Norm-guided Adaptive Visual Embedding for Zero-Shot Sketch-Based Image Retrieval

Wenjie Wang[1] , **Yufeng Shi**[1] , **Shiming Chen**[1] , **Qinmu Peng**[1] , **Feng Zheng**[3] and **Xinge You**[1,2*]
[1]School of Electronic Information and Communication, Huazhong University of Science and Technology
[2]Shenzhen Research Institute of Huazhong University of Science and Technology
[3]Department of Computer Science and Engineering, Southern University of Science and Technology
{wangwj5,yufengshi17, youxg}@hust.edu.cn

Abstract

Zero-shot sketch-based image retrieval (ZS-SBIR), which aims to retrieve photos with sketches under the zero-shot scenario, has shown extraordinary talents in real-world applications. Most existing methods leverage language models to generate class-prototypes and use them to arrange the locations of all categories in the common space for photos and sketches. Although great progress has been made, few of them consider whether such pre-defined prototypes are necessary for ZS-SBIR, where locations of unseen class samples in the embedding space are actually determined by visual appearance and a visual embedding actually performs better. To this end, we propose a novel Norm-guided Adaptive Visual Embedding (NAVE) model, for adaptively building the common space based on visual similarity instead of language-based pre-defined prototypes. To further enhance the representation quality of unseen classes for both photo and sketch modality, modality norm discrepancy and noisy label regularizer are jointly employed to measure and repair the modality bias of the learned common embedding. Experiments on two challenging datasets demonstrate the superiority of our NAVE over state-of-the-art competitors.

1 Introduction

Sketch-based image retrieval (SBIR), which conveniently allows users to search desired photos with free-hand sketches, has a broader prospect with the explosive growth of mobile internet and touch screens. Since it is hard to guarantee that the training set can cover all query categories at the application stage, a more realistic setting termed zero-shot sketch-based image retrieval (ZS-SBIR) has emerged. ZS-SBIR[Shen *et al.*, 2018], which combines zero-shot learning and SBIR, aims to retrieve photos with the query sketches whose categories are no longer limited to the classes that have shown in the training set. ZS-SBIR not only needs to narrow the domain gap between modalities, but needs to facilitate the knowledge transfer from seen categories to unseen ones.

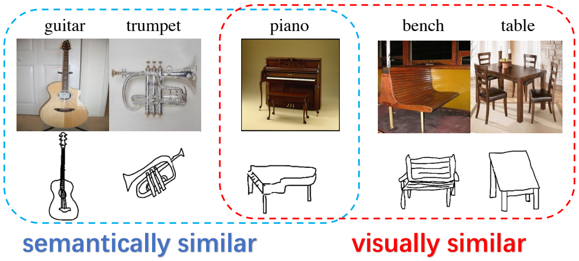

Figure 1: An illustration of the conflict between semantic similarity and visual similarity. ‘Piano’ is semantically similar to ‘guitar’ and ‘trumpet’, since it has a high co-occurrence frequency with the instruments in the corpus. However, it looks more like furniture, which is close to a ‘bench’ or ‘table’ from the visual perspective on both sketch and photo modalities.

Most previous ZS-SBIR approaches intend to learn a projection that maps sketches/photos to a latent embedding space, where sketches and photos from the same category are mapped close to a pre-defined common class-prototype. Specifically, they first utilize pre-trained language models (e.g., Word2Vec [Mikolov *et al.*, 2013] or Ji-Cn [Jiang and Conrath, 1997]), to translate labels of seen classes into high-dimensional vectors and consider them as the class-prototypes in the common space. Then, they learn to project both sketches and photos close to their corresponding class-prototypes. Based on this paradigm, various architectures (e.g., GANs [Dutta and Biswas, 2019], graph [Shen *et al.*, 2018; Zhang *et al.*, 2020], and cycle reconstruction [Dutta and Akata, 2019; Deng *et al.*, 2020]) have been employed to estimate the projection. They expect that such learned mappings can leverage the side-information from language model to project unseen class sketches/photos close to their semantically similar seen classes, and thus the distances among the sketches and photos from the same category are minimized. However, the locations of unseen class samples in the latent space are actually determined by their visual similarity with seen classes rather than the semantic relationship, which is the result of knowledge transfer from seen classes [Hsu *et al.*, 2018; Han *et al.*, 2019]. And we argue that such pre-defined class-prototypes are actually not proper for ZS-SBIR task.

There is the conflict between semantic similarity and visual similarity. The semantic similarity from language models is

[*] Corresponding author

based on the word co-occurrence frequency in large corpus, and thus the semantically similar classes may not be similar in visual appearances (an example is given in Figure 1). Therefore, models that aim to fit the language-based similarity would inevitably neglect the inter-class visual similarity knowledge, leading to a visually confusing embedding projection and poor performance on unseen classes. Meanwhile, the visual representation space from a classification-based CNN is already a meaningful space [Chen *et al.*, 2018; Zeiler and Fergus, 2014], where the distance between visually similar classes (e.g., table and bench) is smaller than the distance between visually dissimilar ones (e.g., table and trumpet). And we experimentally find that directly using such visual embedding as the common space performs well on ZS-SBIR with much discriminative representations generated for both unseen sketches and photos. It inspires us to rethink whether language-based pre-defined class-prototypes are necessary for ZS-SBIR.

To this end, instead of using language-based side-information, we propose a novel method named Norm-guided Adaptive Visual Embedding (NAVE), which adaptively builds the common embedding based on the visual similarity of seen classes and ImageNet. Due to the domain gap, there is also the conflict of visual inter-class similarity between sketch and photo modality. Without proper guidance, the common embedding may be biased and only meaningful for one modality, negatively affecting the knowledge transfer. We thus introduce the modality-mean-feature-norm discrepancy and a noisy label regularizer to measure and repair the modality bias of the learned embedding space. With their interaction, our NAVE can adaptively learn a balanced common embedding where the inter-class similarity is visually meaningful for both sketch and photo modalities, boosting the knowledge transferring for both modalities. Consequently, the final ZS-SBIR performance is improved.

The main contributions of this work can be summarized:

- We analyze the conflict of language-based similarity with visual similarity and its effect on unseen classes for ZS-SBIR task. To the best of our knowledge, this is the first work to consider whether pre-defined language-based prototypes are necessary for ZS-SBIR task.
- We propose the NAVE model for ZS-SBIR task that adaptively builds a common embedding space based on visual similarity instead of language-based pre-defined class-prototypes.
- To make the common embedding visually meaningful for both sketches and photos, we propose modality-mean-feature-norm discrepancy and noisy label regularizer to jointly measure and repair the modality bias of the learned embedding.
- Extensive experimental results on two popular benchmarks demonstrate that our NAVE outperforms the state-of-the-arts by a significant margin.

2 Related Work

Sketch-based Image Retrieval (SBIR). The main challenge in traditional SBIR is the domain gap between sketch and photo. Most existing SBIR methods intend to narrow the gap by learning a shared embedding space. Furthermore, multifarious tools (e.g., Siamese architecture [Qi *et al.*, 2016], pairwise loss [Liu *et al.*, 2017],) are adopted to obtain a better retrieval metric in the common space.

Zero-shot Learning (ZSL). ZSL aims to recognize objects from novel categories that are not shown in training set with additional side-information. Most ZSL approaches utilize visual attribute side-information (e.g., ”has wings”) to generate the class prototypes in the semantic space [Fu *et al.*, 2018]. And some early works [Socher *et al.*, 2013; Frome *et al.*, 2013] choose language model as an alternative. As a classification task, ZSL needs to assign samples with labels, where class-prototypes play the agents of labels in the semantic space. While ZS-SBIR, which is actually an open-set retrieval task, only aims to retrieve the visually similar photo with sketches. If the model can directly grasp the generic visual features of two modalities and generalize them to unseen classes, it is not a requisite for ZS-SBIR models to pre-define such semantic prototypes.

Zero-shot Sketch-based image retrieval (ZS-SBIR). Most existing ZS-SBIR methods follow the common space paradigm and pre-define class-prototypes with language models. Graph [Zhang *et al.*, 2020; Shen *et al.*, 2018], cycle consistency [Dutta and Akata, 2019; Deng *et al.*, 2020] and content-style disentanglement [Dutta and Biswas, 2019]) are employed to learn the projection to map sketches/photos close to such prototypes. [Dutta and Akata, 2019] proposes a selection layer to refine the prototypes and reduce the dimensionality of retrieval features. Graph-based method [Zhang *et al.*, 2020] adjust the language-based adjacency matrix with visual information. [Liu *et al.*, 2019] proposes a teacher-student framework to preserve discriminative representations from ImageNet and coordinates the representations close to language-based prototypes. [Dutta *et al.*, 2020] takes into account the class imbalance problem in ZS-SBIR. Although some refinements for the language prototypes have been utilized, few of them consider whether such prototypes are necessary for ZS-SBIR.

3 Language Prototypes in ZS-SBIR

To investigate whether pre-defined language-based prototypes are beneficial to knowledge transfer in ZS-SBIR task, we adopt the baseline illustrated in [Dutta *et al.*, 2020] and train it with various semantic class-prototypes extracted from different pre-trained language models. We follow the standard ZS-SBIR data partitioning on Sketchy Ext. [Liu *et al.*, 2017] / TU-Berlin Ext. [Zhang *et al.*, 2016] to randomly select 25/30 classes as *unseen* classes $\mathcal{C}^u$. And the rest 100/220 classes are considered as *seen* classes $\mathcal{C}^s$ for training.

The baseline model use two branches to extract feature representations of photos/sketches, namely $f^{(m)} = \mathcal{F}_m(x^{(m)};\theta_m)$, where $x^{(m)}$ is the sketch/photo input and $m \in \{sk, ph\}$. And pre-trained language models are utilized to vectorize the labels of *seen* classes $y \in \mathcal{C}^s$ as their corresponding class-prototypes $h(y)$. The distances among these prototypes in the common space preserve the knowledge from language models. Then, the model is trained to

Model Name	Sketchy Ext.				TU-Berlin Ext.			
	dim	SBIR	SBSR	PBPR	dim	SBIR	SBSR	PBPR
Fast [Joulin *et al.*, 2017]	300	0.323	0.413	0.436	300	0.302	0.418	0.478
Glove [Pennington *et al.*, 2014]		0.324	0.417	0.434		0.294	0.411	0.465
Word2Vector [Mikolov *et al.*, 2013]		0.318	0.411	0.433		0.305	0.415	0.462
Classifier		**0.452**	**0.578**	**0.657**		**0.406**	**0.581**	**0.631**
Ji-Cn [Jiang and Conrath, 1997]	354	0.297	0.388	0.452	664	0.295	0.387	0.406
Lch[Leacock and Chodorow, 1998]		0.285	0.362	0.431		0.274	0.397	0.451
Lin [Lin and others, 1998]		0.222	0.321	0.386		0.204	0.346	0.372
Path Length		0.301	0.394	0.466		0.257	0.419	0.414
Wup [Wu and Palmer, 1994]		0.258	0.355	0.408		0.205	0.345	0.339
Classifier		**0.459**	**0.574**	**0.677**		**0.411**	**0.574**	**0.644**

Table 1: mAP@all of SBIR, SBSR and PBIR on **unseen** set of models on Sketchy Ext. and TU-Berlin Ext. using different kinds of pre-defined class-prototypes from language models and the trainable classifier without any side-information.

map sketches/photos close to their corresponding prototypes with a distance-based cross-entropy loss:

$$\mathcal{L}(x^{(m)},\ y_i) = -log\frac{\exp(-d(f^{(m)}, h(y_i)))}{\sum_{j\in C^s}\exp(-d(f^{(m)}, h(y_j)))}, \quad (1)$$

where $d(f^{(m)}, h(y_i))$ is the squared Euclidean distance between the extracted features and its corresponding prototype. Thus, the model is expected to leverage language side-information to learn a projection, which can map unseen classes sketch/photos clustered around the language similar seen classes.

We use three text-based models (FastText, Word2Vec and Glove) and five hierarchy models (Ji-Cn, Lch, Lin, Path and Wup) to generate language prototypes. The text-based models translate seen labels into 300-dimensional vectors according to the word co-occurrence frequency in the corpus. The hierarchy models extract 354/664-dimensional prototypes for Sketchy/TU-Berlin based on WordNet. We also adopt a classifier to analyze whether language prototypes can benefit knowledge transferring to unseen classes. We use the classifier to train the branches with standard cross-entropy loss. Therefore, without the influence of the language side-information, the classifier model builds the inter-class similarity in the embedding space only based on visual similarity.

Table 1 presents mAP@all of sketch-based image retrieval (ZS-SBIR), sketch-based sketch retrieval (ZS-SBSR) and photo-based photo retrieval (ZS-PBPR) on **unseen** classes samples. We use ZS-SBSR and ZS-PBIR to evaluate the representation quality of unseen classes sketches/photos. As can be seen, different kinds of language prototypes indeed have different effects on the final result, which meets the conclusions in [Dutta and Akata, 2019]. Moreover, we observe that the classifier model outperforms all language-based ones on ZS-SBIR, ZS-SBSR and ZS-PBPR. It indicates that the classifier model can boost the knowledge transfer to generate more discriminative representations of both unseen sketches and photos. However, due to the conflict between visual similarity and semantic similarity, forcing the model to fit the language similarity inevitably makes the model to learn a visually confusing inter-class relationship. It negatively affects the knowledge transfer from seen to unseen classes and thus hampers the representation performance on unseen classes.

4 Proposed Approach

Based on the observation above, we propose Norm-guided Adaptive Visual Embedding (NAVE) for ZS-SBIR. As illustrated in Figure 2, without using any language side-information, our NAVE adaptively learns the common embedding space based on visual similarity from seen classes and ImageNet. A parameter-shared Siamese network is applied to map sketches and photos to the common embedding space. To acquire better representations of unseen photos, we follow previous [Liu *et al.*, 2019] using a teacher-student architecture to preserve the model's capability of recognizing rich photo features from ImageNet. To mediate the visual similarity conflict between sketch and photo modality, we introduce the modality-mean-feature-norm discrepancy and a noisy label regularizer to measure and repair of the learned embedding's modality bias. Therefore, our NAVE can build a more balanced embedding space, where the inter-class similarity is visually meaningful for both sketches and photos. It boosts the knowledge transfer for both modalities and improves the final ZS-SBIR performance.

Preliminaries. In ZS-SBIR setting, the whole dataset is divided into training (*seen*) and test (*unseen*) set according to the categories $C^{u/s}$. Let $\mathcal{S}_{tr} = \{(x^{sk}, y) \mid y \in \mathcal{C}^s\}$ represents the training sketch set and $\mathcal{P}_{tr} = \{(x^{ph}, y) \mid y \in \mathcal{C}^s\}$ be the training photo set, where $x^{sk/ph}$ and y are sketch/photo and label respectively. We denote $\mathcal{S}_{te} = \{(x^{sk}, y) | y \in \mathcal{C}^u\}$ as test sketch set and $\mathcal{P}_{te} = \{(x^{ph}, y) \mid y \in \mathcal{C}^u\}$ as test photo set respectively. To satisfy the zero-shot setting, the training classes and test classes do not share any category, $i.e.$, $\mathcal{C}^s \bigcap \mathcal{C}^u = \emptyset$. During testing, given a sketch query sample x^{sk} from $\mathcal{S}_{te}$, the ZS-SBIR model is expected to retrieve corresponding photos from the test photo set $\mathcal{P}_{te}$.

4.1 Visual Feature Extractor

We use the Siamese CNN pre-trained on ImageNet as the backbone to extract sketch and photo representations, $i.e.$, $f^{sk/ph} = \mathcal{F}(x^{sk/ph};\ \theta)$. Sharing parameters in Siamese

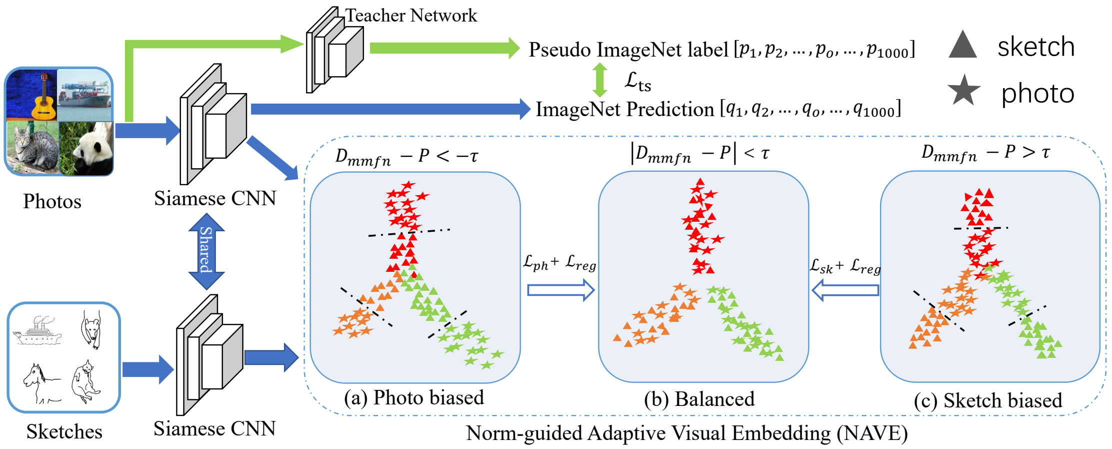
Teacher Network
Pseudo ImageNet label $[p_1, p_2, \ldots, p_o, \ldots, p_{1000}]$
$\mathcal{L}_{\text{ts}}$
ImageNet Prediction $[q_1, q_2, \ldots, q_o, \ldots, q_{1000}]$
sketch
photo
Photos
Siamese CNN
Shared
Sketches
Siamese CNN
$D_{mmfn} - P < -\tau$
$|D_{mmfn} - P| < \tau$
$D_{mmfn} - P > \tau$
$\mathcal{L}_{ph} + \mathcal{L}_{reg}$
$\mathcal{L}_{sk} + \mathcal{L}_{reg}$
(a) Photo biased
(b) Balanced
(c) Sketch biased
Norm-guided Adaptive Visual Embedding (NAVE)

Figure 2: Illustration of our proposed NAVE. A Siamese network is employed to extract features for sketches and photos, and a teacher network is adopted to maintain the rich photo features learned on ImageNet. For each mini-batch, we apply modality-mean-feature-norm discrepancy D_{mmfn} to measure the bias of current embedding. If the norm discrepancy exceeds the given range, e.g., (a) and (c), a noisy label regularizer $\mathcal{L}_{reg}$ is implemented to the larger-norm modality's objective until the discrepancy comes back into the given range (b), which constraints the learned embedding to be balanced and visually meaningful for both sketches and photos.

network can alleviate over-fitting, which has been proven in [Deng *et al.*, 2020]. After obtaining the visual representations of each modality, a fully-connected classification layer with the standard cross-entropy loss is adopted:

$$\mathcal{L}^{m}_{cls}(x^{(m)}, y_i) = -\log \frac{\exp(z^{(m)}_{y_i})}{\sum_{j\in\mathcal{C}^s} \exp(z^{(m)}_j)}, \tag{2}$$

where $m \in \{sk, ph\}$, $z = W^\top f^{(m)} + b$, W and b are the weight and bias of the classification layer.

Meanwhile, we adopt a teacher network that is pre-trained on ImageNet and keep it fixed during the training phase. It can better preserve the model's capability learned on ImageNet to recognize rich photo features learned on ImageNet, and thus benefits the generalization to unseen classes photos. For each photo x^{ph}, the teacher network generates a pseudo soft ImageNet label $\mathbf{p}$ which reflects the probability of each classes $\mathcal{C}^o$ in ImageNet. Then, the teacher-student objective is adopted to encourage the backbone to make the same ImageNet prediction for x^{ph}:

$$\mathcal{L}_{ts}(x^{ph},\ \mathbf{p}) = -\sum_{m\in\mathcal{C}^o} p_m \log(q_m), \tag{3}$$

where $\mathbf{q} = Softmax(W_o^\top f^{ph} + b_o)$ is the softmax output of the ImageNet classification layer. Different from [Liu *et al.*, 2019] that fine-tunes the teacher signals close to the language side-information, our method views the ImageNet supervision as a prior and adaptively learn a balanced embedding that is visually meaningful both photos and sketches.

4.2 Norm-guided Modality Balance

Since sketches are made up of sparse shape strokes while photos contain both texture and shape cues, there is domain gap between two modalities. Therefore, the visual inter-class similarity of the two modalities is not the same and also not proper for the other modality to fit. The ImageNet-trained CNNs have shown to be strongly biased towards recognizing texture features compared with shapes [Geirhos *et al.*, 2018]. Therefore, the inter-class relationship from ImageNet is not visually meaningful for sketch modality. Meanwhile, the similarity of sketches is also not proper for photo models to fit, as it makes the photo model ignore the texture features, leading to poor recognition performance [Nam and Kim, 2019]. We aim to learn a 'sweet spot' embedding to mediate the visual similarity conflict between sketches and photos, and thus more knowledge for both modalities can be transferred to unseen classes. However, without proper guidance, it is hard to guarantee that the learned embedding is balanced for two modalities or biased.

Inspired from recent smaller-norm-less-informative assumption [Xu *et al.*, 2019], we adopt the features norms to measure the bias of the embedding. Specifically, we introduce the **m**odality-**m**ean-**f**eature-**n**orm discrepancy D_{mmfn} to measure the modality bias of the embedding learned during training:

$$D_{mmfn} = \left(\frac{1}{n^{ph}}\sum_{i=1}^{n^{ph}} \|f^{ph}_i\|^2_2 - \frac{1}{n^{sk}}\sum_{j=1}^{n^{sk}} \|f^{sk}_j\|^2_2\right), \tag{4}$$

where n^{sk} and n^{ph} is the batch size of sketch and photo modality. If the discrepancy D_{mmfn} is out of the given range:

$$|D_{mmfn} - P| > \tau, \tag{5}$$

the current embedding is considered as biased towards the larger-norm modality, where expected discrepancy P and threshold τ are hyper-parameters. Since the norm reflects the model's inference confidence on the samples, a biased embedding with large discrepancy is not beneficial for either the small norm modality or the large norm one. If the embedding is not suitable for one modality, the model will project the features with less confidence or small norm, decreasing the knowledge to unseen classes. Meanwhile, too large norm indicates that the model is over-confident to the seen classes, which is also harmful to generalize knowledge to unseen classes. Consequently, some measures should be taken to repair the biased embedding to be more balanced.

4.3 Noisy Label Regularizer

Noisy label is introduced as the regularizer, which can prevent the model from being over-confident [Xie *et al.*, 2016] without spoiling the inter-class similarity knowledge [Müller *et al.*, 2019]. When the norm discrepancy is out of the given range, we add the noisy label regularizer to larger-norm modality's objective. For each sample $(x^{(m)}, y_i)$ from the regularized modality, the noisy label is randomly drawn from a uniform distribution over all *seen* classes except the ground-truth c_i:

$$\begin{cases} \tilde{c} \sim \mathbf{U}(\{c_1, c_2, ..., c_s\} - \{c_i\}) \\ \tilde{y}_{\tilde{c}} = 1 \\ \tilde{y}_k = 0, \ \forall \, k \neq \tilde{c}. \end{cases} \tag{6}$$

Then the regularizer loss $\mathcal{L}_{reg}$ is computed with the generated noisy label $(x^{(m)}, \tilde{y})$:

$$\mathcal{L}^m_{reg} = \mathcal{L}^m_{cls}(x^{(m)}, \ \tilde{y}), \tag{7}$$

where $\mathcal{L}^m_{cls}$ is the standard cross-entropy loss (*i.e.*, Eq.2) and $m \in \{sk, ph\}$. Each regularized sample is therefore equipped with a ground-truth label and a disturbing one. For the regularized modality, the disturbing labels regularize the model to project features with less confidence and small norm. As a coin has two sides, the common space has to embed both modalities. The disturbing labels, meanwhile, allow the model to tune the embedding biased towards the other modality with larger norm features. Therefore, the embedding bias is adjusted.

4.4 Objective Function

The basic training objective of the two modalities are:

$$\begin{cases} \mathcal{L}_{sk} = \mathcal{L}^{sk}_{cls} \\ \mathcal{L}_{ph} = \mathcal{L}^{ph}_{cls} + \mathcal{L}_{ts}. \end{cases} \tag{8}$$

The regularizer $\mathcal{L}_{reg}$ is adaptively added to one modality's objective under the guidance of D_{mmfn}. For example, when $D_{mmfn} - P < -\tau$, the current model is biased towards the larger norm sketch modality and the regularizer is added:

$$\mathcal{L}_{sk} = \mathcal{L}_{sk} + \lambda_{reg}\mathcal{L}_{reg}, \tag{9}$$

where λ_{reg} is a hyper-parameter. More operation details in the whole training procedure are presented in **Algorithm**1.

5 Experiments

5.1 Datasets and Settings

Datasets and Setup. We validate our NAVE on two widely-used benchmarks: Sketchy Ext. [Liu *et al.*, 2017] and TU-Berlin Ext. [Zhang *et al.*, 2016]. Sketchy Ext. consists of 75,479 sketches and 73,002 photos from 125 categories. TU-Berlin Ext. contains 20,000 sketches evenly distributed over 250 categories and additional 204,489 photos. Following the data partitioning in [Liu *et al.*, 2019], we randomly pick 25 classes from Sketchy and 30 classes from TU-Berlin as test set, and regard the rest 100/220 classes as the training set.

Algorithm 1 NAVE algorithm

Input: training set $\{\mathcal{S}_{tr}, \mathcal{P}_{tr}\}$, hyper-parameter of regularizer λ_{reg}, expected discrepancy P and threshold τ.
Parameter: $\mathcal{F}(x^m; \theta)$, classification layer (W, b), ImageNet layer (W_o, b_o).

```
for each iteration do
    Sample sketch batches $\{(x^{sk}, y)\}_{i=1}^{n_{sk}}$ from $\mathcal{S}^{tr}$ and photo batches $\{(x^{ph}, y)\}_{i=1}^{n_{ph}}$ from $\mathcal{P}_{tr}$
    for each mini-batch do
        $\mathcal{L}_{sk} = \mathcal{L}^{sk}_{cls}, \ \mathcal{L}_{ph} = \mathcal{L}^{ph}_{cls} + \mathcal{L}_{ts}$
        Calculate $D_{mmfn}$ according to Eq.(4)
        if $(D_{mmfn} - P < -\tau)$ then
            $\mathcal{L}_{sk} = \mathcal{L}_{sk} + \lambda_{reg}\mathcal{L}_{reg}$;
        else if $(D_{mmfn} - P > \tau)$ then
            $\mathcal{L}_{ph} = \mathcal{L}_{ph} + \lambda_{reg}\mathcal{L}_{reg}$;
        end if
        Update parameters with $\nabla\mathcal{L}_{sk}$ and $\nabla\mathcal{L}_{ph}$;
    end for
end for
return feature extractor $\mathcal{F}(x^m \ ; \theta)$.
```

Implementation Details. We implement our NAVE with PyTorch and train it on two Nvidia 1080 Ti GPUs. We use Adam optimizer for training with an initial learning rate $lr = 0.0001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a 0.5 learning rate decay per epoch. The best hyper-parameters are $\lambda_{reg} = 0.2$, $P = 5$, $\tau = 3$ for Sketchy Ext., and $\lambda_{reg} = 0.1$, $P = 3$, $\tau = 3$ for TU-Berlin Ext. We follow the previous work [Liu *et al.*, 2019] to take SE-ResNet50 pre-trained on ImageNet as the feature extractor backbone and the same teacher network for a fair comparison.

5.2 Comparison with Existing Methods

To verify the superiority of our proposed NAVE, we make the comparison with six recently published ZS-SBIR methods. All methods use ImageNet pre-trained network for weight initialization and utilize language model to extract side-information. Like most existing methods, we adopt mean average precision (mAP@all) and precision considering top 100 (Prec@100) for performance evaluation.

As shown in Table 2, our model outperforms all the language-model equipped competitors, which shows the effectiveness of our visual embedding model. Among the competitors, SAKE [Liu *et al.*, 2019] is second only to our NAVE. It wins on the teacher-student architecture that preserves the ImageNet knowledge for photo modality, which also partially accounts for our adoption of the architecture. However, SAKE still uses additional constraints to calibrate the teacher signals close to the language-based side-information, which results in a visually confusing space. Instead, our NAVE adaptively builds the common space based on visual similarity of sketches and photos. It boosts the knowledge transfer for both modalities and thus improves the final ZS-SBIR performance of our NAVE.

Method	Dimension	Sketchy Ext.		TU-Berlin Ext.	
		mAP@all	Prec@100	mAP@all	Prec@100
ZSIH[Shen *et al.*, 2018]	64	0.258	0.342	0.220	0.291
SEM-PCYC[Dutta and Akata, 2019]	64	0.349	0.463	0.297	0.426
Style-guide[Dutta and Biswas, 2019]	64	0.376	0.484	0.254	0.355
NAVE (Ours)	64	**0.508**	**0.632**	**0.412**	**0.519**
SketchGCN[Zhang *et al.*, 2020]	1024	0.382	0.538	0.324	0.505
SAKE[Liu *et al.*, 2019]	512	0.547	0.692	0.475	0.599
AMDReg[Dutta *et al.*, 2020]	512	0.551	0.715	0.447	0.574
NAVE (Ours)	512	**0.613**	**0.725**	**0.493**	**0.607**

Table 2: Performance comparison (mAP@all and Prec@100) of the propose method with state-of-the-art ZS-SBIR methods on Sketchy Ext. and TU-Berlin Ext. datasets.

Dataset	Expected discrepancy P						
	-25	-15	0	5	10	20	30
Sketchy	0.542	0.563	0.601	0.610	0.603	0.593	0.588
TU-Berlin	0.458	0.475	0.490	0.488	0.485	0.480	0.477

Table 3: mAP@all on Sketchy Ext. and TU-Berlin Ext. with different expected discrepancy P.

#	Description	Sketchy	TU-Berlin
(a)	Baseline($\mathcal{L}_{cls}$)	0.475	0.417
(b)	Baseline($\mathcal{L}_{cls}$) + $\mathcal{L}_{ts}$	0.535	0.467
(c)	Baseline($\mathcal{L}_{cls}$) + $\mathcal{L}_{ts}$ + $\mathcal{L}_{reg}$	0.587	0.473
(d)	NAVE (full model)	**0.613**	**0.493**

Table 4: Ablation Studies on our NAVE mAP@all results of several baselines are shown above.

5.3 Effect of Norm Guidance

To illustrate the relationship between modality norm discrepancy D_{mmfn} and the final retrieval performance, we train our NAVE with different expected discrepancy P on 512-dimensional features. All models are trained under the same data partitioning setting and $\tau = 3$. Table 3 shows ZS-SBIR mAP@all for models with different expected discrepancy P. As shown in Table 3, when the discrepancy is too small, the embedding is biased towards sketch modality and the retrieval accuracy greatly drops. When the embedding is biased towards photos with a large discrepancy, the accuracy drops slightly. We suspect that since photos with complex textures are much harder to recognize than sketches, slightly biased towards photo modality would push the model to pay more attention on photos while has a relatively negligible impact on sketch modality. Hence the photo modality matters more to the final result. The best performance is obtained when P is within $(0, 10)$ for both Sketchy Ext. and TU-Berlin Ext. It indicates that a relatively balanced visual embedding with small discrepancy will improve the final ZS-SBIR performance.

5.4 Ablation Study

In Table 4, we conduct ablation study to demonstrate the effectiveness of each component in our NAVE. Four variants are designed including: (a) Train the model only with the classification loss $\mathcal{L}_{cls}$ (i.e., baseline); (b) add the teacher-student loss $\mathcal{L}_{ts}$ to photo modality's objective; (c) add the noisy label regularizer $\mathcal{L}_{reg}$ to both modalities during the whole training phase without the norm discrepancy guidance; (d) our full NAVE which adaptively adds the noisy label regularizer under the guidance of norm discrepancy. All ablation experiments are conducted on 512-dimensional retrieval features. Table 4 reports their mAP@all results on Sketchy Ext. and TU-Berlin Ext. datasets.

Compared with (b) and (a), we can find that the teacher-student objective indeed improves the performance, since it can maintain the capability of recognizing the rich photo features from ImageNet. In (c), we boldly add the noisy label regularizer to the objective without the guidance of norm discrepancy. As shown in Table 4, the noisy label can improve the baseline's performance, as it regularizes the model from being over-confident to the seen classes. However, due to the visual conflict between sketches and photos, the learn embedding without proper guidance is probably biased and limits the knowledge transfer to unseen classes of both modalities. Finally, among all variants, our full NAVE achieves the highest mAP@all on both Sketchy and TU-Berlin. It proves that with the norm discrepancy guidance, a more balanced embedding is obtained and benefits to generalize more useful knowledge for both modalities and the final performance.

6 Conclusion

In this work, we first analyzed the effect of language-based prototypes to ZS-SBIR and observed that such prototypes actually hamper the knowledge transfer for inferring the unseen categories. To this end, we proposed the Norm-guided Adaptive Visual Embedding (NAVE) method, which builds the common embedding based on visual similarity rather than language-based pre-defined prototypes. A norm discrepancy guidance and a noisy label regularizer were introduced to measure and repair the embedding bias. Experiments on two datasets verified our proposed model outperforms existing ZS-SBIR methods without any language side-information.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61772220), and Science, Technology and Innovation Commission of Shenzhen Municipality (No. JCYJ20180305180637611, JCYJ20180305180804836 and JSGG20180507182030600).

References

[Chen *et al.*, 2018] Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *CVPR*, pages 1043–1052, 2018.

[Deng *et al.*, 2020] Cheng Deng, Xinxun Xu, Muli Yang, and Dacheng Tao. Progressive cross-modal semantic network for zero-shot sketch-based image retrieval. *IEEE Transactions on Image Processing*, 29:8892–8902, 2020.

[Dutta and Akata, 2019] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*, pages 5089–5098, 2019.

[Dutta and Biswas, 2019] Titir Dutta and Soma Biswas. Style-guided zero-shot sketch-based image retrieval. In *BMVC*, page 209, 2019.

[Dutta *et al.*, 2020] Titir Dutta, Anurag Singh, and Soma Biswas. Adaptive margin diversity regularizer for handling data imbalance in zero-shot sbir. In *ECCV*, pages 349–364. Springer, 2020.

[Frome *et al.*, 2013] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, pages 2121–2129, 2013.

[Fu *et al.*, 2018] Yanwei Fu, Tao Xiang, Yu-Gang Jiang, Xiangyang Xue, Leonid Sigal, and Shaogang Gong. Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content. *IEEE Signal Processing Magazine*, 35(1):112–125, 2018.

[Geirhos *et al.*, 2018] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2018.

[Han *et al.*, 2019] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*, pages 8401–8409, 2019.

[Hsu *et al.*, 2018] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *ICLR*, 2018.

[Jiang and Conrath, 1997] Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.

[Joulin *et al.*, 2017] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext. zip: Compressing text classification models. In *ICLR*, 2017.

[Leacock and Chodorow, 1998] C Leacock and M Chodorow. Combining local context and wordnet sense similarity for word sense identification. wordnet, an electronic lexical database. *The MIT Press*, 1998.

[Lin and others, 1998] Dekang Lin et al. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304, 1998.

[Liu *et al.*, 2017] Li Liu, Fumin Shen, Yuming Shen, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, pages 2862–2871, 2017.

[Liu *et al.*, 2019] Qing Liu, Lingxi Xie, Huiyu Wang, and Alan L Yuille. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *ICCV*, pages 3662–3671, 2019.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, pages 3111–3119, 2013.

[Müller *et al.*, 2019] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *NeurIPS*, pages 4694–4703, 2019.

[Nam and Kim, 2019] Hyeonseob Nam and Hyo-Eun Kim. Batch-instance normalization for adaptively style-invariant neural networks. In *NeurIPS*, pages 4694–4703, 2019.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[Qi *et al.*, 2016] Yonggang Qi, Yi-Zhe Song, Honggang Zhang, and Jun Liu. Sketch-based image retrieval via siamese convolutional neural network. In *ICIP*, pages 2460–2464, 2016.

[Shen *et al.*, 2018] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *CVPR*, pages 3598–3607, 2018.

[Socher *et al.*, 2013] Richard Socher, Milind Ganjoo, and Andrew Ng. Zero-shot learning through cross-modal transfer. *NeurIPS*, 26:935–943, 2013.

[Wu and Palmer, 1994] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *ACL*, pages 133–138, 1994.

[Xie *et al.*, 2016] Lingxi Xie, Jingdong Wang, Zhen Wei, Meng Wang, and Qi Tian. Disturblabel: Regularizing cnn on the loss layer. In *CVPR*, pages 4753–4762, 2016.

[Xu *et al.*, 2019] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *ICCV*, pages 1426–1435, 2019.

[Zeiler and Fergus, 2014] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833, 2014.

[Zhang *et al.*, 2016] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. Sketchnet: Sketch classification with web images. In *CVPR*, pages 1105–1113, 2016.

[Zhang *et al.*, 2020] Zhaolong Zhang, Yuejie Zhang, Rui Feng, Tao Zhang, and Weiguo Fan. Zero-shot sketch-based image retrieval via graph convolution network. In *AAAI*, pages 12943–12950, 2020.