Progressive Domain-Independent Feature Decomposition Network for Zero-Shot Sketch-Based Image Retrieval

Xinxun Xu , **Muli Yang** , **Yanhua Yang**[*] and **Hao Wang**
Xidian University
{xinxun.xu,muliyang.xd,haowang.xidian}@gmail.com, yanhyang@xidian.edu.cn

Abstract

Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR) is a specific cross-modal retrieval task for searching natural images given free-hand sketches under the zero-shot scenario. Most existing methods solve this problem by simultaneously projecting visual features and semantic supervision into a low-dimensional common space for efficient retrieval. However, such low-dimensional projection destroys the completeness of semantic knowledge in original semantic space, so that it is unable to transfer useful knowledge well when learning semantic features from different modalities. Moreover, the domain information and semantic information are entangled in visual features, which is not conducive for cross-modal matching since it will hinder the reduction of domain gap between sketch and image. In this paper, we propose a Progressive Domain-independent Feature Decomposition (PDFD) network for ZS-SBIR. Specifically, with the supervision of original semantic knowledge, PDFD decomposes visual features into domain features and semantic ones, and then the semantic features are projected into common space as retrieval features for ZS-SBIR. The progressive projection strategy maintains strong semantic supervision. Besides, to guarantee the retrieval features to capture clean and complete semantic information, the cross-reconstruction loss is introduced to encourage that any combinations of retrieval features and domain features can reconstruct the visual features. Extensive experiments demonstrate the superiority of our PDFD over state-of-the-art competitors.

1 Introduction

With the explosive growth of image contents in our real world, image retrieval has been playing an important role in many fields, such as e-commerce, medical diagnosis and remote sensing. Conventional image retrieval requires providing textual descriptions, which are difficult to be obtained in many real-world cases. On mobile devices, image retrieval with free-hand sketches, for delivering targeted candidates visually and concisely, has attracted increasing attention and formed the area of Sketch-Based Image Retrieval (SBIR).

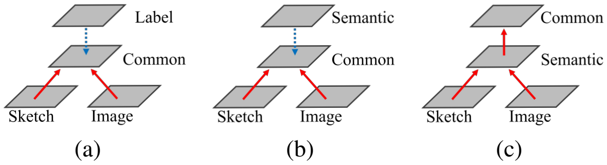


Figure 1: The ways in Figure 1(a) and Figure 1(b) simultaneously project visual features and label/semantic supervision into a low-dimensional common space for efficient retrieval. The Figure 1(c) shows our way that first aligns sketch and image to semantic embedding explicitly and then projects them into common space.

Since it is difficult to guarantee that all categories can be trained in realistic scenarios, unsatisfactory performance is often yielded when testing on unseen categories. In view of this, a more realistic setting is emerged, namely ZS-SBIR, which combines Zero-Shot Learning (ZSL) and SBIR for real-world applications. The ZS-SBIR is a retrieval task under zero-shot scenario, and people can quickly draw sketches on the mobile for retrieval, so it attracted increasing attention in real world. However, ZS-SBIR is extremely challenging since it simultaneously needs to deal with cross-modal matching, significant domain gap, as well as limited knowledge of unseen classes. The traditional SBIR methods [Liu *et al.*, 2017a; Zhang *et al.*, 2018] cannot directly address these problems effectively since they over-fit the source domain and meanwhile neglect the unseen categories. On the contrary, traditional ZSL [Kodirov *et al.*, 2017; Zhang and Saligrama, 2016] methods often focus on solving single-modal problems. Therefore, ZS-SBIR tries to solve these problems sufficiently by combining the advantages of ZSL and SBIR.

As shown in Figure 1(a) and Figure 1(b), previous work attempted to overcome these challenges through simultaneously projecting sketch/image features and label/semantic supervision to a low-dimensional common space. However, these strategies deteriorate the original semantic knowledge, because the low-dimensional projection maps complete semantic embedding from original semantic space to semanti-

[*] Contact Author

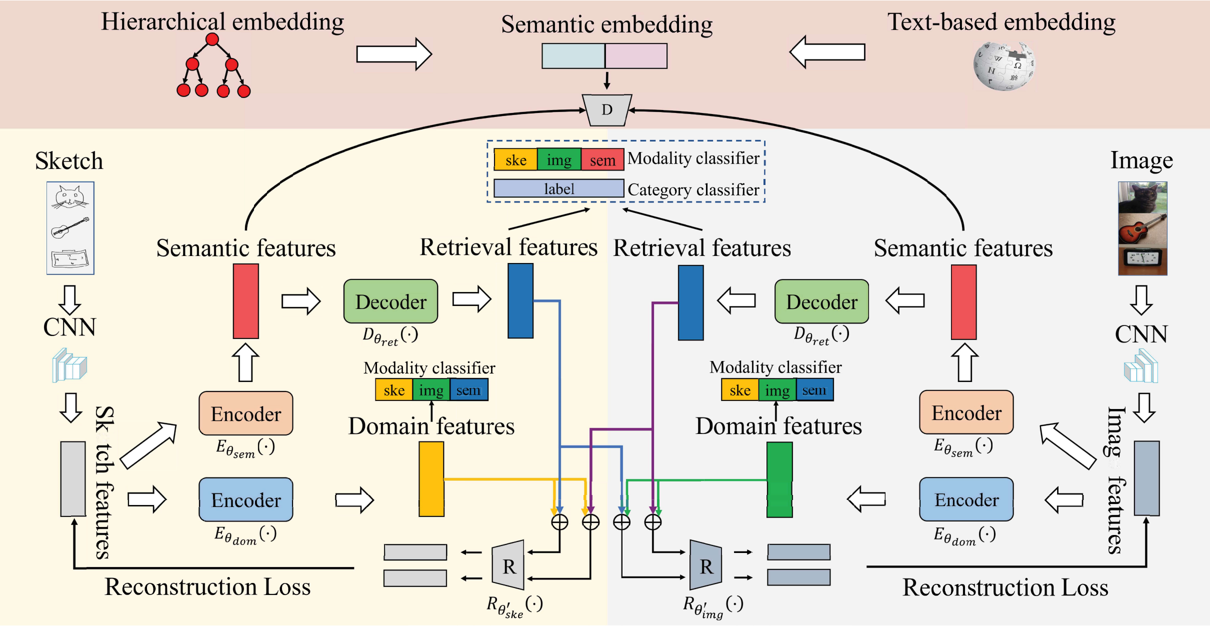


Figure 2: Flowchart of our proposed PDFD. First, it decomposes the visual features into semantic features and domain features, where the semantic features are learned in an adversarial fashion, while the domain features are learned under the constraint of modality classifier. Subsequently, the semantic features are projected into a common space as retrieval features. Moreover, the semantic embedding derived from the text-based embedding and hierarchical embedding serve as true examples to the discriminator. Meanwhile, the cross-reconstruction loss guarantees that the retrieval features only contain high-level knowledge, which is beneficial to reducing the interference of domain features.

cally incomplete low-dimensional space. Hence, as shown in Figure 1(c), we present progressive projection strategy that first learns semantic features with original semantic supervision, and then projects them into a common retrieval space, which is beneficial to knowledge transfer due to the strong semantic supervision can be maintained. Another issue is that, the domain information and semantic information are entangled in visual features, and the distribution of two different domains are highly heterogeneous, which seriously hinders the reduction of domain gap between sketch and image, making cross-modal matching difficult. Since semantic information remains the same expression in different domains, we argue that only semantic information is crucial for cross-modal matching. To this end, we decompose the visual features to attain domain-independent retrieval features that only contain semantic information.

In this paper, we have proposed a Progressive Domain-Independent Feature Decomposition (PDFD) network for ZS-SBIR task. First, PDFD decomposes visual features into semantic features and domain features, where the semantic features are adversarially learned with the supervision of original semantic embedding, while the domain features are learned with a modality classifier. Subsequently, the learned semantic features are projected into a common space as retrieval features under the category and modality supervision. Besides, in order to reduce the domain interference for cross-modal matching, we introduce cross-reconstruction loss to encourage the retrieval features capture clean and complete semantic information. It is expected that such retrieval features can reconstruct the sketch or image visual features by combining with sketch-domain features or image-domain features. In this network, the parameters of decoders and encoders for sketches and images are shared.

The main contributions of this work are summarized:

- We propose a feature decomposition model to effectively reduce the domain gap by generating domain-independent retrieval features with a novel cross-reconstruction loss.
- The proposed progressive projection strategy preserves the strong semantic supervision when generating retrieval features, which is beneficial to knowledge transfer under the zero-shot scenario.
- Extensive experiments conducted on two popular large-scale datasets demonstrate that our proposed PDFD significantly outperforms state-of-the-art methods.

2 Related Work

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

2.1 Sketch-Based Image Retrieval

The existing SBIR approaches can be mainly divided into two categories: hand-crafted features based methods and deep learning based ones. The hand-crafted features based methods attempt to bridge the domain gap by using edge-maps extracted from images, such as gradient field HOG descriptor [Hu and Collomosse, 2013] and Learned Key Shapes (LKS) [Saavedra *et al.*, 2015]. As for the deep learning based methods, Yu *et al.* [2017] first adopted CNN to learn better feature representation for sketches and images. Besides, siamese architecture [Qi *et al.*, 2016] achieves a better metric of retrieval by minimizing the loss function for samples from the same category and maximizing the loss function for samples from different categories.

2.2 Zero-Shot Learning

Existing zero-shot approaches can be classified into two categories: embedding-based and generative-based approaches. In the first category, some approaches learn non-linear multimodal embedding [Akata *et al.*, 2015; Xian *et al.*, 2016]. As for generative-based approaches, Wei *et al.* [2019] utilize a generator to synthesize unseen features and learn the semantic relations in a common space. It's worth noting that, most of ZSL methods utilize auxiliary information, *e.g.*, a text-based embedding [Mikolov *et al.*, 2013] or a hierarchical embedding [Miller, 1995] for label embedding, which is beneficial to knowledge transfer.

2.3 Zero-Shot Sketch-Based Image Retrieval

The first work [Shen *et al.*, 2018] of ZS-SBIR utilizes a multi-modal learning network to mitigate heterogeneity between two different modalities. The recent work SEM-PCYC [Dutta and Akata, 2019] proposes a paired cycle-consistent generative model based on semantically alignment, which maintains a cycle consistency that only requires supervision at category level. Besides, SAKE [Liu *et al.*, 2019; Kodirov *et al.*, 2017] proposes a teacher-student network to maximally preserving previously acquired knowledge to reduce the domain gap between the seen source domain and unseen target domain.

3 Methodology

3.1 Problem Definition

We first provide a formal definition of the ZS-SBIR task. Let $\mathcal{D}_{tr} = \{\mathcal{X}^{seen}, \mathcal{S}^{seen}, \mathcal{Y}^{seen}\}$ be a training set that contains sketch-image pairs $\mathcal{X}^{seen} = \{x_i^{ske}, x_i^{img}\}_{i=1}^{N_s}$, semantic embedding $\mathcal{S}^{seen} = \{s_i^{seen}\}_{i=1}^{N_s}$, and category labels $\mathcal{Y}^{seen} = \{y_i^{seen}\}_{i=1}^{N_s}$ with N_s samples. The test set is denoted as $\mathcal{D}_{te} = \{\mathcal{X}^{un}, \mathcal{Y}^{un}\}$ that contains sketches $\mathcal{X}^{un} = \{x_i^{un}\}_{i=1}^{N_u}$ and category labels $\mathcal{Y}^{un} = \{y_i^{un}\}_{i=1}^{N_u}$ with N_u samples, which satisfies the zero-shot setting $\mathcal{Y}^{seen} \cap \mathcal{Y}^{un} = \varnothing$. During the test, given an unseen sketch x_i^{un} in $\mathcal{X}^{un}$, the objective is to retrieve corresponding natural images from the image gallery which contain all images including training images.

The architecture of our proposed PDFD is illustrated in Figure 2, which contains two branches for sketch and image, respectively. Each branch first decomposes visual features into domain features and semantic features. Subsequently, decoders with shared parameters are trained to project semantic features into retrieval features for subsequent ZS-SBIR task.

3.2 Visual Features Decomposition

In zero-shot learning, it is important to provide knowledge supervision when learning semantic features. Our proposed PDFD utilizes text-based embedding and hierarchical embedding to provide such supervision.

Semantic Knowledge Embedding

In PDFD, we adopt two widely-used text-based embedding, *i.e.*, Word2Vec [Mikolov *et al.*, 2013] and GloVe [Pennington *et al.*, 2014] to obtain text representations. As for hierarchical embedding in PDFD, the similarity between words is calculated in WordNet[1] with path similarity and Jiang-Conrath [Jiang and Conrath, 1997] similarity.

Semantic Features

As illustrated in Figure 2, each branch has a semantic encoder $E_{\theta_{sem}}$, common discriminator $D_{\theta_{dis}}$, and semantic embedding that combine text-based embedding and hierarchical embedding. Given a training sketch-image pair, their visual features are extracted from VGG16 [Simonyan and Zisserman, 2014] network pre-trained on ImageNet [Deng *et al.*, 2009] dataset (before the last pooling layer). Then the semantic features are learned in an adversarial fashion, which means that the learned semantic features are expected to be as similar as the semantic embedding by 'fooling' the discriminator $D_{\theta_{dis}}$. Specifically, the objective can be formulated as:

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

where x^{ske}, x^{img}, s^{seen}, $E_{\theta_{sem}}(\cdot)$ and $D_{\theta_{dis}}(\cdot)$ denote the sketch features, image features, semantic embedding, semantic generation function and discriminator function, respectively. Besides, the semantic generation function and discriminator function are parameterized by θ_{sem} and θ_{dis}. Here, $E_{\theta_{sem}}$ minimizes the objective against an opponent $D_{\theta_{dis}}$ that tries to maximize it.

Domain Features

Since semantic features and domain features are separated, we argue that semantic, image-domain features and sketch-domain features should be distinguished from each other. Thus, we categorize these three kinds of features into three different modalities. Here, the domain encoder $E_{\theta_{dom}}$ is adopted to attain the domain features with the constraint of the modality classifier. The modality classification loss can be formulated as:

$$\begin{aligned}\mathcal{L}_{dmcls} =& -\mathbb{E}[\log P(y^{ske}|x_{ske}^{dom})] \\ &- \mathbb{E}[\log P(y^{img}|x_{img}^{dom})],\end{aligned} \tag{2}$$

where y^{ske} and y^{img} are labels indicating whether the corresponding features belong to sketch and image. Moreover, x_{ske}^{dom}, and x_{img}^{dom} denote the domain features from sketch and image branch respectively. They can be formulated as:

$$x_{ske}^{dom} = E_{\theta_{dom}}(x^{ske}), \tag{3}$$

$$x_{img}^{dom} = E_{\theta_{dom}}(x^{img}). \tag{4}$$

It is worth noting that the domain features also have been constrained to cross-reconstruction loss, which will be introduced in Section 3.3.

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

3.3 Retrieval Features Generation

After learning semantic features and domain features, PDFD generates retrieval features under two kinds of constraints.

Classification Constraint

It should be noted that the semantic features learned from two branches are only constrained by adversarial loss, which only ensures that the semantic features possess semantic knowledge. However, it can not guarantee the features to be class-discriminative. Therefore, category classifier is introduced after the two branches. The category classification loss can be formulated as:

$$\mathcal{L}_{ccls} = -\mathbb{E}[\log P(y|x^{ret}_{ske})] - \mathbb{E}[\log P(y|x^{ret}_{img})], \tag{5}$$

where y is the category label of x^{ske} and x^{img}. Moreover, x^{sem}_{ske}, x^{sem}_{img}, x^{ret}_{ske} and x^{ret}_{img} denote the semantic features and retrieval features generated from sketch and image branch respectively. The generation of retrieval features can be formulated as:

$$x^{ret}_{ske} = D_{\theta_{ret}}(x^{sem}_{ske}), \tag{6}$$

$$x^{ret}_{img} = D_{\theta_{ret}}(x^{sem}_{img}), \tag{7}$$

where $D_{\theta_{ret}}(\cdot)$ is the generation function of retrieval features.

On the other hand, retrieval features should be domain-independent, such that they ought to be classified the semantic modality. We adopt the same modality classifier used in generation of domain features to ensure that, where the modality classification loss is written as:

$$\begin{aligned}\mathcal{L}_{rmcls} = &- \mathbb{E}[\log P(y^{sem}|x^{ret}_{ske})] \\ &- \mathbb{E}[\log P(y^{sem}|x^{ret}_{img})],\end{aligned} \tag{8}$$

where y^{sem} is the label indicating whether the corresponding features belong to semantic modality.

Cross Reconstruction Constraint

To ensure learning clean and complete retrieval features, we argue that such features should reconstruct the original sketch/image visual features by combined with sketch/image-domain features. To this end, the cross-reconstruction loss is introduced to ensure the reconstructed features are similar to the original features. These reconstructed features are formulated as:

$$\tilde{x}^{ske}_1 = R_{\theta'_{ske}}(x^{ret}_{ske} + x^{dom}_{ske}), \tag{9}$$

$$\tilde{x}^{ske}_2 = R_{\theta'_{ske}}(x^{ret}_{img} + x^{dom}_{ske}), \tag{10}$$

$$\tilde{x}^{img}_1 = R_{\theta'_{img}}(x^{ret}_{ske} + x^{dom}_{img}), \tag{11}$$

$$\tilde{x}^{img}_2 = R_{\theta'_{img}}(x^{ret}_{img} + x^{dom}_{img}), \tag{12}$$

where $R_{\theta'_{ske}}(\cdot)$ and $R_{\theta'_{img}}(\cdot)$ denote the reconstruction function on the sketch branch and image branch, respectively. Besides, $\tilde{x}^{ske}_1$ and $\tilde{x}^{ske}_2$ denote the reconstructed sketch features; $\tilde{x}^{img}_1$ and $\tilde{x}^{img}_2$ denote the reconstructed image features. Furthermore, the cross-reconstruction losses in sketch and image branch are written as:

$$\mathcal{L}_{rec\_ske} = ||\tilde{x}^{ske}_1 - x^{ske}||^2_2 + ||\tilde{x}^{ske}_2 - x^{ske}||^2_2, \tag{13}$$

$$\mathcal{L}_{rec\_img} = ||\tilde{x}^{img}_1 - x^{img}||^2_2 + ||\tilde{x}^{img}_2 - x^{img}||^2_2. \tag{14}$$

The total cross-reconstruction loss can be formulated as:

$$\mathcal{L}_{rec} = \mathcal{L}_{rec\_ske} + \mathcal{L}_{rec\_img}. \tag{15}$$

3.4 Objective and Optimization

Since the modality loss is constrained both on domain features and semantic features, we can formulate the total modality loss as:

$$\mathcal{L}_{mcls} = \mathcal{L}_{dmcls} + \mathcal{L}_{rmcls}. \tag{16}$$

Finally, the full objective of our proposed PDFD is:

$$\begin{aligned}\mathcal{L} = &\lambda_{adv} * \mathcal{L}_{adv} + \lambda_{ccls} * \mathcal{L}_{ccls} \\ &+ \lambda_{rec} * \mathcal{L}_{rec} + \lambda_{mcls} * \mathcal{L}_{mcls},\end{aligned} \tag{17}$$

where λ_{adv}, λ_{ccls}, λ_{rec} and λ_{mcls} are coefficients for balancing the overall performance.

4 Experiment

4.1 Datasets and Setup

There are two widely-used large-scale sketch datasets Sketchy [Sangkloy *et al.*, 2016] and TU-Berlin [Eitz *et al.*, 2012] for ZS-SBIR.

Sketchy originally consists of 75,479 sketches and 12,500 images from 125 categories. Liu *et al.* [2017a] extended the image gallery by collecting extra 60,502 images from ImageNet [Deng *et al.*, 2009] dataset, such that the total number of images is 73,002 in the extended version.

TU-Berlin originally consists of 20,000 unique free-hand sketches evenly distributed over 250 object categories. Compared to Sketchy, TU-Berlin only has category-level matches rather than instance-level matches.

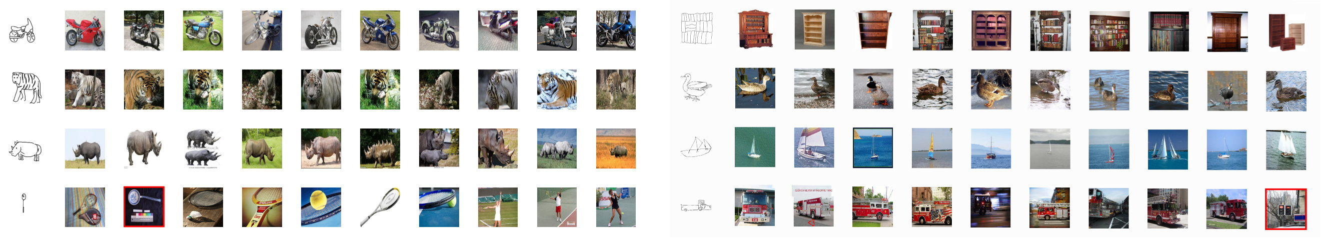
Following the same zero-shot data partitioning in SEM-PCYC [Dutta and Akata, 2019], we also follow the same evaluation criterion in most previous works [Dutta and Akata, 2019; Shen *et al.*, 2018] in terms of mean average precision (mAP@all) and precision considering the top 100 (Prec@100) retrievals.

4.2 Implementation Details

PDFD is trained with Adam [Kingma and Ba, 2014] optimizer on PyTorch with an initial learning rate $lr = 0.0001$, $\beta_1 = 0.5$, $\beta_2 = 0.99$. The input size of the image/sketch is 224×224. We use the grid search method to select the best coefficients, which are $\lambda_{adv} = 1.0$, $\lambda_{rec} = 1.0$, $\lambda_{mcls} = 1.0$, $\lambda_{ccls} = 0.01$ when training on Sketch and $\lambda_{adv} = 1.0$, $\lambda_{rec} = 0.5$, $\lambda_{mcls} = 0.4$, $\lambda_{ccls} = 0.4$ when training on TU-Berlin. VGG16 [Simonyan and Zisserman, 2014] pre-trained on ImageNet is adopted as a feature extractor. The word text-based embedding [Mikolov *et al.*, 2013] is adopted to extract 300-dimensional word vectors. Under the zero-shot setting, we only consider the seen classes when constructing the hierarchy embedding [Miller, 1995] for obtaining the class embedding. Therefore, the hierarchical embedding for Sketchy and TU-Berlin respectively contain 354 and 664 nodes.

	Methods	Feature Dimension	Sketchy		TU-Berlin	
			mAP@all	Prec@100	mAP@all	Prec@100
SBIR	Siamese CNN [Qi *et al.*, 2016]	64	0.132	0.175	0.109	0.141
	SaN [Yu *et al.*, 2017]	512	0.115	0.125	0.089	0.108
	GN Triplet [Sangkloy *et al.*, 2016]	1024	0.204	0.296	0.175	0.253
	3D Shape [Wang *et al.*, 2015]	64	0.067	0.078	0.054	0.067
	DSH (binary) [Liu *et al.*, 2017a]	64	0.171	0.231	0.129	0.189
	GDH (binary) [Zhang *et al.*, 2018]	64	0.187	0.259	0.135	0.212
ZSL	DeViSE [Frome *et al.*, 2013]	300	0.067	0.077	0.059	0.071
	JLSE [Zhang and Saligrama, 2016]	100	0.131	0.185	0.109	0.155
	SAE [Kodirov *et al.*, 2017]	100	0.216	0.293	0.167	0.221
	ZSH (binary) [Yang *et al.*, 2016]	64	0.159	0.214	0.141	0.171
ZS-SBIR	ZSIH (binary) [Shen *et al.*, 2018]	64	0.258	0.342	0.223	0.294
	CVAE [Kiran Yelamarthi *et al.*, 2018]	4096	0.196	0.284	0.005	0.001
	SEM-PCYC [Dutta and Akata, 2019]	64	0.349	0.463	0.297	0.426
	SEM-PCYC(binary) [Dutta and Akata, 2019]	64	0.344	0.399	0.293	0.392
	SAKE(binary) [Liu *et al.*, 2019]	64	0.364	0.487	0.359	0.481
	CSDB [Dutta and Biswas, 2019]	64	0.484	0.375	0.355	0.254
	PDFD (ours)	64	**0.623**	**0.726**	**0.460**	**0.595**
	PDFD (ours binary)	64	**0.638**	**0.755**	**0.386**	**0.542**
	SAKE [Liu *et al.*, 2019]	512	0.547	0.692	0.475	0.599
	PDFD (ours)	512	**0.661**	**0.781**	**0.483**	**0.600**

Table 1: ZS-SBIR performance of our proposed PDFD compared with existing SBIR, ZSL and ZS-SBIR approaches.



(a) Retrieval results on Sketchy.

(b) Retrieval results on TU-Berlin.

Figure 3: The top 10 images retrieved by our PDFD on the two datasets. The red borders indicate mis-retrieved images.

4.3 Comparison with Peer Methods

Apart from ZS-SBIR methods, some existing SBIR and ZSL approaches are also involved in retrieval comparison. The performances of all the comparisons are shown in Table 1, where we can observe that most ZS-SBIR methods outperform SBIR and ZSL methods while GN Triplet [Sangkloy *et al.*, 2016] and SAE [Kodirov *et al.*, 2017] reach the best performance in SBIR and ZSL, respectively. The main reason is that SBIR and ZSL methods are unable to take both domain gap and knowledge transfer into consideration. Therefore, the ZS-SBIR methods have better performance as they possess both the ability of reducing the domain gap and transferring the semantic knowledge. Due to the larger number of classes in TU-Berlin, all involved methods perform relatively worse on this dataset compared with Sketchy.

Most of the ZS-SBIR methods are conducted to retrieve based on 64-dimensional features, so PDFD generates 64-dimensional retrieval features for retrieval and outperforms the best competitor [Dutta and Biswas, 2019] by more than 13% on Sketchy and 10% on TU-Berlin. However, the SAKE [Liu *et al.*, 2019] adopts the 512-dimensional features for retrieval, and then applies iterative quantization algorithm [Liu *et al.*, 2017b] to obtain the 64-dimensional binary codes. For a fair comparison, PDFD also obtains 512-dimensional retrieval features and 64-dimensional binary codes. The result shows that our model significantly

Text-based embedding		Hierarchical embedding		Sketchy	TU-Berlin
Glove	Word2Vector	Path	Ji-Cn [Jiang and Conrath, 1997]		
✓				0.583	0.387
	✓			0.584	0.388
		✓		0.603	0.392
			✓	0.603	0.393
✓		✓		0.615	0.447
✓			✓	0.615	**0.460**
	✓	✓		0.622	0.447
	✓		✓	**0.623**	0.458

Table 2: The mAP@all results of PDFD using different semantic embedding and their combinations for ZS-SBIR.

#	Description	Sketchy	TU-Berlin
1	Baseline	0.377	0.338
2	Baseline + Progressive ($\mathcal{L}_{ccls}$)	0.481	0.374
3	Baseline + Progressive ($\mathcal{L}_{ccls}$) + Decomposition ($\mathcal{L}_{mcls}$)	0.510	0.396
4	Baseline + Progressive ($\mathcal{L}_{ccls}$) + Decomposition ($\mathcal{L}_{rec}$)	0.613	0.449
5	Baseline + Progressive ($\mathcal{L}_{ccls}$) + Decomposition ($\mathcal{L}_{mcls}$ + $\mathcal{L}_{rec}$)	**0.623**	**0.460**

Table 3: The mAP@all results of ablation study on our PDFD with several baselines for ZS-SBIR.

outperforms SAKE by around 11% on Sketchy and 1% on TU-Berlin when adopting 512-dimensional retrieval features.

All of these demonstrate the effectiveness of PDFD for domain gap reduction and semantic knowledge transfer. The retrieved images for sketches are shown in Figure 3. The red borders indicate the incorrectly retrieved images.

4.4 Effect of Semantic Knowledge Embedding

Since different types of semantic embedding have different impact on performance, we analyze the effects of different semantic embedding as well as different combinations of them on retrieval performance based on 64-dimensional retrieval features. Table 2 shows the quantitative results. As we can see, the combination of Word2Vec and Jiang-Conrath [Jiang and Conrath, 1997] hierarchical similarity reaches the highest mAP of 62.3% on Sketchy, while on TU Berlin dataset, the combination of Glove and Jiang-Conrath reaches the highest mAP of 46.0%. Note that we adopt the same embedding setting for all ablation studies.

4.5 Ablation Study

Five ablation studies are conducted to validate the effectiveness of our proposed PDFD as exhibited in Table 3, which are: 1) A baseline that simultaneously projects visual features and semantic supervision into a low-dimensional common space; 2) Adding our progressive projection strategy to the baseline that first learns semantic features with the original semantic supervision, and then projects them as retrieval features under the category classification loss $\mathcal{L}_{ccls}$; 3) Further adding our proposed feature decomposition to attain domain-independent retrieval features under the modality classification loss $\mathcal{L}_{mcls}$; 4) Replacing the modality classification loss $\mathcal{L}_{mcls}$ in "3)" with $\mathcal{L}_{rec}$ to validate the effectiveness of cross-reconstruction; 5) Full PDFD model. Moreover, all the retrieval features in ablation studies are 64-dimensional.

As shown in Table 3, our full model outperforms all baselines. The progressive projection strategy in "2)" improves baseline performance by around 10% on Sketchy and 4% on TU-Berlin, as this strategy benefits semantic knowledge transfer. By further decomposing visual features into semantic features and domain features under the modality classification loss $\mathcal{L}_{mcls}$, we can derive domain-independent retrieval features and improve the cross-modal retrieval performance. Moreover, the proposed cross-reconstruction loss $\mathcal{L}_{rec}$ encourages learning retrieval features with clean and complete semantic information, which improves the performance by a large margin. Finally, with all proposed modules, the full model reaches the highest mAP@all of 62.3% on Sketchy and mAP@all of 46.0% on TU-Berlin.

5 Conclusion

We have presented a novel progressive domain-independent feature decomposition network to address the problem of ZS-SBIR more effectively. On one hand, a progressive projection strategy is exploited to preserve the semantic information with the strong supervision of original semantic knowledge for learning semantic features. On the other hand, the cross-reconstruction loss is imposed to reduce the domain gap by ensuring that the retrieval features capture clean and complete semantic information. Experiments on two large-scale datasets show that our proposed PDFD significantly outperforms existing state-of-the-art methods in ZS-SBIR task.

Acknowledgments

This work is supported in part by the Key R&D Program-The Key Industry Innovation Chain of Shaanxi under Grant 2019ZDLGY03-02-01, and in part by the National Key R&D Program of China under Grant 2017YFE0104100 and 2016YFE0200400.

References

[Akata *et al.*, 2015] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(7):1425–1438, 2015.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[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*, 2019.

[Eitz *et al.*, 2012] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Trans. Graph.*, 31(4):44–1, 2012.

[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.

[Hu and Collomosse, 2013] Rui Hu and John Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Comput. Vis. Image Understand.*, 117(7):790–806, 2013.

[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.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[Kiran Yelamarthi *et al.*, 2018] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, pages 300–317, 2018.

[Kodirov *et al.*, 2017] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, pages 3174–3183, 2017.

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

[Liu *et al.*, 2017b] Li Liu, Mengyang Yu, and Ling Shao. Learning short binary codes for large-scale image retrieval. *IEEE Trans. Image Process.*, 26(3):1289–1299, 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, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv:1301.3781*, 2013.

[Miller, 1995] George A Miller. Wordnet: a lexical database for english. *ACM Commun.*, 38(11):39–41, 1995.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher 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 a siamese convolutional neural network. In *ICIP*, pages 2460–2464, 2016.

[Saavedra *et al.*, 2015] Jose M Saavedra, Juan Manuel Barrios, and S Orand. Sketch based image retrieval using learned keyshapes(lks). In *BMVC*, volume 1, page 7, 2015.

[Sangkloy *et al.*, 2016] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Trans. Graph.*, 35(4):119, 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.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[Wang *et al.*, 2015] Meng Wang, Chaokun Wang, Jeffrey Xu Yu, and Jun Zhang. Community detection in social networks: an in-depth benchmarking study with a procedure-oriented framework. In *VLDB*, pages 998–1009, 2015.

[Wei *et al.*, 2019] Kun Wei, Muli Yang, Hao Wang, Cheng Deng, and Xianglong Liu. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *ICCV*, pages 3741–3749, 2019.

[Xian *et al.*, 2016] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *CVPR*, pages 69–77, 2016.

[Yang *et al.*, 2016] Yang Yang, Yadan Luo, Weilun Chen, Fumin Shen, Jie Shao, and Heng Tao Shen. Zero-shot hashing via transferring supervised knowledge. In *ACM-MM*, pages 1286–1295, 2016.

[Yu *et al.*, 2017] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-net: A deep neural network that beats humans. *Int. J. Comput. Vis.*, 122(3):411–425, 2017.

[Zhang and Saligrama, 2016] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding. In *CVPR*, pages 6034–6042, 2016.

[Zhang *et al.*, 2018] Jingyi Zhang, Fumin Shen, Li Liu, Fan Zhu, Mengyang Yu, Ling Shao, Heng Tao Shen, and Luc Van Gool. Generative domain-migration hashing for sketch-to-image retrieval. In *ECCV*, pages 297–314, 2018.