SEM}$.
multi-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

\[ F^S = \mathcal{G}_S(S; \theta_S), \quad F^I = \mathcal{G}_I(I; \theta_I), \]  

where \( \mathcal{G}_S \) and \( \mathcal{G}_I \) are sketch encoder parameterized with \( \theta_S \) and image encoder parameterized with \( \theta_I \), 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

\[ L_{SW} = \sum_l I[l \notin \text{BN}] \cdot \|\theta^S_l - \theta^I_l\|^2_2, \tag{2} \]

where \( \theta^S_l \) and \( \theta^I_l \) 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 \( I[l \notin \text{BN}] \) 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 \( \theta_s \) and image network \( g_i \) parameterized with \( \theta_i \), the loss can be formulated as

\[ L = L_s(g_s(x_s; \theta_s), y_s) + L_i(g_i(x_i; \theta_i), y_i) + \lambda \|\theta_s - \theta_i\|^2_2, \tag{3} \]

where \( y_s \) and \( y_i \) are labels of input sketches \( x_s \) and images \( x_i \), and \( \lambda \) 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 \( \theta \) is calculated as

\[ \frac{\partial L}{\partial \theta} = \frac{\partial L_s(g_s(x_s; \theta), y_s)}{\partial \theta} + \frac{\partial L_i(g_i(x_i; \theta), y_i)}{\partial \theta}, \tag{4} \]

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 Figure 3(a) 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 Figure 3(b)) 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 \( \theta_s \) can be described as

\[ \frac{\partial L}{\partial \theta_s} = \frac{\partial L_s(g_s(x_s; \theta_s), y_s)}{\partial \theta_s} + \lambda \|\theta_s - \theta_i\|^2_2, \tag{5} \]

Similarly, the gradient of \( \theta_i \) is

\[ \frac{\partial L}{\partial \theta_i} = \frac{\partial L_i(g_i(x_i; \theta_i), y_i)}{\partial \theta_i} + \lambda \|\theta_s - \theta_i\|^2_2. \tag{6} \]

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 Figure 3(a). Moreover, the acceptable validation accuracy
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3123315, IEEE Transactions on Pattern Analysis and Machine Intelligence

(i.e., blue line in Figure 3(b)) 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.

Then we confirm the advantage of using soft weight-sharing strategy from Figure 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 Figure 5) and over-fitting is occurred (as illustrated in Figure 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 \( g_R \) parameterized with \( \theta_R \) 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

\[
R^S = g_R(F^S; \theta_R), \quad R^I = g_R(F^I; \theta_R). \tag{7}
\]

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

\[
O^C = g_C(R; \theta_C), \quad O^T = g_T(R; \theta_T), \tag{8}
\]

where \( O^C \) and \( O^T \) stand for the outputs of classification branch \( g_C \) parameterized with \( \theta_C \) and knowledge distillation branch \( g_T \) parameterized with \( \theta_T \), 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

\[
\mathcal{L}_{DIS} = \frac{1}{N} \sum_{i=1}^{N} -Y_i^C \log P_i^C + \frac{1}{N} \sum_{i=1}^{N} -Y_i^T \log P_i^T, \tag{9}
\]

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 Figure 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

\[
R^{SA}_{i} = G_{i}(Y^{S}_{i}; \theta_{A}), \\
\tilde{R}^{A}_{j} = \alpha R^{SA}_{j} + (1 - \alpha) R^{A}_{j}, \alpha \sim U(0, 1), \\
R^{P}_{j} = \arg \max_{k, Y^{C}_{k}=Y^{C}_{j}} \text{Dist}(\tilde{R}^{A}_{j}, R_{k}), \\
R^{N}_{j} = \arg \min_{k, Y^{C}_{k}=\neq Y^{C}_{j}} \text{Dist}(\tilde{R}^{A}_{j}, R_{k}),
\]

(10)

where \(G_{i}\) is the single fully connected layer parameterized with \(\theta_{A}\) to translate semantic vector \(Y^{S}_{i}\) into feature embedding \(R^{SA}_{i}\). Then the new semantic-based anchor \(\tilde{R}^{A}_{j}\) is generated via uniform interpolation. Based on it, \(R^{P}_{j}\) and \(R^{N}_{j}\) are selected as the hardest positive sample and the hardest negative one, respectively. Then the semantic metric loss can be described as

\[
L_{SEM} = \sum_{j=1}^{B} \delta(\text{Dist}(\tilde{R}^{A}_{j}, R^{P}_{j}) - \text{Dist}(\tilde{R}^{A}_{j}, R^{N}_{j})).
\]

(11)

We adopt standard Euclidean distance as the Dist function, and Softplus activation as the \(\delta\) function. Here, we advocate the generation of anchor \(\tilde{R}^{A}_{j}\) 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 overall objective of our approach can be formulated as

\[
L = L_{SW} + L_{DIS} + L_{SEM},
\]

(12)

where \(L_{SW}, L_{DIS}\) and \(L_{SEM}\) 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.

### 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 Figure 6.
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 Google News Dataset [47].

We train the model with Adam [48] optimizer and set its weight decay as $5 \times 10^{-4}$. The batch size and maximum number of training epochs are 32 and 10, respectively. The learning rate starts at $1 \times 10^{-4}$ and exponentially decays to $3 \times 10^{-6}$ during training. $\lambda_{DIS}$ and $\lambda_{SEM}$ are set to 1 while $\lambda_{SW}$ 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 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., GNTriplet [8], DSH [41] and EMS [29], two zero-shot methods, i.e., ZSH [45] and SAE [18] and five prior works on ZS-SBIR, i.e., ZSIH [43] and SAE [18], GN-Triplet [8], DSH [41] and EMS [29], two zero-shot methods, i.e., ZSH [45] and SAE [18] and five prior works on ZS-SBIR, i.e., ZSIH [43] and SAE [18], Doodle [14], SEMPCY [15] and SAKIE [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 Figure 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, which produces highly abstract sketches.

<table>
<thead>
<tr>
<th>#Sketches</th>
<th>Sketchy</th>
<th>TU-Berlin</th>
<th>QuickDraw</th>
</tr>
</thead>
<tbody>
<tr>
<td>75,471</td>
<td>20,000</td>
<td>330,000</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>80</td>
<td>3,000</td>
<td></td>
</tr>
<tr>
<td>73,002</td>
<td>204,489</td>
<td>204,000</td>
<td></td>
</tr>
<tr>
<td>600-700</td>
<td>~764</td>
<td>~1,854</td>
<td></td>
</tr>
<tr>
<td>125</td>
<td>250</td>
<td>110</td>
<td></td>
</tr>
<tr>
<td>100/104</td>
<td>220</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>25/21</td>
<td>30</td>
<td>30</td>
<td></td>
</tr>
</tbody>
</table>

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. 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Dimension</th>
<th>TU-Berlin</th>
<th>Sketchy Split1</th>
<th>Sketchy Split2</th>
<th>QuickDraw</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>mAP@all</td>
<td>Prec@100</td>
<td>mAP@all</td>
<td>Prec@100</td>
</tr>
<tr>
<td>SBIR</td>
<td>64</td>
<td>0.189</td>
<td>0.241</td>
<td>0.211</td>
<td>0.310</td>
</tr>
<tr>
<td>DSH [41]</td>
<td>64†</td>
<td>0.122</td>
<td>0.198</td>
<td>0.164</td>
<td>0.227</td>
</tr>
<tr>
<td>EMS [29]</td>
<td>512</td>
<td>0.259</td>
<td>0.369</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EM [10]</td>
<td>64†</td>
<td>0.165</td>
<td>0.252</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ZSL</td>
<td>300</td>
<td>0.161</td>
<td>0.210</td>
<td>0.210</td>
<td>0.302</td>
</tr>
<tr>
<td>SAF [18]</td>
<td>64†</td>
<td>0.139</td>
<td>0.174</td>
<td>0.165</td>
<td>0.217</td>
</tr>
<tr>
<td>ZSH [12]</td>
<td>64†</td>
<td>0.220</td>
<td>0.291</td>
<td>0.254</td>
<td>0.340</td>
</tr>
<tr>
<td>CAAE [13]</td>
<td>4096</td>
<td>-</td>
<td>-</td>
<td>0.196</td>
<td>0.284</td>
</tr>
<tr>
<td>CVAE [13]</td>
<td>4096</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.225</td>
</tr>
<tr>
<td>Doodle [14]</td>
<td>256</td>
<td>0.109</td>
<td>-</td>
<td>-</td>
<td>0.460</td>
</tr>
<tr>
<td>SEM-PCYC [15]</td>
<td>64</td>
<td>0.297</td>
<td>0.426</td>
<td>0.349</td>
<td>0.463</td>
</tr>
<tr>
<td>SAK [20]</td>
<td>512</td>
<td>0.293</td>
<td>0.392</td>
<td>0.344</td>
<td>0.399</td>
</tr>
<tr>
<td>SAK [20]</td>
<td>64†</td>
<td>0.359</td>
<td>0.481</td>
<td>0.364</td>
<td>0.487</td>
</tr>
<tr>
<td>SAK [20]</td>
<td>64†</td>
<td>0.381</td>
<td>0.506</td>
<td>0.488</td>
<td>0.644</td>
</tr>
</tbody>
</table>

TABLE 3: Performance comparison of using different backbones in TCN with 64 dimensional features. Here, Attention denotes spatial attention module [10], which aims to focus on fine-grained details.

<table>
<thead>
<tr>
<th>Backbone</th>
<th>Attention</th>
<th>Sketchy Split1</th>
<th>TU-Berlin</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG16</td>
<td>√</td>
<td>0.408</td>
<td>0.524</td>
</tr>
<tr>
<td>VGG16</td>
<td></td>
<td>0.414</td>
<td>0.523</td>
</tr>
<tr>
<td>VGG19</td>
<td></td>
<td>0.435</td>
<td>0.543</td>
</tr>
<tr>
<td>VGG19</td>
<td>√</td>
<td>0.442</td>
<td>0.549</td>
</tr>
<tr>
<td>ResNet-50</td>
<td></td>
<td>0.582</td>
<td>0.711</td>
</tr>
<tr>
<td>SE-ResNet-50</td>
<td></td>
<td>0.499</td>
<td>0.627</td>
</tr>
<tr>
<td>CSE-ResNet-50</td>
<td></td>
<td>0.509</td>
<td>0.625</td>
</tr>
</tbody>
</table>

TABLE 4: Ablation studies of individual component in TCN with 64 dimensional features. 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.

<table>
<thead>
<tr>
<th>Component</th>
<th>Variant</th>
<th>Sketchy Split1</th>
<th>TU-Berlin</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>mAP@all</td>
<td>Prec@100</td>
</tr>
<tr>
<td>w/o</td>
<td></td>
<td>0.368</td>
<td>0.449</td>
</tr>
<tr>
<td>Weight-Sharing</td>
<td>Hard (0 : 1)</td>
<td>0.403</td>
<td>0.508</td>
</tr>
<tr>
<td></td>
<td>Hard (1 : 1)</td>
<td>0.399</td>
<td>0.520</td>
</tr>
<tr>
<td></td>
<td>Hard (10 : 1)</td>
<td>0.367</td>
<td>0.488</td>
</tr>
<tr>
<td></td>
<td>Soft</td>
<td>0.582</td>
<td>0.711</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.557</td>
<td>0.691</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.567</td>
<td>0.702</td>
</tr>
<tr>
<td>Semantic Metric</td>
<td>BHT</td>
<td>0.571</td>
<td>0.703</td>
</tr>
<tr>
<td></td>
<td>BHT+SR (1 : 1)</td>
<td>0.570</td>
<td>0.702</td>
</tr>
<tr>
<td></td>
<td>BHT+SR (10 : 1)</td>
<td>0.569</td>
<td>0.700</td>
</tr>
<tr>
<td></td>
<td>SBHT</td>
<td>0.582</td>
<td>0.711</td>
</tr>
<tr>
<td>Teacher Network</td>
<td>w/o</td>
<td>0.480</td>
<td>0.649</td>
</tr>
<tr>
<td></td>
<td>With</td>
<td>0.582</td>
<td>0.711</td>
</tr>
</tbody>
</table>

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.

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 Figure 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 Figure 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
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.
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. Firstly, 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. Secondly, 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.

**Weight-Sharing Strategy.** To verify the effectiveness of each component, we conduct detailed ablation studies in Table 4. Firstly, the performance is poor when adopting no weight-sharing strategies. We argue that the model is over-fitting.

![Fig. 8: Retrieval results of the random selected 8 categories on Sketchy dataset.](image)

**TABLE 5:** The performance improvement (relative increase) of the random selected 8 categories at testing stage with 64 dimensional features on Sketchy. WS stands for the comparison of soft weight-sharing vs. without weigh-sharing and SM represents the comparison of semantic metric vs. without semantic metric.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Cup</th>
<th>Swan</th>
<th>Harp</th>
<th>Squirrel</th>
<th>Snail</th>
<th>Ray</th>
<th>Pineapple</th>
<th>Volcano</th>
<th>Rifle</th>
<th>Scissors</th>
</tr>
</thead>
<tbody>
<tr>
<td>WS</td>
<td>32.9%</td>
<td>8.4%</td>
<td>189.3%</td>
<td>14.7%</td>
<td>87.8%</td>
<td>106.3%</td>
<td>48.3%</td>
<td>59.6%</td>
<td>23.7%</td>
<td>52.4%</td>
</tr>
<tr>
<td>SM</td>
<td>1.3%</td>
<td>11.5%</td>
<td>10.6%</td>
<td>-5.7%</td>
<td>-1.4%</td>
<td>-9.7%</td>
<td>0.4%</td>
<td>10.0%</td>
<td>3.6%</td>
<td>1.7%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Parrot</th>
<th>Windmill</th>
<th>Teddy Bear</th>
<th>Tree</th>
<th>Wine Bottle</th>
<th>Deer</th>
<th>Chicken</th>
<th>Airplane</th>
<th>Wheelchair</th>
<th>Tank</th>
</tr>
</thead>
<tbody>
<tr>
<td>WS</td>
<td>29.7%</td>
<td>84.8%</td>
<td>279.3%</td>
<td>141.7%</td>
<td>316.5%</td>
<td>36.2%</td>
<td>77.9%</td>
<td>34.3%</td>
<td>30.1%</td>
<td>50.7%</td>
</tr>
<tr>
<td>SM</td>
<td>14.9%</td>
<td>-0.7%</td>
<td>4.7%</td>
<td>9.5%</td>
<td>7.2%</td>
<td>1.4%</td>
<td>17.5%</td>
<td>11.1%</td>
<td>8.7%</td>
<td>8.5%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Umbrella</th>
<th>Butterfly</th>
<th>Camel</th>
<th>Horse</th>
<th>Bell</th>
</tr>
</thead>
<tbody>
<tr>
<td>WS</td>
<td>264.1%</td>
<td>-5.0%</td>
<td>68.2%</td>
<td>64.0%</td>
<td>239.7%</td>
</tr>
<tr>
<td>SM</td>
<td>-7.8%</td>
<td>8.4%</td>
<td>-1.4%</td>
<td>10.9%</td>
<td>2.1%</td>
</tr>
</tbody>
</table>
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 the analysis of singular values of each modality in Figure 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 Figure 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 Figure 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.

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 ensure the transferability 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. Fur-
thermore, 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 map can be observed in Figure 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 noting 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 Figure 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 Figure 11(a). The Euclidean distance between semantic anchors of adjacent iteration is very small (i.e., less than 0.0035 in Figure 11(b)), which indicates that the semantic anchor changes slightly during training stage.

**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.

TABLE 6: Classification accuracy of digit images. Here, S and T stand for source and target domain, respectively. And the * denotes it is implemented by the public code.

<table>
<thead>
<tr>
<th>Method</th>
<th>MNIST (S) → MNIST-M (T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source Only</td>
<td>52.2%</td>
</tr>
<tr>
<td>DANN*</td>
<td>92.4%</td>
</tr>
<tr>
<td>DANN*+TCN</td>
<td>93.9%</td>
</tr>
</tbody>
</table>

**Length of Hashing Code.** In Figure 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 Figure 13, we analyze the effect of hyper-parameters $\lambda_{SWS}$. For simplicity, we fix the coefficients $\lambda_{DIS}$ and $\lambda_{SEM}$, 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 $\lambda_{SWS}$ and reaches the peak value at 1,000.

**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 BSD500 dataset [33]. 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.
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 transferability.

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.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 62132016, Grant 62171343, and Grant 62071361, in part by the Fundamental Research Funds for the Central Universities ZDRC2102, and in part by the Australian Research Council Projects DE-190101473 and IC-190100031.

REFERENCES


This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3123315, IEEE Transactions on Pattern Analysis and Machine Intelligence