Computer Vision and Image Understanding 218 (2022) 103412

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Zero-shot sketch-based image retrieval with structure-aware asymmetric disentanglement

Jiangtong Li, Zhixin Ling, Li Niu *, Liqing Zhang *

MOE Key Lab of Artificial Intelligence, Department of Computer Science and Engineering Shanghai Jiao Tong University, Shanghai, China

ARTICLE INFO

Communicated by Nikos Paragios

Keywords:
Disentangled representation
Domain translation
Zero-shot learning
Sketch-based image retrieval

ABSTRACT

The goal of Sketch-Based Image Retrieval (SBIR) is using free-hand sketches to retrieve images of the same category from a natural image gallery. However, SBIR requires all test categories to be seen during training, which cannot be guaranteed in real-world applications. So we investigate more challenging Zero-Shot SBIR (ZS-SBIR), in which test categories do not appear in the training stage. After realizing that sketches mainly contain structure information while images contain additional appearance information, we attempt to achieve structure-aware retrieval via asymmetric disentanglement. For this purpose, we propose our STRucture-aware Asymmetric Disentanglement (STRAD) method, in which image features are disentangled into structure features and appearance features while sketch features are only projected to structure space. Through disentangling structure and appearance space, bi-directional domain translation is performed between the sketch domain and the image domain. Extensive experiments demonstrate that our STRAD method remarkably outperforms state-of-the-art methods on three large-scale benchmark datasets.

1. Introduction

In the conventional SBIR setting, it assumes that the images and sketches in training and test sets share the same set of categories. However, in real-world applications, the categories of test sketches/images may be out of the scope of training categories, leading to a more challenging task called zero-shot sketch-based image retrieval (ZS-SBIR) (Shen et al., 2018), which assumes that test categories do not appear in the training stage. In the remainder of this paper, we refer to training (*resp.*, test) categories as seen (*resp.*, unseen) categories (Dupont, 2018). Traditional SBIR methods (Yu et al., 2017) suffer from sharp performance drop in ZS-SBIR setting, because traditional SBIR methods may take a shortcut by correlating sketches/images with their category labels and retrieving the images from the same category as the query sketch (Yelamarthi et al., 2018). This shortcut is very effective when test data share the same categories as training data, but can hardly generalize to unseen categories as depicted in Fig. 1.

To overcome the drawbacks of traditional SBIR methods in ZS-SBIR setting, several ZS-SBIR methods have been proposed to boost the performance on unseen categories, which can be categorized into the following groups: (1) Yelamarthi et al. (2018) used aligned sketch-image pairs (a sketch is drawn based on a given image and thus has roughly the same outline as this image) to learn better correlation between sketches and images. However, the aligned sketch-image pairs are either unavailable or very expensive; (2) Some works (Dutta and Akata, 2019; Dey et al., 2019; Zhang et al., 2020) employed category-level semantic information to reduce the gap between seen categories and unseen categories. Whereas, category-level semantic information like word vectors (Mikolov et al., 2013) are sometimes inaccessible or ambiguous. (3) Liu et al. (2019) located the catastrophic forgetting phenomenon and preserved the knowledge of model pretrained on ImageNet (Deng et al., 2009). Despite its competitive performance, it relies on auxiliary WordNet (Fellbaum, 2012) knowledge and its performance gain is mainly from the pre-training strategy; (4) Dutta et al. (2020) proposed to disentangle the representations of two domains (*i.e.*, sketch and image) into domain-independent and domain-specific representations. Nevertheless, this symmetric disentanglement approach is not well-tailored for SBIR task and ignores the asymmetric relation between sketch and image domain, that is, sketches only contain structure information (*e.g.*, outline, shape) while images additionally contain appearance information (*e.g.*, color, texture, and background).

Being aware of the asymmetric relation between sketch domain and image domain, we conjecture that the key of sketch-based image retrieval might be successfully matching the structure information between sketches and images. In this paper, we attempt to learn the structural correspondence between sketches and images on seen categories, which could generalize to unseen categories and facilitate ZS-SBIR task. For example, as shown in Fig. 1, the structural correspondence between sketches and images within the same category can be learned based on seen categories, *e.g.*, the structural similarity between

* Corresponding authors.
E-mail addresses: ustcnewly@sjtu.edu.cn (L. Niu), zhang-lq@cs.sjtu.edu.cn (L. Zhang).

https://doi.org/10.1016/j.cviu.2022.103412
Received 23 July 2021; Received in revised form 26 February 2022; Accepted 5 March 2022
Available online 17 March 2022

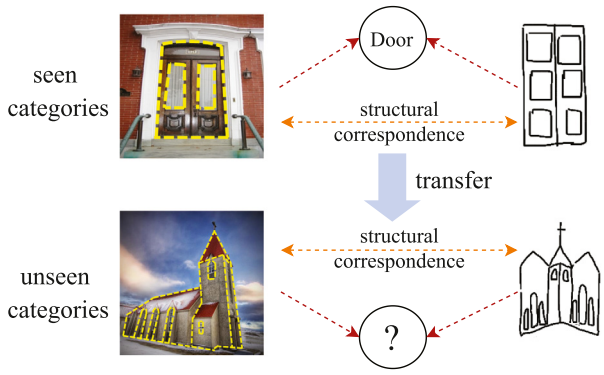

Fig. 1. Traditional SBIR methods which correlate sketches/images with their category labels cannot generalize well to unseen categories. Our STRAD method learns structural correspondence between sketches and images from seen categories, which can be transferred to unseen ones.

"door" sketch and "door" image *w.r.t.* the global/local contour and the layout of different components. In the test stage, given sketches and images from an unseen category "church", without knowing which category they belong to, we can still verify whether they belong to the same category on the premise of their structural similarity learnt from seen categories.

To obtain the structure features of sketches and images, we propose a STRucture-aware Asymmetric Disentanglement (STRAD) method to disentangle image feature into structure feature (*e.g.*, outline, shape) and appearance feature (*e.g.*, color, texture, and background) while projecting sketch feature into structure feature only. Our asymmetric disentanglement method is different from symmetric disentanglement in StyleGuide (Dutta et al., 2020), because our disentangled representations have explicit meanings (*i.e.*, structure and appearance) and are specifically designed for SBIR task. Note that although several previous methods (Yu et al., 2016; Liu et al., 2017) also intended to obtain the structure information of images by directly extracting edge maps from images, the low-level edge maps obtained in this way may not be very reliable due to possible noisy and redundant information. In contrast, our method could extract high-level robust structure features from both sketches and images.

Our proposed STRAD method is illustrated in Fig. 2. We first use a pre-trained model to extract features from sketches (*resp.*, images) as sketch (*resp.*, image) features. Then, the image features are disentangled into structure features and appearance features, while the sketch features are also projected to the structure space. Furthermore, bi-directional domain translation is performed through the structure features and appearance features. Concretely, for image-to-sketch translation, we project the image features to structure features and then generate sketch features from the structure features. For sketch-to-image translation, we project the sketch features to structure features, which are combined with variational appearance features to compensate for the appearance uncertainty when generating image features from the sketch features.

Finally, we perform retrieval in all three spaces (*i.e.*, structure space, image space, and sketch space), to combine the best of three worlds. Apparently, the retrieval in structure space and sketch space is structure-aware. Image feature is generated from structure feature and variational appearance feature. Since variational appearance feature is category-agnostic, the retrieval in image space is also structure-aware. The effectiveness of our proposed STRAD method is verified by comprehensive experimental results on three large-scale benchmark datasets. Our main contributions are summarized as follows:

- Based on the asymmetric relation between image domain and sketch domain, we design a novel STRucture-aware Asymmetric Disentanglement (STRAD) method to learn structural correspondence between these two domains.
- We design a hybrid retrieval strategy in three spaces, where different space has its own speciality and they can complement with each other to satisfy different retrieval requirement.
- Comprehensive results on three popular benchmark datasets show that our method significantly outperforms the state-of-the-art methods.

2. Related work

2.1. Zero-shot sketch-based image retrieval (ZS-SBIR)

Zero-shot sketch-based image retrieval (ZS-SBIR) was proposed by Shen et al. (2018). To reduce the intra-class variance in sketches and stabilize the training process, semantic information was leveraged in many models. In detail, Dutta and Akata (2019) proposed SEM-PCYC to combine the text information along with image and sketch feature generation and align the semantical representation in latent-space. Dey et al. (2019) proposed a new large-scale sketch-image dataset for ZS-SBIR and proposed a triplet loss-based network. Zhang et al. (2020) utilized the graph convolution network to align the image and sketch in the same latent-space. To reduce the gap between seen and unseen categories, a generative model along with aligned data pairs was proposed by Yelamarthi et al. (2018). To adapt the pre-trained model to ZS-SBIR without forgetting the knowledge from ImageNet, semantic-aware knowledge preservation was used in SAKE (Liu et al., 2019). To disentangle the representations of two domains (*i.e.*, sketch and image) into domain-independent and domain-specific representations, a symmetrical disentangle framework was proposed by Dutta et al. (2020). However, all of the above methods did not consider the special relation between images and sketches, and treated them equally in their models.

2.2. Disentangled representation

Disentangled representation learning aims to divide the latent representation into multiple units, with each unit corresponding to one latent factor (*e.g.*, position, scale, identity). Each unit is only affected by its corresponding latent factor, but not influenced by other latent factors.

Disentangled representation learning methods can be categorized into unsupervised methods and supervised methods according to whether supervision for latent factors is available. For unsupervised disentanglement, abundant methods have been developed, including InfoGAN (Chen et al., 2016), MTAN (Liu et al., 2018a). For supervised disentanglement, Kingma et al. (2014) used disentangled representation to enhance semi-supervised learning. Zheng et al. (2019) proposed DG-Net to integrate discriminative and generative learning using disentangled representation. Hadad et al. (2018) proposed a two-step disentanglement method, which disentangles the label information from the original representation and enables feature reconstruction from decomposed features. Besides, supervised disentanglement has been applied to different tasks, like face recognition (Liu et al., 2018b), image generation (Harsh Jha et al., 2018), and style transfer (Yang et al., 2019). Our work applies asymmetric disentangled representation learning to facilitate structure-aware retrieval.

2.3. Domain translation

Many domain translation approaches, like Pix2Pix (Isola et al., 2017), CycleGAN (Zhu et al., 2017) have been proposed, which can translate figures between two domains (*e.g.*, sketch domain and image domain). In this subsection, we mainly discuss the domain translation methods (Lee et al., 2018; Gonzalez-Garcia et al., 2018) based on disentangled representation. Overall speaking, they disentangle latent representation into domain-specific representation and domain-invariant representation. In our problem, structure (*resp.*, appearance) features

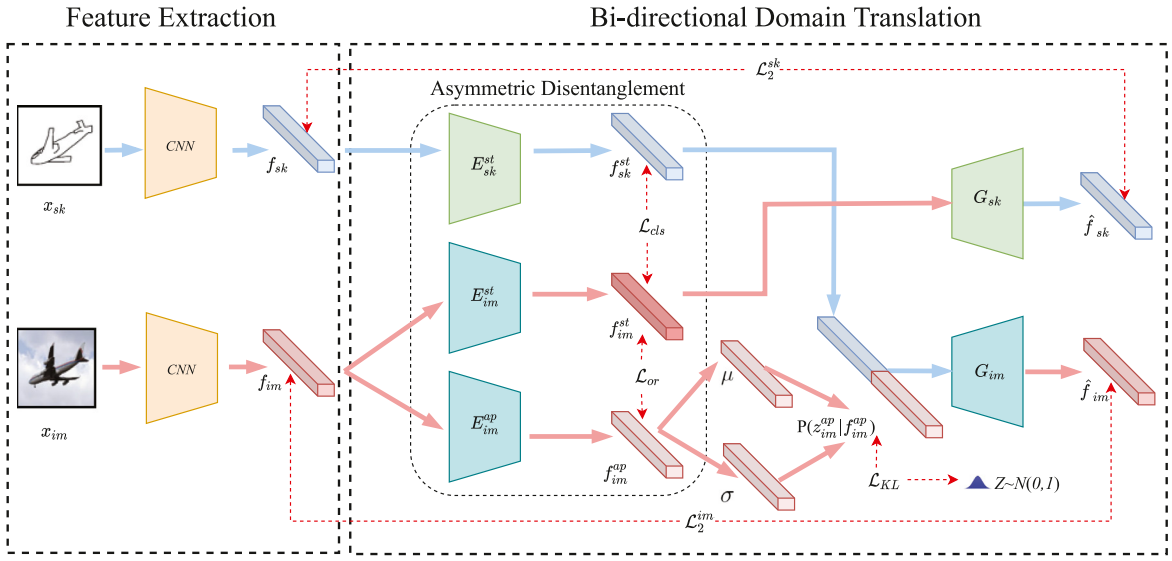

Fig. 2. An overview of our STRAD method. We first adopt VGG-16 (Simonyan and Zisserman, 2015) to extract features from images and sketch. Then we disentangle the image feature into appearance feature and structure feature, through which bi-directional domain translation is performed between the image domain and the sketch domain.

can be treated as domain-invariant (*resp.*, specific) representation. The translation between two domains in previous works (Lee et al., 2018; Gonzalez-Garcia et al., 2018) is generally symmetric. In contrast, the translation between sketch domain and image domain in our problem is asymmetric because image domain has additional domain-specific representation compared with sketch domain.

3. Methodology

3.1. Problem definition

In this paper, we focus on zero-shot sketch-based image retrieval, where only the sketches and images from seen categories are used for training. In the test stage, our proposed method is expected to use the sketches to retrieve the images, the categories of which are unseen during training.

Formally, given a sketch dataset $S_{sk} = \{(\mathbf{x}_{sk,i}, y_i)|y_i \in \mathcal{Y}\}$ and an image dataset $S_{im} = \{(\mathbf{x}_{im,j}, y_j)|y_j \in \mathcal{Y}\}$, where $\mathcal{Y}$ is category label set, and $(\mathbf{x}_{sk,i}, y_i)$ (*resp.*, $(\mathbf{x}_{im,j}, y_j)$) represents a sketch (*resp.*, image) with its corresponding category label, we follow the zero-shot setting in Shen et al. (2018) to split all categories $\mathcal{Y}$ into $\mathcal{Y}^{tr}$ and $\mathcal{Y}^{te}$, in which no overlap exists between two label sets, *i.e.*, $\mathcal{Y}^{tr} \cap \mathcal{Y}^{te} = \emptyset$. Based on the partition of label set $\mathcal{Y}$, we can split the sketch (*resp.*, image) dataset into S^{tr}_{sk} and S^{te}_{sk} (*resp.*, S^{tr}_{im} and S^{te}_{im}). In the training stage, our model can only process the data in S^{tr}_{sk} and S^{tr}_{im}. In the test, given a sketch x_{sk}, our model needs to retrieve the images belonging to the same category from test image gallery S^{te}_{im}.

3.2. Our framework

An overview of our method is illustrated in Fig. 2. In this section, we will introduce our method from Asymmetric Disentanglement and Bi-direction Domain Translation.

3.2.1. Asymmetric disentanglement

Given an image $\mathbf{x}_{im}$ and a sketch $\mathbf{x}_{sk}$ from the same category y, we first use fixed backbone model, *i.e.*, VGG-16 (Simonyan and Zisserman, 2015), to produce their image feature $\mathbf{f}_{im}$ and sketch feature $\mathbf{f}_{sk}$. For image feature $\mathbf{f}_{im}$, we adopt two image encoders E^{st}_{im} and E^{ap}_{im} to disentangle $\mathbf{f}_{im}$ into image structure feature $\mathbf{f}^{st}_{im}$ and image appearance feature $\mathbf{f}^{ap}_{im}$. Besides, to project sketch feature $\mathbf{f}_{sk}$ to the same structure space as $\mathbf{f}^{st}_{im}$, a sketch encoder E^{st}_{sk} is adopted to obtain sketch structure feature $\mathbf{f}^{st}_{sk}$. The above process is formulated as follows,

$$\mathbf{f}^{ap}_{im} = E^{ap}_{im}(\mathbf{f}_{im}); \quad \mathbf{f}^{st}_{im} = E^{st}_{im}(\mathbf{f}_{im}); \quad \mathbf{f}^{st}_{sk} = E^{st}_{sk}(\mathbf{f}_{sk}). \tag{1}$$

To capture the structural correspondence between images and sketches, we expect the structure features from both images and sketches to be aligned in the same space. Moreover, in the structure space shared by sketch and image, we expect intra-class coherence and inter-class separability across different domains (*i.e.*, sketch domain and image domain). Specifically, we expect to pull close the image/sketch structure features of the same category and push apart the image/sketch structure features from different categories. It has been proved in Liu et al. (2019) that a simple classification loss can accomplish the above task well. Therefore, we employ a structure classifier on both image structure features and sketch structure features to distinguish their category labels, by using the cross-entropy classification loss:

$$\begin{aligned}\mathcal{L}_{cls} = &- \log \frac{\exp(\boldsymbol{w}_y^{\mathrm{T}}\mathbf{f}^{st}_{im} + b_y)}{\sum_{k\in\mathcal{Y}^{tr}} \exp(\boldsymbol{w}_k^{\mathrm{T}}\mathbf{f}^{st}_{im} + b_k)} \\ &- \log \frac{\exp(\boldsymbol{w}_y^{\mathrm{T}}\mathbf{f}^{st}_{sk} + b_y)}{\sum_{k\in\mathcal{Y}^{tr}} \exp(\boldsymbol{w}_k^{\mathrm{T}}\mathbf{f}^{st}_{sk} + b_k)},\end{aligned} \tag{2}$$

where $\boldsymbol{w}_k$ and b_k are learnable parameters in the structure classifier corresponding to category k. Recall that y is the category label of $\mathbf{x}_{im}$ and $\mathbf{x}_{sk}$. Although the structure space regulated by Eq. (2) seems sufficient for retrieval, the following modules to be introduced can further help capture the structural correspondence and improve the generalization ability.

After enforcing the structure features of sketches and images to the same structure space, we further expect that the appearance features of images only contain complementary information (*e.g.*, color, texture, and background) to structure features. To reinforce the disentanglement of image features, we impose an orthogonal constraint between structure and appearance features of images based on cosine similarity (Shukla et al., 2019):

$$\mathcal{L}_{or} = \cos(\mathbf{f}^{ap}_{im}, \mathbf{f}^{st}_{im}) = \frac{\mathbf{f}^{ap}_{im} \cdot \mathbf{f}^{st}_{im}}{\|\mathbf{f}^{ap}_{im}\|_2 \|\mathbf{f}^{st}_{im}\|_2}, \tag{3}$$

where $\cdot$ means the dot product between two vectors. Note that $\mathbf{f}^{ap}_{im}$ and $\mathbf{f}^{st}_{im}$ are the output of ReLU activation, so $\cos(\mathbf{f}^{ap}_{im}, \mathbf{f}^{st}_{im})$ is always non-negative and minimizing Eq. (3) will push $\cos(\mathbf{f}^{ap}_{im}, \mathbf{f}^{st}_{im})$ towards zero.

3.2.2. Bi-directional domain translation

To learn better disentangled representations and fully utilize the disentangled image features, we perform bi-directional domain translation between the sketch and the image domain.

For image-to-sketch translation, we aim to translate image feature $\mathbf{f}_{im}$ to sketch feature through image structure feature $\mathbf{f}_{im}^{st}$. We employ a decoder G_{sk} to generate sketch feature $\hat{\mathbf{f}}_{sk}$ by $\hat{\mathbf{f}}_{sk} = G_{sk}(\mathbf{f}_{im}^{st})$. Considering that $\hat{\mathbf{f}}_{sk}$ and $\mathbf{f}_{sk}$ belong to the same category, we enforce the generated sketch features to be close to the real sketch features from the same category by

$$\mathcal{L}_2^{sk} = \|\mathbf{f}_{sk} - \hat{\mathbf{f}}_{sk}\|_2. \tag{4}$$

For sketch-to-image translation, we aim to translate sketch feature $\mathbf{f}_{st}$ to image feature through its sketch structure feature $\mathbf{f}_{sk}^{st}$. However, images contain extra appearance information (*e.g.*, color, texture, and background) compared with sketches, so it is necessary to compensate appearance uncertainty when translating from structure features to image features. Therefore, in the training stage, appearance feature $\mathbf{f}_{im}^{ap}$ could be integrated with sketch structure feature $\mathbf{f}_{sk}^{st}$ to generate image feature.

During testing, given a sketch, we also hope to generate its image feature to enable retrieval in the image space. Nevertheless, we do not have the corresponding appearance feature. A common solution is stochastic sampling. We introduce a variational estimator V_{im}^{ap} to approximate the variational Gaussian distribution $P(\mathbf{z}_{im}^{ap}|\mathbf{f}_{im}^{ap})$ based on $\mathbf{f}_{im}^{ap}$, that is, $(\boldsymbol{\mu}_{im}^{ap}, \boldsymbol{\sigma}_{im}^{ap}) = V_{im}^{ap}(\mathbf{f}_{im}^{ap})$. Then, we use Kullback–Leibler divergence to enforce $P(\mathbf{z}_{im}^{ap}|\mathbf{f}_{im}^{ap})$ to be close to prior distribution $\mathcal{N}(\mathbf{0}, \mathbf{1})$ to support stochastic sampling:

$$\mathcal{L}_{KL} = KL(\mathcal{N}(\boldsymbol{\mu}_{im}^{ap}, \boldsymbol{\sigma}_{im}^{ap}) \parallel \mathcal{N}(\mathbf{0}, \mathbf{1})). \tag{5}$$

After using reparameterization trick (Kingma and Welling, 2014) to sample variational appearance feature $\mathbf{z}_{im}^{ap}$, *i.e.*, $\mathbf{z}_{im}^{ap} = \boldsymbol{\mu}_{im}^{ap} + \epsilon\boldsymbol{\sigma}_{im}^{ap}$, where ϵ is sampled from $\mathcal{N}(0, 1)$, we employ a decoder G_{im} to generate $\hat{\mathbf{f}}_{im} = G_{im}([\mathbf{z}_{im}^{ap}, \mathbf{f}_{sk}^{st}])$ based on $\mathbf{z}_{im}^{ap}$ and $\mathbf{f}_{sk}^{st}$, where $[\cdot, \cdot]$ means concatenating two vectors. Considering that $\hat{\mathbf{f}}_{im}$ has the same category label as $\mathbf{f}_{im}$ and its appearance uncertainty comes from $\mathbf{f}_{im}$, we enforce $\hat{\mathbf{f}}_{im}$ to be close to $\mathbf{f}_{im}$ with

$$\mathcal{L}_2^{im} = \|\mathbf{f}_{im} - \hat{\mathbf{f}}_{im}\|_2. \tag{6}$$

By performing image-to-sketch translation, we expect that the image structure features contain the necessary structure information to generate sketch features. By performing sketch-to-image translation, we expect that the appearance features contain the necessary information of appearance uncertainty to compliment the sketch structure features when generating image features. Therefore, bi-directional domain translation could cooperate with classification loss and orthogonal loss to assist feature disentanglement.

Finally, the full objective function can be expressed as

$$\mathcal{L} = \mathcal{L}_{or} + \mathcal{L}_{cls} + \mathcal{L}_{KL} + \mathcal{L}_2^{im} + \mathcal{L}_2^{sk}. \tag{7}$$

3.3. Retrieval strategy

In the test stage, we perform retrieval in three spaces: structure, sketch, and image spaces. Given a sketch $\mathbf{x}_{sk}$ with sketch feature $\mathbf{f}_{sk}$ and an image $\mathbf{x}_{im}$ with image feature $\mathbf{f}_{im}$, we compare them in three spaces to combine the best of all worlds.

(1) **Structure space**: We project $\mathbf{f}_{im}$ and $\mathbf{f}_{sk}$ into the structure space by $\mathbf{f}_{im}^{st} = E_{im}^{st}(\mathbf{f}_{im})$ and $\mathbf{f}_{sk}^{st} = E_{sk}^{st}(\mathbf{f}_{sk})$, so we can calculate the distance in structure space $\mathcal{D}_{st} = 1 - cos(\mathbf{f}_{im}^{st}, \mathbf{f}_{sk}^{st})$.

(2) **Sketch space**: For image $\mathbf{x}_{im}$, based on its image structure feature $\mathbf{f}_{im}^{st}$, we employ the sketch decoder G_{sk} to generate its sketch feature $\hat{\mathbf{f}}_{sk} = G_{sk}(\mathbf{f}_{im}^{st})$, so we can calculate the distance in sketch space $\mathcal{D}_{sk} = 1 - cos(\hat{\mathbf{f}}_{sk}, \mathbf{f}_{sk})$.

(3) **Image space**: For sketch $\mathbf{x}_{sk}$, based on sketch structure feature $\mathbf{f}_{sk}^{st}$ and variational appearance feature $\mathbf{z}_i$ sampled from $\mathcal{N}(\mathbf{0}, \mathbf{1})$, we employ image decoder G_{im} to generate image feature. We can generate N image features by sampling N times and use the average to represent the final image feature $\hat{\mathbf{f}}_{im}$:

$$\hat{\mathbf{f}}_{im} = \frac{1}{N}\sum_{i=1}^{N} G_{im}([\mathbf{z}_i, \mathbf{f}_{sk}^{st}]). \tag{8}$$

where $[\cdot, \cdot]$ means concatenation of two vectors and $\mathbf{z}_i$ is sampled from $\mathcal{N}(0, 1)$.

So we can calculate the distance in image space $\mathcal{D}_{im} = 1 - cos(\mathbf{f}_{im}, \hat{\mathbf{f}}_{im})$. Finally, we calculate the weighted average of three distances for the best retrieval:

$$\mathcal{D}_{fusion} = \lambda_1(\mathcal{D}_{im} + \mathcal{D}_{sk}) + \lambda_2\mathcal{D}_{st}, \tag{9}$$

where λ_1 and λ_2 are hyper-parameters to balance different spaces and we apply the constraint $2\lambda_1 + \lambda_2 = 1$. Since sketch feature is generated based on structure feature, retrieval in sketch space is obviously structure-aware. Image feature is generated based on structure feature and variational appearance feature $\mathbf{z}_i$, which is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{1})$ for any category and category-agnostic. Therefore, retrieval in image space mainly depends on structure information and is also structure-aware.

4. Experiment

4.1. Dataset

We evaluate our STRAD method and all the other baselines on three large-scale benchmark datasets: Sketchy (Sangkloy et al., 2016), TU-Berlin (Eitz et al., 2012), and QuickDraw (Dey et al., 2019). Following the same setting in Shen et al. (2018), Sketchy and TU-Berlin are extended with images obtained from Liu et al. (2017).

As for the seen/unseen category split, following Xu et al. (2019), we use the 104/21 split for Sketchy and the 194/56 split in TU-Berlin for fair comparison in the main submission. For other kinds data splits, we conduct more experiments in Supplementary Material. Due to the limitation of space, the details of datasets, seen/unseen category split, and implementation are left to Supplementary Material.

4.2. Comparison with existing methods

We compare our model with 15 prior methods, which can be categorized into three categories: sketch-based image retrieval (SBIR), zero-shot learning (ZSL), and zero-shot SBIR (ZS-SBIR). The SBIR baselines include Siamese (Qi et al., 2016), and SaN (Yu et al., 2017). The ZSL baselines include ESZSL (Romera-Paredes and Torr, 2015), SAE (Kodirov et al., 2017), and CMT (Socher et al., 2013). For a fair comparison, we use fine-tuned VGG-16 as backbone for all methods except SaN, which specifically designs a backbone for SBIR. Among all the previous works on ZS-SBIR, either two different backbones or a single backbone can be used to extract image and sketch features. In this paper, we conduct experiments in both settings. For the "double backbone" setting, we compare with CVAE (Yelamarthi et al., 2018; Xu et al., 2019), SEM-PCYC (Dutta and Akata, 2019), Doodle (Dey et al., 2019), StyleGuide (Dutta et al., 2020), PCMSN (Deng et al., 2020), SketchGCN (Zhang et al., 2020), and PDFD (Xu et al., 2020). For the "single backbone" setting, we compare with SAKE (Liu et al., 2019) because they use single backbone in their paper. The difference between the "single backbone" (SB) and "double backbone" (DB) is whether we separately fine-tune the backbone on image and sketch. For the "single backbone" setting, we fine-tune the pre-trained on the mixture of images and sketches, and produce only one backbone model to extract image/sketch features For the "double backbone" setting, we fine-tune the pre-trained separately on images and sketches, and produce two backbone model to extract image and sketch features separately. During training, we fix the backbone for all methods except for Doodle (Dey et al., 2019) and SAKE (Liu et al., 2019). For Doodle (Dey

Table 1
Comparison of our STRAD method and baselines on Sketchy, TU-Berlin, and QuickDraw datasets. (D) is short for "double backbone" setting and (S) is short for "single backbone" setting. Best results are denoted in boldface in both settings.

	Method	Dim	Sketchy Ext.			TU-Berlin Ext.			QuickDraw Ext.		
			P@200	mAP@200	mAP@all	P@200	mAP@200	mAP@all	P@200	mAP@200	mAP@all
SBIR	SaN (Yu et al., 2017)	512	0.153	0.058	0.055	0.101	0.042	0.047	0.042	0.009	0.039
	Siamese (Qi et al., 2016)	64	0.256	0.153	0.158	0.083	0.037	0.041	0.040	0.007	0.038
ZSL	ESZSL (Romera-Paredes and Torr, 2015)	1024	0.209	0.118	0.109	0.085	0.034	0.041	0.063	0.018	0.059
	SAE (Kodirov et al., 2017)	300	0.261	0.145	0.152	0.104	0.046	0.040	0.068	0.019	0.066
	CMT (Socher et al., 2013)	300	0.273	0.158	0.151	0.108	0.049	0.047	0.063	0.017	0.061
	SSE (Zhang and Saligrama, 2015)	100	0.202	0.125	0.131	0.026	0.003	0.006	0.079	0.029	0.081
ZS-SBIR (D)	Simple DB	4096	0.321	0.190	0.229	0.146	0.068	0.061	0.071	0.031	0.077
	CVAE (Yelamarthi et al., 2018)	4096	0.393	0.263	0.291	0.152	0.075	0.077	0.064	0.018	0.063
	Xu et al. (2019)	512	0.428	0.311	0.352	0.191	0.115	0.111	0.072	0.030	0.073
	SEM-PCYC (Dutta and Akata, 2019)	64	0.438	0.316	0.372	0.195	0.117	0.108	0.094	0.037	0.112
	Doodle (Dey et al., 2019)	4096	0.432	0.301	0.317	0.193	0.112	0.107	0.098	0.037	0.109
	StyleGuide (Dutta et al., 2020)	4096	0.430	0.285	0.348	0.178	0.118	0.103	0.083	0.035	0.101
	PCMSN (Deng et al., 2020)	64	0.439	0.325	0.361	0.199	0.122	0.112	0.095	0.041	0.108
	SketchGCN (Zhang et al., 2020)	2048	0.451	0.348	0.388	0.213	0.138	0.111	0.097	0.040	0.116
	PDFD (Xu et al., 2020)	512	0.478	0.361	0.424	0.229	0.148	0.119	0.112	0.048	0.132
	STRAD	1024	**0.502**	**0.379**	**0.458**	**0.245**	**0.154**	**0.124**	**0.126**	**0.054**	**0.141**
ZS-SBIR (S)	Simple SB	4096	0.497	0.367	0.399	0.262	0.162	0.122	0.166	0.071	0.173
	SAKE (Liu et al., 2019)	512	0.519	0.394	0.433	0.287	0.181	0.151	0.184	0.084	0.201
	STRAD	1024	**0.537**	**0.413**	**0.463**	**0.301**	**0.199**	**0.163**	**0.191**	**0.086**	**0.221**

et al., 2019), we strictly follow their paper to train the backbone during the training. For SAKE (Liu et al., 2019), we strictly follow the training strategy introduced in their paper.[1] Due to the calculation of mAP@200 in Dey et al. (2019), Zhang et al. (2020) is different from ours, their reported mAP@200 are different from our reproduced results. We also report the performance of fine-tuned single backbone (*resp.*, double backbones) as "Simple SB" (*resp.*, "Double DB") in Table 1, where cosine distance between image and sketch features is used for retrieval.

Based on Table 1, we find all SBIR and ZSL baselines underperform the ZS-SBIR baselines due to their poor generalization ability from seen categories to unseen ones. On the TU-Berlin, the results of several methods (Liu et al., 2019; Dutta and Akata, 2019; Dutta et al., 2020; Zhang et al., 2020) are worse than those reported in their papers due to different seen/unseen category splits. In particular, the number of unseen categories under our split is twice larger than that in Liu et al. (2019), Zhang et al. (2020). Our split criterion also prevents information leakage from ImageNet-1k to unseen categories. The overall results on TU-Berlin are lower than those on Sketchy due to larger number of unseen categories (56 *v.s.* 21). Furthermore, the overall results on TU-Berlin are higher than those on QuickDraw since sketches of QuickDraw were drawn by amateurs.

In "double backbone" setting, our STRAD excels the state-of-the-art methods by 2.4% on Sketchy, 1.8% on TU-Berlin, and 1.4% on QuickDraw in terms of P@200. In "single backbone" setting, we find that "Simple SB" outperforms most methods in "double backbone" setting, which reveals that it might be the best solution for SBIR task to use a single model to pull close the image and sketch space. Starting from "Simple SB", the performance gain of STRAD is smaller than that in "double backbone" setting, because a single backbone has already filtered out most differences between images and sketches. However, there still remains extra appearance information in image features, so STRAD also outperforms SAKE (Liu et al., 2019) and achieves the best results on all datasets.

To further demonstrate the effectiveness of our method, we report the results on TU-Berlin following the splits in Liu et al. (2019), the results in generalized ZS-SBIR setting, and the results of ablation study in Supplementary Material.

[1] Since SAKE (Liu et al., 2019) starts from backbone model pre-trained on ImageNet-1k to prevent knowledge forgetting, we do not change their setting.

Table 2
Comparison of our STRAD method and the three feature spaces on Sketchy, TU-Berlin, and QuickDraw datasets in the ZS-SBIR setting. "Im", "Sk" and "St" represent the retrieval performance in "sketch space", "image space" and "structure space". "Im + Sk", "Im + St", and "Sk + St" represent the retrieval performance in the combination of "image and sketch spaces", "image and structure spaces", and "sketch and structure spaces". Best results are denoted in boldface in both settings.

	Feature	Sketchy Ext.	TU-Berlin Ext.	QuickDraw Ext.
Double Backbone	Im	0.464	0.214	0.104
	Sk	0.477	0.221	0.112
	St	0.481	0.229	0.118
	Im + Sk	0.489	0.227	0.117
	Im + St	0.492	0.235	0.120
	Sk + St	0.497	0.239	0.123
	STRAD	**0.502**	**0.245**	**0.126**
Single Backbone	Im	0.498	0.269	0.164
	Sk	0.511	0.281	0.171
	St	0.518	0.288	0.178
	Im + Sk	0.521	0.289	0.177
	Im + St	0.528	0.293	0.185
	Sk + St	0.531	0.297	0.188
	STRAD	**0.537**	**0.301**	**0.191**

4.3. Ablation study

4.3.1. Effect of different feature spaces

To further study the usefulness of the three different features, we perform retrieval with these three features on two backbone settings (*i.e.* "single backbone" and "double backbone") on Table 2. The results show that although structure space works best among all three spaces, its cooperation with the other two could further boost the performance.

4.3.2. Usage of image appearance feature

To further reveal the usage of the image appearance feature, we add the classification loss towards the image appearance feature (f_{im}^{ap}). Considering the sketches do not have appearance feature and the classification loss is also applied to structure feature, we use the sketch structure feature to match with image appearance feature when retrieving using appearance feature. In Table 3, we show the retrieval performance with sketch space, image space, structure space, appearance space and the combination of different feature spaces on "double backbone" setting.

Table 3
Comparison of our STRAD method and the three feature spaces with or without the "$\mathcal{L}_{cls}$ on f_{im}^{ap}" on Sketchy, TU-Berlin, and QuickDraw datasets in the ZS-SBIR setting. "Im", "Sk", "St", "Ap", and "Comb" represent the retrieval performance in "sketch space", "image space", "structure space", "appearance space", and the combination of all space. "diff" represents the performance gap of "Comb" without and with the "$\mathcal{L}_{cls}$ on f_{im}^{ap}". Best results are denoted in boldface.

	Feature	Sketchy Ext.	TU-Berlin Ext.	QuickDraw Ext.
- $\mathcal{L}_{cls}$ on f_{im}^{ap}	Im	0.464	0.214	0.104
	Sk	0.477	0.221	0.112
	St	0.481	0.229	0.118
	Ap	–	–	–
	Comb	**0.502**	**0.245**	**0.126**
+ $\mathcal{L}_{cls}$ on f_{im}^{ap}	Im	0.463	0.215	0.103
	Sk	0.476	0.220	0.113
	St	0.480	0.230	0.117
	Ap	0.279	0.102	0.062
	Comb	0.381	0.171	0.097
	diff	−0.121	−0.074	−0.029

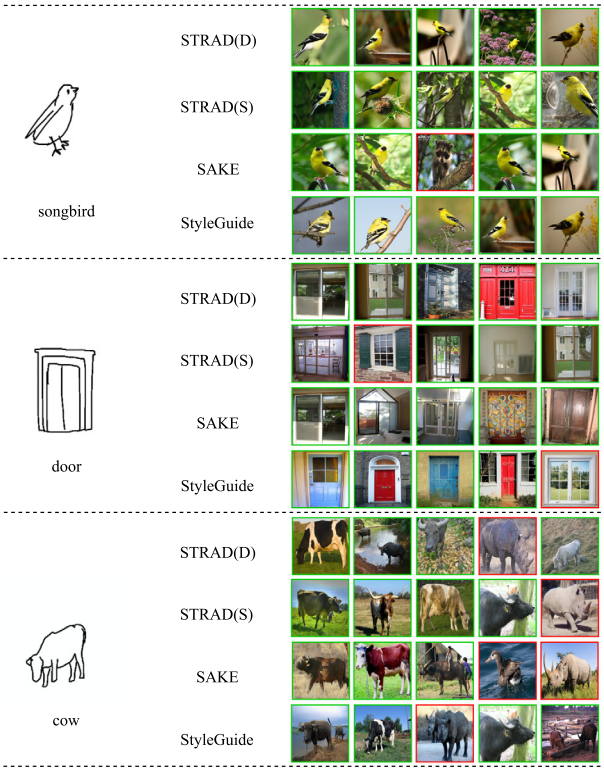

Fig. 3. Top-5 images retrieved by STRAD (double backbone), STRAD (single backbone), SAKE (Liu et al., 2019), and StyleGuide (Dutta et al., 2020) methods on Sketchy. The green (*resp.*, red) border indicates the correct (*resp.*, incorrect) retrieval results.

From the experimental results, we can find that after applying classification loss to image appearance feature (f_{im}^{ap}), the image appearance feature could be used for retrieval, however, the performance on "appearance space" is quite low. Besides, if we add the appearance space to the combination of difference spaces, we also find that the overall performance decrease by a large margin. We suspect that the KL loss is in conflict with the classification loss, which makes the category information within the appearance feature is less than that in structure feature. Considering that the KL loss is quite important for feature disentanglement and image feature generation, applying the classification loss on image appearance feature (f_{im}^{ap}) would be harmful for the overall performance.

Table 4
Comparison of our STRAD method when the backbone is fixing or trainable on Sketchy, TU-Berlin, and QuickDraw datasets in term of P@200 . "diff" represents the performance gap between "Fixing" and "Trainable". "S" and "D" represent the single backbone and double backbone setting.

Feature	Backbone	Sketchy Ext.	TU-Berlin Ext.	QuickDraw Ext.
STRAD (D)	Fixing	0.502	0.245	0.126
	Trainable	0.427	0.178	0.082
	diff	−0.075	−0.067	−0.044
STRAD (S)	Fixing	0.537	0.301	0.191
	Trainable	0.442	0.217	0.122
	diff	−0.095	−0.084	−0.069

4.3.3. Effect of backbone

To further reveal the effect of backbone, in Table 4, we compare the performance when the backbone is fixing or trainable in terms of P@200. From the results, we can find that when the backbone is trainable, the retrieval performance drops by a large margin, which is mainly caused by the catastrophic forgetting. For the catastrophic forgetting problem, SAKE (Liu et al., 2019) incorporates the knowledge distillation during training, however, since our main contribution is not knowledge preserving, we simply fix the backbone during training.

4.4. Case study

4.4.1. Comparison with existing methods

In Fig. 3, we show top-5 retrieval results of STRAD (double backbone), STRAD (single backbone), SAKE (Liu et al., 2019), and StyleGuide (Dutta et al., 2020), based on which we have the following observations. (a) Our STRAD is adept at capturing the correspondence between the retrieved images and the given sketch *w.r.t.* both local structure information (*e.g.*, door-case) and global structure information (*e.g.*, global grid structure of door). Besides, in both "single backbone" and "double backbone" settings, our STRAD is able to retrieve the images with cluttered background (*e.g.*, a bird with intricate leaves/flowers behind), which benefits from our structure-aware retrieval in three spaces. The above advantages come from the combination of three retrieval spaces. (b) For baselines, SAKE (Liu et al., 2019) with single backbone can retrieve images with cluttered background while StyleGuide (Dutta et al., 2020) with double backbones can barely tolerate the complex backgrounds. One possible reason is that using the same backbone for the sketch domain and image domain is to share a large amount of parameters for feature extraction, which would eliminate the background differences between these two domains. More case study of comparison among three retrieval space and the failure cases can be found in Supplementary Material.

4.4.2. Comparison among three retrieval spaces

In this section, we present some retrieval results of our STRAD (D) on Sketchy dataset, as shown in Fig. 4. For each test category of Sketchy dataset, we take one sketch as an example and present top-10 retrieved images in the structure space, sketch space, image space, and the combination of all three spaces from the bottom row to the top row. The advantages of these three feature space can be summarized as

- Cluttered background can be handled in the image space;
- Clean and full objects are prone to be retrieved in the sketch space;
- Structure space owns the ability to match local information.

More details about the analyses of these three feature spaces can be found in Sec. 7.1 of the supplementary.

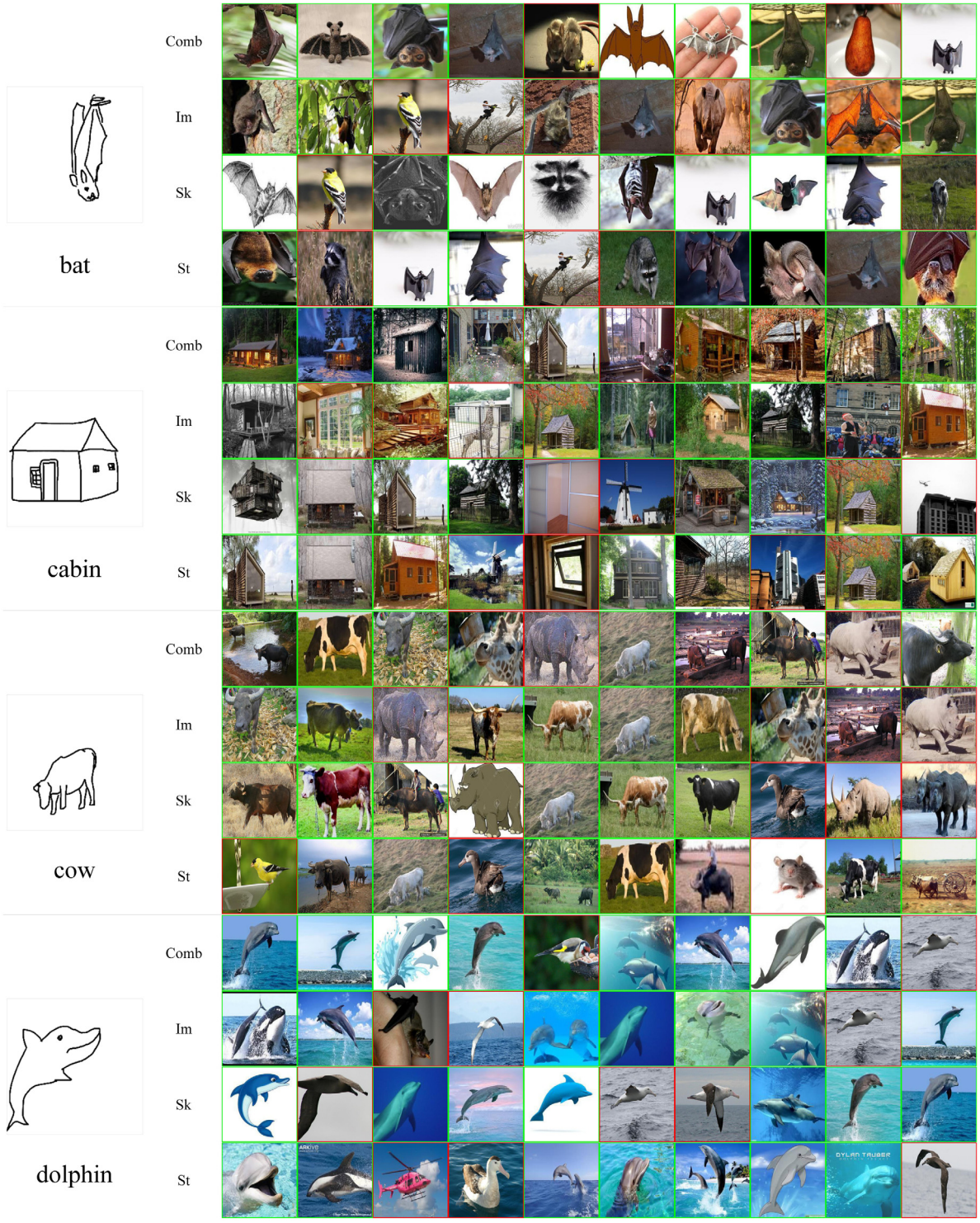

Fig. 4. Top-10 images from four categories (*i.e.* bat, cabin, cow, and dolphin) retrieved from image feature space (Im), sketch feature space (Sk), structure feature space (St), and the combination of these three feature spaces (Comb) on Sketchy. The green (*resp.*, red) border indicates the correct (*resp.*, incorrect) retrieval results.

4.4.3. Visualization of image disentanglement

In this section, we visualize some images that have similar structure or appearance features to demonstrate the effectiveness of the disentanglement in Fig. 5. For the facilitation of similarity measurement, we visualize the images that have similar structure or appearance features based on the t-SNE visualization for image structure feature and image appearance in Fig. 5. In this figure, we can find that in image structure space, similar features usually have similar structure information, like the outline poses or shapes. However, in the image appearance space, similar features usually have similar appearance information, like the color and background.

4.4.4. Comparison among different datasets

To reveal how the quality of sketches and images affect the retrieval performance, we visualize the retrieval results retrieved on different datasets on the same category in Fig. 6. For each dataset, we randomly

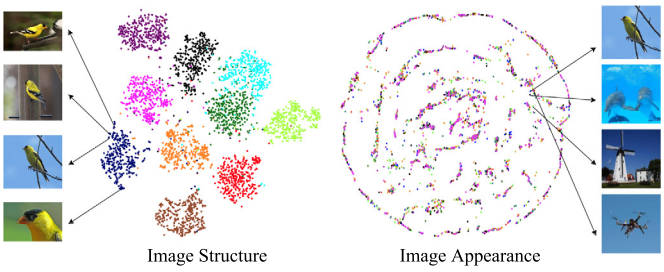

Fig. 5. The t-SNE visualization of "image structure features" and "image appearance features" on Sketchy test set long with some corresponding images. Best viewed in color.

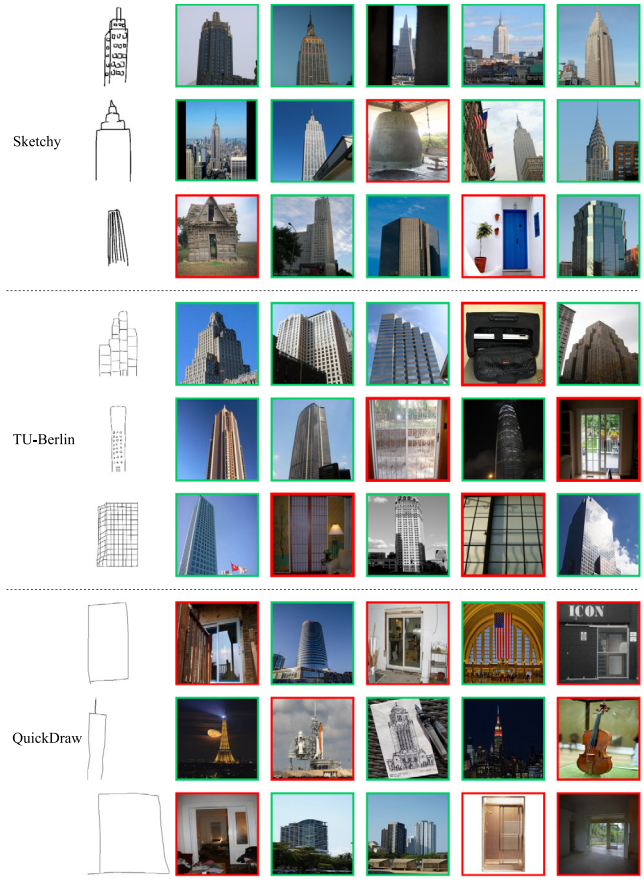

Fig. 6. Top-5 images from the same category (*i.e.* skyscraper) retrieved from three different datasets (*i.e.* Sketchy, TU-Berlin, and QuickDraw). The green (*resp.*, red) border indicates the correct (*resp.*, incorrect) retrieval results.

select three sketches to retrieve their corresponding images. From the visualization, we can find that the Sketchy dataset has the best quality of sketch, whereas the QuickDraw owns the worst one, which make the retrieval performance in Sketchy the highest. Further, some images in QuickDraw have been wrongly labeled. For example, in the first row, forth column of "QuickDraw", the flag in the air is labeled as "skyscraper", which may also affect the retrieval evaluation and performance.

5. Conclusion

In this work, we have studied the problem zero-shot sketch-based image retrieval (ZS-SBIR) from a new viewpoint, *i.e.*, using asymmetric disentangled representation to facilitate structure-aware retrieval. We have proposed our STRucture-aware Asymmetric Disentanglement (STRAD) model, with retrieval performed in combination of three complementary spaces. Comprehensive experiments on three large-scale benchmark datasets have demonstrated the generalization ability of our model from seen categories to unseen ones.

CRediT authorship contribution statement

Jiangtong Li: Conceptualization, Methodology, Writing – original draft. **Zhixin Ling:** Experiment. **Li Niu:** Writing – review & editing. **Liqing Zhang:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work was supported by the National Science Foundation of China (62076162), and the Shanghai Municipal Science and Technology Major Project, China (2021SHZDZX0102, 20511100300). We thank Wu Wen Jun Honorary Doctoral Scholarship, AI Institute, Shanghai Jiao Tong University, China.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cviu.2022.103412.

References

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P., 2016. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In: Proceedings of the Advances in Neural Information Processing Systems. NeurIPS, pp. 2172–2180.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 248–255.

Deng, C., Xu, X., Wang, H., Yang, M., Tao, D., 2020. Progressive cross-modal semantic network for zero-shot sketch-based image retrieval. IEEE Trans. Image Process. 29, 8892–8902.

Dey, S., Riba, P., Dutta, A., Llados, J., Song, Y.-Z., 2019. Doodle to search: Practical zero-shot sketch-based image retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 2179–2188.

Dupont, E., 2018. Learning disentangled joint continuous and discrete representations. In: Proceedings of the Advances in Neural Information Processing Systems. NeurIPS, pp. 710–720.

Dutta, A., Akata, Z., 2019. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 5089–5098.

Dutta, T., Singh, A., Biswas, S., 2020. StyleGuide: Zero-shot sketch-based image retrieval using style-guided image generation. IEEE Trans. Multimed. 1. http://dx.doi.org/10.1109/TMM.2020.3017918.

Eitz, M., Hays, J., Alexa, M., 2012. How do humans sketch objects? ACM Trans. Graph. 31 (4), 1–10.

Fellbaum, C., 2012. WordNet. Encycl Appl Linguist.

Gonzalez-Garcia, A., van de Weijer, J., Bengio, Y., 2018. Image-to-image translation for cross-domain disentanglement. In: Proceedings of the Advances in Neural Information Processing Systems. NeurIPS, pp. 1287–1298.

Hadad, N., Wolf, L., Shahar, M., 2018. A two-step disentanglement method. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 772–780.

Harsh Jha, A., Anand, S., Singh, M., Veeravasarapu, V., 2018. Disentangling factors of variation with cycle-consistent variational auto-encoders. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 805–820.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. CVPR, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR.5967–5976,

Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M., 2014. Semi-supervised learning with deep generative models. In: Proceedings of the Advances in Neural Information Processing Systems. NeurIPS, pp. 3581–3589.

Kingma, D.P., Welling, M., 2014. Auto-encoding variational bayes. In: Proceedings of the International Conference on Learning Representations. ICLR.

Kodirov, E., Xiang, T., Gong, S., 2017. Semantic autoencoder for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3174–3183.

Lee, H.-Y., Tseng, H.-Y., Huang, J.-B., Singh, M., Yang, M.-H., 2018. Diverse image-to-image translation via disentangled representations. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 35–51.

Liu, L., Shen, F., Shen, Y., Liu, X., Shao, L., 2017. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 2862–2871.

Liu, Y., Wang, Z., Jin, H., Wassell, I., 2018a. Multi-task adversarial network for disentangled feature learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3743–3751.

Liu, Y., Wei, F., Shao, J., Sheng, L., Yan, J., Wang, X., 2018b. Exploring disentangled feature representation beyond face identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 2080–2089.

Liu, Q., Xie, L., Wang, H., Yuille, A.L., 2019. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV, pp. 3662–3671.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In: Proceedings of the Advances in Neural Information Processing Systems. NeurIPS, pp. 3111–3119.

Qi, Y., Song, Y.-Z., Zhang, H., Liu, J., 2016. Sketch-based image retrieval via siamese convolutional neural network. In: Proceedings of the IEEE International Conference on Image Processing. ICIP, pp. 2460–2464.

Romera-Paredes, B., Torr, P., 2015. An embarrassingly simple approach to zero-shot learning. In: Proceedings of the International Conference on Machine Learning. ICML, pp. 2152–2161.

Sangkloy, P., Burnell, N., Ham, C., Hays, J., 2016. The sketchy database: Learning to retrieve badly drawn bunnies. ACM Trans. Graph. 35 (4), 119.

Shen, Y., Liu, L., Shen, F., Shao, L., 2018. Zero-shot sketch-image hashing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3598–3607.

Shukla, A., Bhagat, S., Uppal, S., Anand, S., Turaga, P., 2019. Product of orthogonal spheres parameterization for disentangled representation learning. In: Proceedings of the British Machine Vision Conference. BMVC.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations. ICLR.

Socher, R., Ganjoo, M., Manning, C.D., Ng, A., 2013. Zero-shot learning through cross-modal transfer. In: Proceedings of the Advances in Neural Information Processing Systems. NeurIPS, pp. 935–943.

Xu, X., Deng, C., Yang, M., Wang, H., 2020. Progressive domain-independent feature decomposition network for zero-shot sketch-based image retrieval. In: Proceedings of the International Joint Conference of Artificial Intelligence. IJCAI.

Xu, X., Wang, H., Li, L., Deng, C., 2019. Semantic adversarial network for zero-shot sketch-based image retrieval. arXiv preprint arXiv:1905.02327.

Yang, S., Liu, J., Wang, W., Guo, Z., 2019. TET-GAN: Text effects transfer via stylization and destylization. In: Proceedings of the AAAI Conference on Artificial Intelligence. AAAI, pp. 1238–1245.

Yelamarthi, S.K., Reddy, S.K., Mishra, A., Mittal, A., 2018. A zero-shot framework for sketch based image retrieval. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 316–333.

Yu, Q., Liu, F., Song, Y.-Z., Xiang, T., Hospedales, T.M., Loy, C.-C., 2016. Sketch me that shoe. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 799–807.

Yu, Q., Yang, Y., Liu, F., Song, Y.-Z., Xiang, T., Hospedales, T.M., 2017. Sketch-a-net: A deep neural network that beats humans. Int. J. Comput. Vis. (IJCV) 122 (3), 411–425.

Zhang, Z., Saligrama, V., 2015. Zero-shot learning via semantic similarity embedding. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV, pp. 4166–4174.

Zhang, Z., Zhang, Y., Feng, R., Zhang, T., Fan, W., 2020. Zero-shot sketch-based image retrieval via graph convolution network.. In: Proceedings of the AAAI Conference on Artificial Intelligence. AAAI, pp. 12943–12950.

Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., Kautz, J., 2019. Joint discriminative and generative learning for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 2138–2147.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 2223–2232.