Pattern Recognition 126 (2022) 108528

Contents lists available at ScienceDirect

Pattern Recognition

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

An efficient framework for zero-shot sketch-based image retrieval

Osman Tursun*, Simon Denman, Sridha Sridharan, Ethan Goan, Clinton Fookes

Signal Processing, Artificial Intelligence and Vision Technologies (SAIVT), Queensland University of Technology, Australia

ARTICLE INFO

Article history:
Received 26 January 2021
Revised 14 November 2021
Accepted 9 January 2022
Available online 21 January 2022

Keywords:
Sketch-based image retrieval
Zero-shot learning
Knowledge distillation
Similarity learning

ABSTRACT

Zero-shot sketch-based image retrieval (ZS-SBIR) has recently attracted the attention of the computer vision community due to its real-world applications, and the more realistic and challenging setting that it presents over SBIR. ZS-SBIR inherits the main challenges of multiple computer vision problems including content-based Image Retrieval (CBIR), zero-shot learning and domain adaptation. The majority of previous studies using deep neural networks have achieved improved results by either projecting sketch and images into a common low-dimensional space, or transferring knowledge from seen to unseen classes. However, those approaches are trained with complex frameworks composed of multiple deep convolutional neural networks (CNNs) and are dependent on category-level word labels. This increases the requirements for training resources and datasets. In comparison, we propose a simple and efficient framework that does not require high computational training resources, and learns the semantic embedding space from a vision model rather than a language model, as is done by related studies. Furthermore, at training and inference stages our method only uses a single CNN. In this work, a pre-trained ImageNet CNN (*i.e.*, ResNet50) is fine-tuned with three proposed learning objects: *domain-balanced quadruplet loss, semantic classification loss*, and *semantic knowledge preservation loss*. The *domain-balanced quadruplet* and *semantic classification* losses are introduced to learn discriminative, semantic and domain invariant features by considering ZS-SBIR as an object detection and verification problem. To preserve semantic knowledge learned with ImageNet and exploit it for unseen categories, the *semantic knowledge preservation loss* is proposed. To reduce computational cost and increase the accuracy of the semantic knowledge distillation process, ground-truth semantic knowledge is prepared in a class-oriented fashion prior to training. Extensive experiments are conducted on three challenging ZS-SBIR datasets: Sketchy Extended, TU-Berlin Extended and QuickDraw Extended. The proposed method achieves state-of-the-art results, and outperforms the majority of related works by a substantial margin.

1. Introduction

Searching for images using an image query has increased in popularity as content-based image retrieval (CBIR) techniques have improved in recent years. However, the majority of CBIR research has considered the scenario where both query and gallery images are real photos (*i.e.*scenes [1], faces [2]) or digital images (*i.e.*logo [3]). With the widespread popularity of touch-screen devices, free-hand sketch-based image retrieval (SBIR) tasks have drawn the attention of the computer vision (CV) community as sketches are a convenient, universal, easy and fast method for image description [4–9]. Figure 1 shows examples of SBIR results.

The domain-gap and information-gap between sketch and photo domains presents a challenge for existing CBIR approaches. Sketches contain sparse and abstract information, while photos carry dense and precise information. Deep learning models have been applied to reduce these gaps, either in the latent space or the pixel space. A variety of complex architectures such as multiple independent networks [4,10–12], Semi-Heterogeneous networks [13,14], generative adversarial networks (GAN) [15–18] and networks with domain-invariant layers [5,12] have been proposed to address the domain gap. Although these approaches have shown substantial improvements over hand-crafted features, the increased model complexity requires extra resources during training and inference processes.

Model complexity has increased to improve SBIR performance under zero-shot settings, where testing queries are unseen during training. Related studies found existing SBIR models tend to fail under zero-shot settings [12]. One approach to tackle this problem is to learn a joint embedding space or mapping between the vi-

* Corresponding author.
E-mail addresses: osman.tursun@qut.edu.au (O. Tursun), s.denman@qut.edu.au (S. Denman), s.sridharan@qut.edu.au (S. Sridharan), e.goan@hdr.qut.edu.au (E. Goan), c.fookes@qut.edu.au (C. Fookes).

https://doi.org/10.1016/j.patcog.2022.108528

Fig. 1. Examples of SBIR results. In each row, the figure in the left is query while others are retrieved results.

sual representation and class semantic representation. To achieve this, language models [4,19–21] are often used to extract semantic class embeddings, while auxiliary networks such as auto-encoders [4,19,20], GANs [20,22,23] or graph convolutional neural networks (GNN) [12,21] are trained to learn the joint representation or mapping.

The high training cost is not the only drawback of the aforementioned methods. They require all classes of the training set to have descriptive text labels suitable for use by a language model. However, in practical applications classes may only be labelled with a simple numeric label, or out-of-dictionary word labels, which are not suitable for a language model.

This paper aims to tackle ZS-SBIR with a simple, efficient, and language model-free framework. The recent state-of-the-art (SOTA) work, SAKE [5], is a concise and simple framework. It's feature extraction encoder is a single-stream Convolutional Neural network (CNN) that ensures an efficient and simple inference process. However, during training it also requires a language model and another ImageNet pre-trained CNN to generate valid teacher signals for knowledge distillation. They argue preserving knowledge learned from ImageNet via knowledge distillation is beneficial for ZS-SBIR. Although we also find that rich features learned from ImageNet are essential for ZS-SBIR, we find a language model and online teacher network are not necessary. SAKE generates a teacher signal for each input item, either from the sketch or photo domain. It therefore requires a language model to align teacher signals that are otherwise invalid due to domain-shift. The alignment is based on the semantic similarity matrix of ImageNet labels and target dataset labels, which is constructed using WordNet [24]. In comparison, we generate teacher signals for each class by averaging the activations of a pre-trained ImageNet with images from the photo domain, whose distribution is close to ImageNet. We therefore do not require ground-truth text-labels that are interpretable by a language model, and the teacher signal generation is a one-time offline process.

We also find SAKE treats the ZS-SBIR problem as an object identification task, where the learning objective is a categorical classification loss, while other related works [4,23] consider the problem a verification task, where metric learning *i.e.*, triplet loss, is applied. In this work, we not only unify these two objectives in a single framework, but also propose a domain-balanced quadruplet loss for metric learning.

We have tested the proposed method on two popular SBIR datasets (Sketchy Extended [25] and TU-Berlin Extended [26]) and a newly proposed challenging SBIR dataset, QuickDraw Extended [4]. In all benchmarks, we have achieved state-of-the-art (SOTA) performance by only fine-tuning a ResNet50 [27] model with our three proposed learning objectives: *Domain-balanced Quadruplet, Semantic Classification* and *Semantic Knowledge Preservation* losses.

The main contributions of this study are as follows:

- We propose a simple and efficient framework for zero-shot sketch-based image retrieval. A single feature extractor with a ResNet-50 backbone is trained with three introduced learning objectives: domain-balanced quadruplet loss, semantic classification loss and semantic knowledge preservation loss.
- We minimise the cost and restrictions that come with utilising a semantic knowledge distillation process.
- The proposed method achieves state-of-the-art results on three challenging zero-shot sketch-based image retrieval datasets: Sketchy Extended, TU-Berlin Extended and QuickDraw Extended.

The remainder of the paper is organized as follows. Section 2 presents a literature review where we discuss related ZS-SBIR studies. In Section 3, the proposed model and learning objectives are introduced. Section 4 outlines experiment setups, results and related discussions; and finally Section 5 concludes the paper.

2. Related work

In this section, we primarily discuss general sketch-based image retrieval and zero-shot sketch-based image retrieval. However, there are other closely related topics to these such as fine-grained sketch-based image retrieval [8,28], sketch-based face retrieval [29–31] and sketch-based 3D shape retrieval [29,32] and interested readers are referred to the relevant citations.

Early SBIR studies mainly focus on the challenges raised by the large domain gap between the sketch and photo domain. Both hand-crafted features and deep features have been explored. Hand-crafted features include edge/shape-based features [33–35] with a bag-of-words representation, as in some aspects the strong edges in a photo correspond to the contours of sketches. On the other hand, deep features seek to learn a joint representation the of sketch and photo domain through metric learning [14,28,36,37], style-content disentangled representations, [6,22] and style-transfer [10,15,18,38]. However, related studies discover that the accuracy of these models decreases in a real-life and challenging scenarios where either the queries or gallery images are unseen. To tackle this problem, zero-shot SBIR approaches [4,5,12,19–21] have been proposed.

The majority of ZS-SBIR approaches leverage semantic information embedded in seen data (*i.e.*word labels) to learn a generalised representation for both seen and unseen categories. The main difference between these methods lies in the architecture of the mapping network and the embedding method in the semantic space. GNNs [12,21], Multi Layer Perceptrons (MLP) [4,39] and GANs [19,20,40,41] have all be used for the mapping network. On the other hand, word2vec [4,19–21,39–41] and hierarchical models [19,20,40] are common embedding methods used to construct the semantic space.

In comparison, the recent work, SAKE [5], uses a visual semantic representation learned from ImageNet. However, to avoid an incorrect representation caused by domain-shift, SAKE aligns the visual semantic representation with a semantic similarity matrix constructed using WordNet. We propose an alternative method to extract a visual-semantic representation that is free from alignment. As such, our method does not require a language model, making our proposed approach one of a very small number of ZS-SBIR studies that do not require a language model. Other language model free methods are typically generative approaches based-on GANs [6], or variational auto-encoder encoders (VAE) [42].

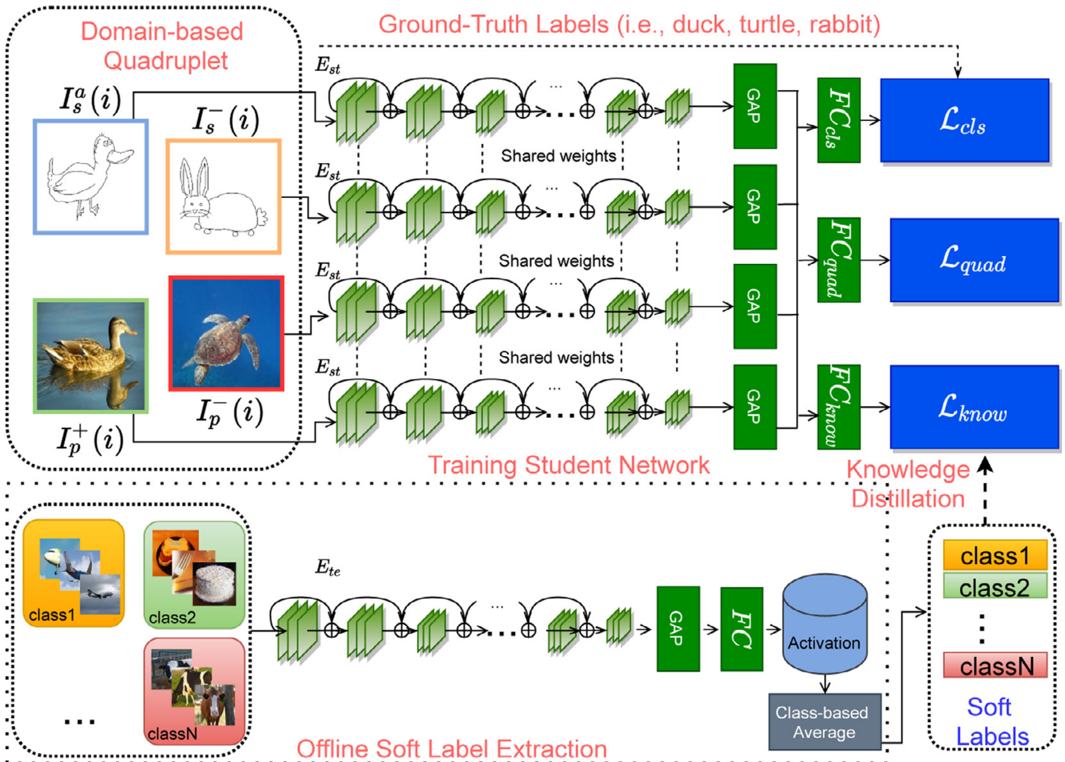


Fig. 2. A diagram of the proposed method. The approach includes an *Online Training Student Network* and an *Offline Soft Label Extraction with Teacher Network*. The student (E_{st}) and teacher (E_{te}) networks use ResNet-50 as the backbone. A quadruplet composed of two images from the sketch domain and two images from the photo domain is the input of the student network, while the inputs to the teacher network are only from the photo domain. The teacher network generates soft labels to prevent the student network from forgetting previously learned knowledge from ImageNet, and the soft label extraction is a one-time offline process.

3. The proposed method

In this section, we present our proposed method for zero-shot sketch-based image retrieval (ZS-SBIR). In the following subsections, we first outline the overall structure of the proposed method, then explain network structures, learning objectives and discuss implementation details.

3.1. Network architecture

The objective of the proposed method is to learn discriminative and domain-invariant CNN encoders that map semantically similar images from the sketch and photo domains into the same region of a common embedding space. An overall diagram of the proposed approach is shown in Fig. 2. The diagram is composed of two parts: the *Online Training Student Network* and the *Offline Soft Label Extraction with Teacher Network*. It includes two networks: the student network, E_{st} and teacher network, E_{te}. E_{st} requires training, while E_{te} does not. E_{te} generates ground-truth labels for one of the learning objectives with which E_{st} is trained. Both E_{st} and E_{te} use ResNet-50 [27] as the backbone. However, for E_{st}, we replace the fully-connected layer of ResNet-50 with three new fully connected layers, FC_{cls}, FC_{quad} and FC_{know}, that correspond to three proposed learning objectives described in Section 3.2.

During the training stage, for an input image I, E_{st} outputs three different activations from it's three fully connected layers. The L2-normalised output of the layer FC_{quad} is used in both in the inference and training stage. The output of two other fully connected layers are only used in the training stage. The notation f represents the L2-normalised output of the feature extraction layer. The layer FC_{quad} is the feature extraction layer and f_{quad} represents its L2-normalised output. Notations ϕ_{cls} and ϕ_{know} represent the softmax activations of the layers FC_{cls} and FC_{know}.

One of the learning objective is a modified quadruplet loss [43]. As such, during training E_{st} operates as a four-stream network where all streams share the same weights. We also trialled the inclusion of an attention block [44] before the last fully connected layer in E_{st} as per [4], although we observed that it didn't result in any improvements during ablation studies.

3.2. Learning objectives

To learn a discriminative and domain-invariant encoder with general semantic knowledge, we introduce the following learning objectives: *Domain-balanced Quadruplet Loss*, $\mathcal{L}_{quad}$, *Classification Loss*, $\mathcal{L}_{cls}$, and *Knowledge Preservation loss*, $\mathcal{L}_{know}$. Each loss is calculated using the output of the corresponding fully connected layer as shown in Fig. 2.

The **Domain-balanced Quadruplet Loss** is a modified version of the triplet loss [2], which has been widely used to maximise the inter-class distance and minimise the intra-class distance in the embedding space for various image retrieval tasks [1,43]. Here, our objective to also minimise the distance between sketches and photos from the same semantic category, while maximising the distance between sketches and photos from different categories in the target embedding space.

With the triplet loss, this inter-class and intra-class distance relationship is formulated with a triplet (T) where a anchor sketch (I_s^a) and a positive photo (I_p^+) are selected from the same category, while a negative photo (I_p^-) is selected from a different category. For example, $T = \{I_s^a, I_p^+, I_p^-\}$ is a triplet where $l(I_s^a) = l(I_p^+)$ and $l(I_s^a) \neq l(I_p^-)$ (notation l represents the ground-truth label). The Euclidean distance between the anchor sketch and the positive (which belong to the same class) photo image is $\delta_{sp}^+ = ||f(I_s^a) - f(I_p^+)||_2^2$, while the Euclidean distance between the anchor sketch and the negative (different class) photo image is $\delta_{sp}^{-} = ||f(I_s^a) - f(I_p^-)||_2^2$. δ_{sp}^{-} should be larger than δ_{sp}^{+} by a margin α, which is set to 0.2 as in [2]. The triplet loss for a batch of N triplets is defined as

$$\mathcal{L}_{trip} = \frac{1}{N}\sum_{i=1}^{N} max(\delta_{sp}^{+}(i) - \delta_{sp}^{-}(i) + \alpha, 0), \tag{1}$$

where i represents ith triplet.

The proposed domain-balanced quadruplet loss deploys an extra negative sketch image, I_s^-, such that the quadruplet is defined as $Q = \left\{I_s^a, I_p^+, I_p^-, I_s^-\right\}$. The additional image is used to calculate the Euclidean distance between the anchor sketch and the additional negative (different class) sketch image, $\delta_{ss}^{-} = ||f(I_s^a) - f(I_s^-)||_2^2$. Therefore, the proposed loss is

$$\mathcal{L}_{quad} = \frac{1}{2N}\sum_{i=1}^{N}(max(\delta_{sp}^{+}(i) - \delta_{sp}^{-}(i) + \alpha, 0) + max(\delta_{sp}^{+}(i) - \delta_{ss}^{-}(i) + \alpha, 0)). \tag{2}$$

We proposed the domain-balanced quadruplet loss for the following reasons:

1. In this work, the input data is from two different domains: sketch and photo; nevertheless only one encoder is trained with multiple losses. As a result, during training with the triplet loss domain imbalance will occur as the input will include more samples from one domain than the other. This can bias the model towards the domain with more samples. For example, in SBIR studies, a triplet is usually composed of one sketch and two photos, thus the model will favour the photo domain over the sketch domain. The chance of domain imbalance will be higher when fine-tuning a model which was pre-trained on the photo domain, such as ImageNet. In comparison, a domain-balanced quadruplet includes an equal number of samples from both the sketch and photo domains. It therefore reduces the chance of domain imbalance.
2. Related studies [43,45] demonstrate that an extra negative image is beneficial for learning discriminative features. However, these studies do not consider domain differences and domain imbalance in their formulations. If the quadruplet loss is used with three photo images, this will increase the chance of domain imbalance as each quadruplet will include only one sketch image, yet three photo images. The existence of this problem is supported by the ablation study where a quadruplet loss with three photo images is shown to be inferior to the triplet loss and the domain-balanced quadruplet loss.

It is important to note that the domain-balanced quadruplet has benefits beyond the metric learning component of the framework. For example, having an even number of samples from each class is equally important for the Semantic Classification Loss which seeks to ensure that meaningful semantic information is extracted from images.

Semantic Classification Loss is introduced to ensure hidden features extracted with E_{st} are composed of signals that are sufficient for identifying the semantic classes of inputs from both the sketch and photo domains. Additionally, with this semantic loss, E_{st} implicitly learns to minimise the intra-class distance.

Specifically, a soft-max cross-entropy loss is utilised. As Eq. (3) shows, every input to E_{st} is a quadruplet, Q, that includes two images from the sketch domain and two images from the photo domain. This even domain sampling ensures domain balance. The output from E_{st} will be sent to FC_{cls} for softmax calculation. Here, we use notation ϕ_{cls} to represent this whole process,

$$\mathcal{L}_{cls} = -\frac{1}{4N}\sum_{i=1}^{N}\sum_{I\in Q} p(l(I_i)) \cdot \log \phi_{cls}(I_i), \tag{3}$$

where p represents the one-hot encoding of the class label l.

Semantic Knowledge Preservation Loss Transfer learning plays a key role in SBIR tasks. Networks pre-trained on ImageNet have been fine-tuned for ZS-SBIR problems in previous works [4–6,12]. However, Liu et al. [5] claim fine-tuning will cause catastrophic forgetting that decreases the ability of the fine-tuned network to adapt back to the original domain. To prevent a network from forgetting previously learned knowledge, Liu et al. generates a teacher signal for each of the training inputs for knowledge distillation. However, this requires extra computational resources during training as inputs are also sent to a teacher network to generate the teacher signals. Moreover, their method also requires a language model for error alignments. Inspired by Liu et al., we implement an efficient knowledge distillation approach which does not require a language model. As shown in Fig. 2, we only use a class-based teacher signal rather than item-based teacher signals. The teacher signals are the softmax of the average activation of the teacher network E_{te} for each semantic class. The teacher signals can be considered as soft labels. The notation $q(l(I_i))$ represents the soft label of the ith image, I_i. To reduce the errors caused by domain shifts, we calculate q with the softmax of the average activation of each class that exists in the photo domain as shown in Fig. 2. As these soft labels are only calculated once in an offline process, the method is highly efficient. We use the cross-entropy loss with soft labels for calculation of the Knowledge loss $\mathcal{L}_{know}$

$$\mathcal{L}_{know} = -\frac{1}{4N}\sum_{i=1}^{N}\sum_{I\in Q} q(l(I_i)) \cdot \log \phi_{know}(I_i) \tag{4}$$

In summary, E_{st} is trained using the $\mathcal{L}$ in Eq. (5), which is a combination of the three proposed objectives. For simplicity, in the default setting the weights of each objective are set to 1.

$$\mathcal{L} = \mathcal{L}_{know} + \mathcal{L}_{cls} + \mathcal{L}_{quad} \tag{5}$$

3.3. Implementation details

PyTorch [46] is used as our implementation framework, and all models are trained with single GTX 1080Ti GPU. We select an ImageNet pretrained ResNet-50 as the backbone for both teacher and student networks. We applied the SGD optimiser with momentum=0.9 and decay=5×10^{-4}. The batch size is 16, but it includes 64 images as each input is a quadruplet. The initial learning rate $\lambda = 1 \times 10^{-4}$, and it decays by a factor of 10 after every ten epochs. We trained all models for up to 25 epochs, which is smaller than what previous works [4,5,22] require, as our model starts to converges after only a few training epochs. The size of each epoch is equal to the number of sketches in the training set. We also use early-stopping based on the mean average precision score (MAP@all) on the validation set (5% of the seen classes is used as validation data). If the model's validation accuracy has not shown improvements for 5 epochs, the model will stop training. The model with the best validation MAP@all score is used for the final evaluation.

4. Experiments

4.1. Datasets

We evaluated our method on well-known large-scale SBIR datasets: *Sketchy Extended, TU-Berlin Extended* and *QuickDraw Extended*. An overall comparison of these datasets is given in Table 1. As shown in Table 1, each dataset is split into *seen* and *unseen* classes. All unseen classes are used for testing, while seen classes are further split into training and validation classes. 95% of seen classes form the training data while the remainder is used for validation.

Table 1
Comparison of public SBIR datasets. These datasets include images from the sketch and photo domains. For zero-shot studies, they are split to train (seen) and test (unseen) classes.

	Sketchy Ext. [14]		TU-Berlin Ext. [10]	QuickDraw Ext. [4]
# Sketch/Class	604 ± 61		80 ± 0	3022 ± 216
# Image/Class	584 ± 76		787 ± 489	1853 ± 308
Name	SK-SH [12]	SK-YE [42]	TUB-SH [12]	QD-DE [4]
Type	Random	ImageNet orthogonal	Random	ImageNet orthogonal
# Seen Classes	100	104	220	80
# Unseen Classes	25	21	30	30

Sketchy Extended is an extended version of the Sketchy dataset [25] by Liu et al. [14]. The Sketchy dataset has 125 categories. Each category is composed of 100 natural images and at least 600 sketches. It's photo domain is extended by adding an extra 60,502 natural images collected from ImageNet. The extended version has an average of 604 sketches and 584 images in each class. It is a balanced dataset as the variance between the number of items in each class is relatively small. To adapt this dataset for zero-shot studies, the dataset is partitioned into seen and unseen sets. There exists two partition protocols in the literature. For brevity, we refer to them as *SK-SH* and *SK-YE*. SK-SH is proposed by Shen et al. [12], who creates an unseen set by randomly selecting 25 classes, and the remaining 100 classes are seen classes. However, some of those randomly selected classes may have already been seen by networks initialised with pre-trained ImageNet weights, and thus this violates the zero-shot setting. SK-YE, introduced by Yelamarthi et al. [42], carefully selects 21 unseen classes that are not present in ImageNet.

TU-Berlin Extended includes 20,000 sketches from the TU-Berlin dataset [26] and an extra 204,489 real images collected by Liu et al. [14]. Its sketch-domain has a uniform class distribution but with only 80 items, while the photo-domain has around 787 items. This is a highly imbalanced and challenging dataset. The partition protocol introduce by Shen et al. [12] is used for creating zero-shot training and testing sets. We refer to this protocol as *TU-SH* where 30 randomly picked classes that include at least 400 photo images are seen classes, and other classes are unseen classes.

QuickDraw Extended is a challenging dataset created by Dey et al. [4]. Compared to the Sketchy Extended and TU-Berlin Extended datasets, it includes more sketches (average of 3022/class) and photos (average of 1853/class). All sketches are drawn by amateurs, so and can thus be very abstract and highly varied. Moreover, all classes are carefully selected to avoid ambiguity and overlap. A partition following a similar protocol to that proposed by Yelamarthi et al. [42] is provided. We named this partition *QD-DE*. With this partition, the dataset is split into 80 seen and 30 unseen classes.

4.2. Evaluation metrics

Precision (P) and mean average precision (mAP) are two main metrics for evaluating the ranked retrieval results for testing queries in related SBIR studies. Precision is calculated for the top k (*i.e.*, 100, 200) ranked results, and mAP values are calculated for the top k or all ranked results. The precision at index k (P@k) is the ratio between the total number of documents and number of relevant documents in the k retrieved results. P@k is also used for calculating AP values of each query as follows:

$$AP@k = \sum_{i=1}^{k} \frac{P@i \times \gamma(i)}{N}, \tag{6}$$

where N is total number of relevant documents and $\gamma(i)$ is 1 if the ith ranked result is relevant, otherwise 0. mAP@k is the mean AP@k of all queries.

4.3. State-of-the-art comparison

We have compared the proposed methods with SOTA methods on ZS-SBIR and its generalised version (search space includes *seen* and *unseen* categories [20]). The results of the ZS-SBIR task on the Sketchy Extended and TU-Berlin Extended datasets are shown in Table 2, and the ZS-SBIR results on the QuickDraw Extended datasets are listed in Table 4. The results of the generalised ZS-SBIR task on the Sketchy Extended and TU-Berlin Extended datasets are shown in Table 3.

In all these experiments, with the default setting, the proposed method shown improvements compared to SOTA methods. The proposed method surpassed other methods that have not utilized a language model by a substantial margin. We also compare the embedding feature sizes used by the methods. We reported results with feature sizes 512 and 64. We achieved the best result with a feature size, 512, which we note is relatively small compared to many other methods, and equal to the feature size of the previous best SOTA method, SAKE [5]. Our results with feature size 64 outperform most of the other methods with large or the same feature dimension. We also mapped the features to binary hash codes as [20] by using the iterative quantisation (ITQ) [49] for the sake of comparison with methods with binary hash codes. Though an apparent decrease is found in the performance after binary hash coding, our results with 64-bit binary hash codes achieve better performance compared to previous SOTA results with binary hash coding.

4.4. Qualitative results

We visualize top-10 and top-5 results of some success and failure cases in Figs. 5 and 6. Figure 5 displays both ZS-SBIR and GZS-SBIR results from Sketchy Ext. and TU-Berlin Ext. datasets. Figure 6 shows top-10 results from the QuickDraw Ext. dataset. The proposed method returns perfect results when the given query is unambiguous, and returns acceptable false-positive results when the query is unclear. Unclear queries are more likely to occur when the search space is as large as in generalised ZS-SBIR cases.

4.5. Ablation studies

Learning Objectives Here, we investigate the impact of each proposed learning objective and attention (att.) [44] on the proposed approach using the TU-Berlin Extended and Sketchy Extended Datasets. As shown in Table 6, we provide results of models trained with several combinations of these losses. The model trained with the triplet loss is the baseline, and it's results show that it is a challenging baseline which outperforms most of the

Table 2
A performance comparison of recent state-of-the-art ZS-SBIR methods.

Method	Backbone	DIM.	Sketchy Ext. (Split: SK-SH)		Sketchy Ext. (Split: SK-YE)		TUBerlin Ext. (Split: TU-SH)	
			mAP @all	P @100	mAP @200	P @200	mAP @all	P @100
ZSIH [12]	2 × AlexNet	64*	25.8	34.2	-	-	22.3	29.4
EMS [47]	CSE-ResNetXt-101	64*	-	-	-	-	16.5	25.2
		512	-	-	-	-	25.9	36.9
CVAE [42]	2 × VGG-16	4096	19.6	28.4	22.5	33.3	-	-
GZS-SBIR [16]	2 × ResNet-152	2048	28.9	35.8	-	-	23.8	33.4
SEM-PCYC [20]	2 × VGG-16	64	34.9	46.3	-	-	29.7	42.6
		64*	34.4	39.9	-	-	29.3	39.2
Doodle2Search [4]	2 × VGG-16	4096	-	-	46.1[a]	37.0	10.9	-
SketchGCN [21]	2 × ResNet-50	2048	-	-	56.8[a]	48.7	32.4	50.5
Style-guide [6]	2 × VGG-16	4096	37.6	48.4	35.8	40.0	25.4	35.6
SAKE [5]	CSE-ResNet-50	64*	36.4	48.7	35.6	47.7	35.9	48.1
		512	54.7	69.2	49.7	**59.8**	47.5	59.9
STRAD [48]	VGG-16	-	-	-	41.3	53.7	-	-
OCEAN [19]	2 × VGG-16	64	46.2	59.0	-	-	33.3	46.7
PCMSN [40]	2 × VGG-16	64	52.3	61.6	-	-	42.4	51.7
		64*	50.6	61.5	-	-	35.5	45.2
		64	51.0	64.2	46.2	56.1	44.7	57.2
SBTKNet	ResNet-50	64*	47.4	61.0	41.0	51.5	39.4	51.4
		512	**55.3**	**69.8**	**50.2**	59.6	**48.0**	**60.8**

[a] These mAP@200 evaluations use a different formulation to ours. If we follow the same mAP@200 evaluation protocol, our mAP@200 value for Sketchy-Ext. (Split: SK-YE) is 72.24. * represents binary features.

Table 3
Comparison results of generalised ZS-SBIR on Sketchy Extended and TU-Berlin Extended datasets.

Method	Backbone	DIM.	Sketchy Ext. (Split: SK-SH)		TU-Berlin Ext. (Split:TU-SH)	
			mAP@all	P@100	mAP@all	P@100
ZSIH [12]	2 × AlexNet	64	21.9	29.6	14.2	21.8
SEM-PCYC [20]	2 × VGG-16	64	30.7	36.4	19.2	29.8
Style-guide [6]	2 × VGG-16	4096	33.1	38.1	14.9	22.6
SBTKNet	ResNet-50	512	**51.5**	**57.2**	**33.4**	**49.4**

(a) Dolphin (Sketch-ext.) and white shark (ImageNet)

(b) Tree (Sketch-ext.) and park bench (ImageNet)

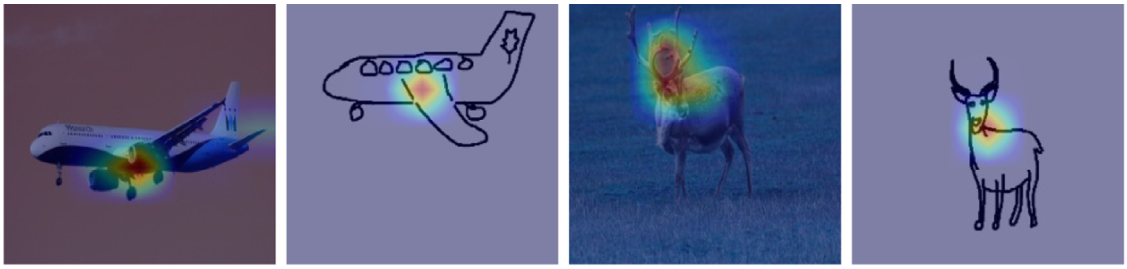
Fig. 3. Examples of classes from the Sktechy-extended and ImageNet datasets.

Table 4
Comparison results of QuickDraw-Extended Dataset.

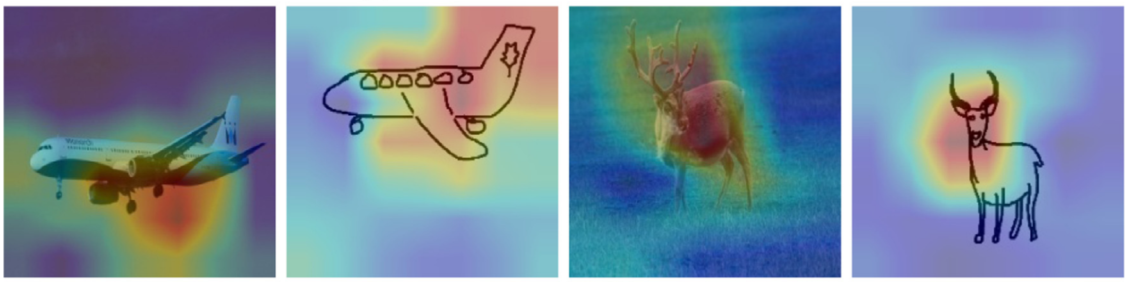
Method	Backbone	DIM.	mAP@all	P@200
CVAE [42]	2 × VGG-16	4096	0.30	0.30
Doodle2Search [4]	2 × VGG-16	4096	7.52	6.75
SBTKNet	ResNet-50	**512**	**11.9**	**16.7**

state-of-the-art method listed in Table 2. Each learning objective improves the results of the baseline. The combination of the domain-balanced quadruplet loss and the semantic classification loss demonstrates significant improvements. The combination of the semantic classification loss and the semantic knowledge preservation loss also achieves results that are comparable with SAKE where a semantic classification loss and a similar but expensive semantic knowledge preservation loss are utilised. Our results could achieve further improvements if a semi-heterogeneous network such as CSE-ResNet-50 is used as the backbone as SAKE.

To show the importance of domain-balance, the proposed domain-balanced quadruplet loss is compared with both the triplet loss and the quadruplet loss. In triplet loss, the number of images from the photo domain is two times the number of the image domain, while for quadruplet loss, this is three times. As shown in Table 6, The domain-balanced quadruplet loss outperforms both the domain-imbalanced quadruplet loss and triplet loss, and the domain-imbalanced quadruplet loss is inferior to the triplet loss.

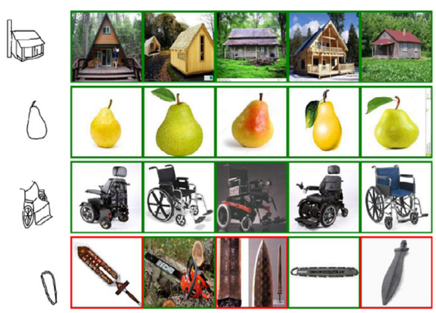


(a) Visualisations of attention layer

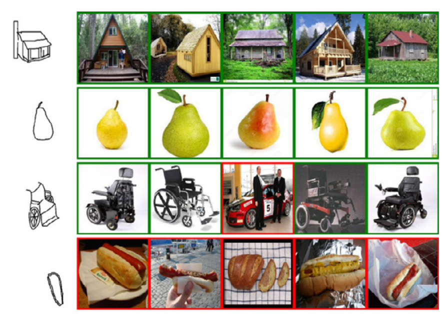


(b) Visualisations of CAM of the last convolutional layer

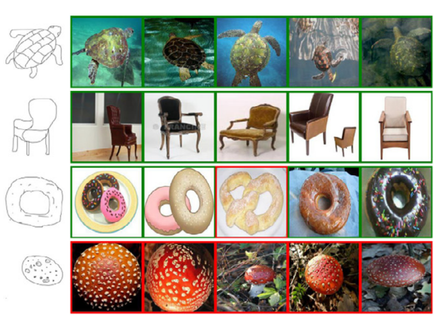
Fig. 4. Visualisation of attention block and the last convolutional layer's activation map.



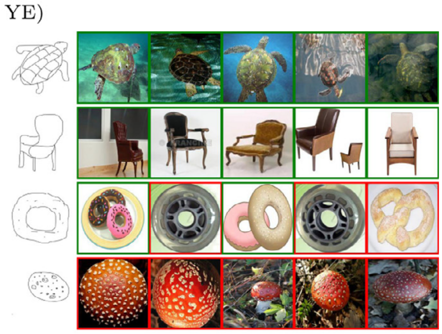
(a) Zeroshot, Sketchy (Split:SK-YE)



(b) Generalized zeroshot, Sketchy (Split:SK-YE)



(c) Zeroshot, TU-Berlin (Split:TUB-SH)



(d) Generalized zeroshot, TU-Berlin (Split:TUB-SH)

Fig. 5. Top-5 ZS-SBIR (a,c) and generalised ZS-SBIR (b,d) results retrieved by our model on Sketchy Ext. (a,b) and TU-Berlin Ext. (c,d) datasets. Correct results are shown with a green border, while false results are shown with a red border. The top two rows are all correct, the third row is partially correct, while the bottom row is all incorrect. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

These results indicate that domain imbalance causes a decrease in performance.

Knowledge Preservation with the Semantic Knowledge Preservation Loss The model trained with the semantic knowledge preservation loss should have an embedding space similar to the embedding space of it's teacher network. To evaluate this, we present a comparison between class labels for samples in the target dataset (*i.e.*Sketch-extended), and the ImageNet labels that correspond to the maximum activation of FC_{know} (i.e. the class to which the sample is assigned) for these samples. In detail, FC_{know} embeddings are extracted for all photo images in the Sketchy-extended dataset. Then the centroid of the embedding space for each class is calculated. The ImageNet text label of the maximum activation of the centre embedding of each class is com-

Table 5
Examples of semantic knowledge preservation using the semantic knowledge preservation loss. Red indicates an exact match, while green and blue represent semantically/visually related, and contextually related matches respectively.

Unseen classes		Seen classes	
Sketchy-ext.	**ImageNet**	**Sketchy-ext.**	**ImageNet**
bat	letter opener	airplane	airliner
cabin	church	alarm clock	analog clock
cow	hog	ant	ant
dolphin	great white shark	ape	howler monkey
door	rocking chair	apple	Granny Smith
giraffe	zebra	armor	chain mail
helicopter	airliner	axe	hatchet
mouse	fox squirrel	banana	banana
pear	Granny Smith	bear	brown bear
raccoon	fox squirrel	bee	bee
rhinoceros	hog	beetle	tiger beetle
saw	hatchet	bell	chime
scissors	letter opener	bench	park bench
seagull	drake	bicycle	bicycle-built-for-two
skyscraper	church	blimp	airship
songbird	macaw	bread	French loaf
sword	letter opener	butterfly	monarch
tree	park bench	camel	Arabian camel
wheelchair	bicycle-built-for-two	candle	candle
windmill	church	cannon	cannon
window	dining table	car (sedan)	racer

pared with its original text label as shown in Table 5. Each class in the *seen classes* column has an exact matching ImageNet label. For *unseen classes*, an exact match is impossible, as these classes do not exist in ImageNet. However, semantically, visually or contextually related outcomes are expected. As a result, for all unseen classes except for the class "bat", the ImageNet label and original label are either semantically or visually close (marked in green) or contextually related (marked in blue). For example, as shown in Fig. 3, "dolphin" and "great white shark" appear similar, while "tree" frequently appears in scenes that contain a "park bench".

Attention Layer We also tested the proposed method with an attention layer before the last convolutional layer as per [4]. However, the attention layer does not bring any improvements to the final results. In Fig. 4, we visualised the attention layer and compared it with the last convolutional layer's activation map (CAM) [50]. It shows that the network, without attention, is able to learn to suppress background information, and thus the attention layer is redundant. When using attention, Fig. 4 shows that the network becomes more selective, and excludes useful information. Alternative attention formulations may overcome this.

Domain-balanced Quadruplet Loss Margin The range of the margin hyper-parameter α of the domain-balanced quadruplet loss is 0 to 1, as the Euclidean distance is calculated with L2-normalised feature vectors. We used 0.2 as the default value for $alpha$ as per [51], nevertheless we also tested other values (0.4, 0.6 and 0.8). As shown in Table 7, the proposed model achieved the best result when α is 0.2. The performance of the model gradually decreases when increasing α.

The Impact of Loss Weights By default, the proposed approach assigns equal weights (each weight is one) to the three proposed losses. To assess model consistency through different weight combinations, we trained and tested the model by changing the weight of each loss. We tested six different weight values: 0, 0.5, 1, 2, 4 and 8. The comparison is given in Table 8. The change in results is small when applying different weight combinations where

Table 6
Ablation study for the proposed approach. The same backbone network trained with the triplet loss is used as baseline.

Quad.	Cls.	Know.	Att.	Sketchy Ext. (Split: SK-YE)				TUBerlin Ext. (Split: TU-SH)			
				mAP @all	mAP @200	P @100	P @200	mAP @all	mAP @200	P @100	P @200
-	-	-	-	47.7	43.6	55.1	52.0	44.8	46.2	56.8	55.0
✓	-	-	-	48.5	44.4	56.0	52.9	46.3	47.5	58.0	56.2
✓	✓	-	-	51.1	48.9	61.7	57.5	46.8	48.1	58.1	56.0
-	✓	✓	-	48.7	47.7	62.6	57.7	44.8	48.1	58.9	56.5
✓	-	✓	-	51.8	49.5	62.7	58.3	47.2	48.7	58.7	56.7
✓	✓	✓	-	52.7	50.2	64.2	59.6	48.0	50.5	60.8	58.6
✓	✓	✓	✓	52.6	50.2	63.9	59.5	48.0	50.5	60.6	58.3
*	-	-	-	45.0	40.7	52.6	49.4	41.4	41.8	52.7	51.1

* Represents quadruplet loss where the quadruplet is composed of one anchor sketch, one positive photo and two negative photos.

Table 7
Comparison of different margin values for domain-balanced quadruplet loss.

Margin	Sketchy Ext. (Split: SK-YE)				TUBerlin Ext. (Split: TU-SH)			
(α)	mAP@all	mAP@200	P@100	P@200	mAP@all	mAP@200	P@100	P@200
0.2	**52.7**	**50.2**	**64.2**	**59.5**	**48.0**	**50.5**	**60.8**	**58.6**
0.4	51.9	48.3	61.7	57.8	46.2	47.5	57.6	55.3
0.6	50.7	46.3	59.6	55.8	45.1	46.1	56.0	53.9
0.8	48.4	42.6	55.5	52.3	42.6	42.2	52.0	50.1

Table 8
Comparison of different weight combinations of learning objectives.

Quad.	Class.	Know.	Sketchy Ext. (Split: SK-YE)				TUBerlin Ext. (Split: TU-SH)			
			mAP @all	mAP @200	P @100	P @200	mAP @all	mAP @200	P @100	P @200
1	1	1	52.7	50.2	64.2	59.6	48.0	50.5	60.8	58.6
0	1	1	48.7	47.7	62.6	57.7	44.8	48.1	58.9	56.5
0.5	1	1	52.6	50.1	64.0	59.5	47.4	50.1	60.3	58.0
2	1	1	52.9	50.1	63.7	59.5	47.3	49.8	60.0	57.8
4	1	1	51.9	48.7	62.3	58.3	47.8	49.5	60.0	57.7
8	1	1	50.1	46.8	60.7	56.6	48.1	49.8	60.3	57.8
1	0	1	51.8	49.5	62.7	58.3	47.2	48.7	58.7	56.7
1	0.5	1	**53.2**	50.3	63.8	59.7	46.9	48.9	59.1	57.0
1	2	1	52.9	50.6	64.6	60.0	48.4	51.3	61.6	59.2
1	4	1	52.7	**50.8**	65.0	**60.2**	**48.8**	**52.5**	**62.8**	**60.3**
1	8	1	51.9	50.0	**64.7**	59.6	48.2	52.1	62.6	60.2
1	1	0	51.1	48.9	61.7	57.5	46.8	48.1	58.1	56.0
1	1	0.5	53.1	**50.8**	64.4	59.9	47.7	50.1	60.3	58.1
1	1	2	52.6	50.2	64.1	59.6	48.0	50.4	60.7	58.3
1	1	4	52.5	49.7	63.9	59.1	48.1	50.7	60.9	58.4
1	1	8	50.9	48.1	62.2	57.9	47.8	50.5	60.7	58.3

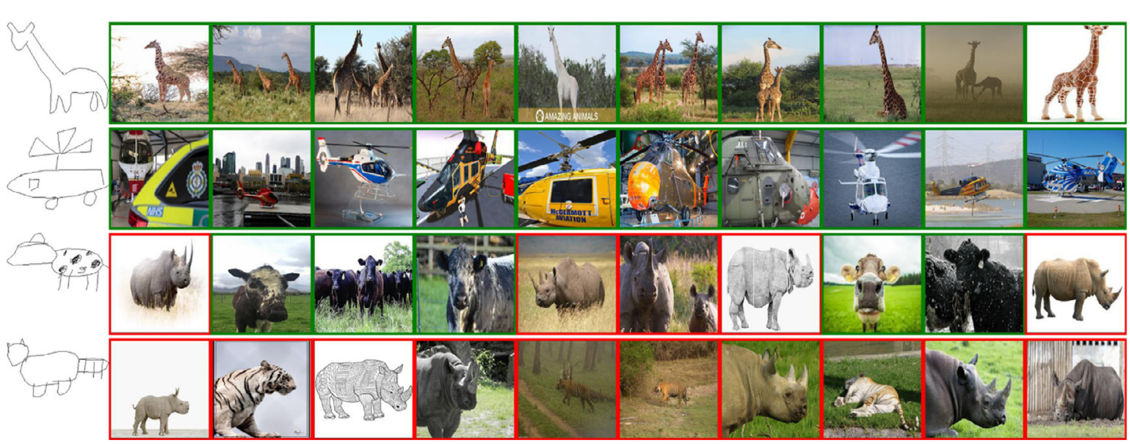


Fig. 6. Top-10 ZS-SBIR results retrieved by the proposed model on the QuickDraw Ext. dataset. Correct results are shown with a green border, while incorrect results are shown with a red border. The top two rows are all correct, the third row is partially correct, while the bottom row is all incorrect. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

all losses have weights larger than zero, though the default setting achieved the best result.

5. Conclusion

In this work, we propose a simple and efficient framework for zero-shot sketch-based image retrieval (ZS-SBIR). The model is trained in an end-to-end fashion with three proposed losses: *domain-balanced quadruplet loss, semantic classification loss* and *semantic knowledge preservation loss*. The domain-balanced quadruplet loss addresses the issue of domain-imbalance that occurs using the vanilla triplet loss that is frequently used to reduce the domain gap and learn a shared low-dimension feature space. In addition, categorical semantic classification is also used to learn semantic features. To enhance the zero-shot ability of the learned model, the semantic knowledge preservation loss is introduced. This loss is formulated to prevent the rich knowledge learned from the ImageNet dataset from being forgotten during fine-tuning of the pre-trained ImageNet model that is used by the network. Experiments on three challenging ZS-SBIR datasets show that the proposed framework is more efficient and effective than related works. Moreover, extensive ablation studies show each introduced loss brings non-trivial improvements, and contributes to the state-of-the-art performance.

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.

References

[1] A. Gordo, J. Almazán, J. Revaud, D. Larlus, Deep image retrieval: learning global representations for image search, in: European Conference on Computer Vision, Springer, 2016, pp. 241–257.
[2] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: a unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.
[3] O. Tursun, S. Denman, S. Sivapalan, S. Sridharan, C. Fookes, S. Mau, Component-based attention for large-scale trademark retrieval, IEEE Trans. Inf. Forensics Secur. (2019).
[4] S. Dey, P. Riba, A. Dutta, J. Llados, Y.-Z. Song, Doodle to search: practical zero-shot sketch-based image retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2179–2188.
[5] Q. Liu, L. Xie, H. Wang, A.L. Yuille, Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3662–3671.
[6] T. Dutta, S. Biswas, Style-guided zero-shot sketch-based image retrieval, in: BMVC, 2019, p. 209.
[7] T. Dutta, A. Singh, S. Biswas, Adaptive margin diversity regularizer for handling data imbalance in zero-shot SBIR, in: European Conference on Computer Vision, Springer, 2020, pp. 349–364.
[8] Y. Wang, F. Huang, Y. Zhang, R. Feng, T. Zhang, W. Fan, Deep cascaded cross-modal correlation learning for fine-grained sketch-based image retrieval, Pattern Recognit. 100 (2020) 107148.
[9] F. Huang, C. Jin, Y. Zhang, K. Weng, T. Zhang, W. Fan, Sketch-based image retrieval with deep visual semantic descriptor, Pattern Recognit. 76 (2018) 537–548.
[10] H. Zhang, S. Liu, C. Zhang, W. Ren, R. Wang, X. Cao, SketchNet: sketch classification with web images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1105–1113.
[11] P. Sangkloy, N. Burnell, C. Ham, J. Hays, The sketchy database: learning to retrieve badly drawn bunnies, ACM Trans. Graphics (TOG) 35 (4) (2016) 1–12.
[12] Y. Shen, L. Liu, F. Shen, L. Shao, Zero-shot sketch-image hashing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3598–3607.
[13] J. Lei, Y. Song, B. Peng, Z. Ma, L. Shao, Y.-Z. Song, Semi-heterogeneous three-way joint embedding network for sketch-based image retrieval, IEEE Trans. Circuits Syst. Video Technol. (2019).
[14] L. Liu, F. Shen, Y. Shen, X. Liu, L. Shao, Deep sketch hashing: fast free-hand sketch-based image retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2862–2871.
[15] J. Zhang, F. Shen, L. Liu, F. Zhu, M. Yu, L. Shao, H. Tao Shen, L. Van Gool, Generative domain-migration hashing for sketch-to-image retrieval, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 297–314.
[16] V. Kumar Verma, A. Mishra, A. Mishra, P. Rai, Generative model for zero-shot sketch-based image retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.
[17] K. Pang, Y.-Z. Song, T. Xiang, T.M. Hospedales, Cross-domain generative learning for fine-grained sketch-based image retrieval, in: BMVC, 2017, pp. 1–12.
[18] L. Guo, J. Liu, Y. Wang, Z. Luo, W. Wen, H. Lu, Sketch-based image retrieval using generative adversarial networks, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 1267–1268.
[19] J. Zhu, X. Xu, F. Shen, R.K.-W. Lee, Z. Wang, H.T. Shen, Ocean: a dual learning approach for generalized zero-shot sketch-based image retrieval, in: 2020 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2020, pp. 1–6.
[20] A. Dutta, Z. Akata, Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval, CVPR, 2019.
[21] Z. Zhang, Y. Zhang, R. Feng, T. Zhang, W. Fan, Zero-shot sketch-based image retrieval via graph convolution network, in: AAAI, 2020, pp. 12943–12950.
[22] J. Li, Z. Ling, L. Niu, L. Zhang, Bi-directional domain translation for zero-shot sketch-based image retrieval, arXiv:1911.13251(2019).
[23] A. Pandey, A. Mishra, V.K. Verma, A. Mittal, H. Murthy, Stacked adversarial network for zero-shot sketch based image retrieval, in: The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 2540–2549.
[24] G.A. Miller, WordNet: An Electronic Lexical Database, MIT press, 1998.
[25] P. Sangkloy, N. Burnell, C. Ham, J. Hays, The sketchy database: learning to retrieve badly drawn bunnies, in: ACM Transactions on Graphics (Proceedings of SIGGRAPH), 2016.
[26] M. Eitz, J. Hays, M. Alexa, How do humans sketch objects? ACM Trans. Graphics (TOG) 31 (4) (2012) 1–10.
[27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[28] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, T.M. Hospedales, Deep spatial-semantic attention for fine-grained sketch-based image retrieval, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5551–5560.
[29] F. Wang, L. Kang, Y. Li, Sketch-based 3D shape retrieval using convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1875–1883.
[30] C. Peng, N. Wang, J. Li, X. Gao, DLFace: deep local descriptor for cross-modality face recognition, Pattern Recognit. 90 (2019) 161–171.
[31] D. Liu, X. Gao, N. Wang, C. Peng, J. Li, Iterative local re-ranking with attribute guided synthesis for face sketch recognition, Pattern Recognit. 109 (2021) 107579.
[32] Y. Lei, Z. Zhou, P. Zhang, Y. Guo, Z. Ma, L. Liu, Deep point-to-subspace metric learning for sketch-based 3D shape retrieval, Pattern Recognit. 96 (2019) 106981.
[33] R. Hu, J. Collomosse, A performance evaluation of gradient field hog descriptor for sketch based image retrieval, Comput. Vision Image Understanding 117 (7) (2013) 790–806.
[34] J.M. Saavedra, Sketch based image retrieval using a soft computation of the histogram of edge local orientations (S-HELO), in: 2014 IEEE International Conference on Image Processing (ICIP), IEEE, 2014, pp. 2998–3002.
[35] J.M. Saavedra, J.M. Barrios, S. Orand, Sketch based image retrieval using learned keyshapes (LKS), in: BMVC, vol. 1, 2015, p. 7.
[36] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T.M. Hospedales, C.-C. Loy, Sketch me that shoe, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 799–807.
[37] Y. Qi, Y.-Z. Song, H. Zhang, J. Liu, Sketch-based image retrieval via siamese convolutional neural network, in: 2016 IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 2460–2464.
[38] C. Bai, J. Chen, Q. Ma, P. Hao, S. Chen, Cross-domain representation learning by domain-migration generative adversarial network for sketch based image retrieval, J. Vis. Commun. Image Represent 71 (2020) 102835.
[39] U. Chaudhuri, B. Banerjee, A. Bhattacharya, M. Datcu, A simplified framework for zero-shot cross-modal sketch data retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 182–183.
[40] C. Deng, X. Xu, H. Wang, M. Yang, D. Tao, Progressive cross-modal semantic network for zero-shot sketch-based image retrieval, IEEE Trans. Image Process. 29 (2020) 8892–8902.
[41] X. Xu, K. Lin, H. Lu, L. Gao, H.T. Shen, Correlated features synthesis and alignment for zero-shot cross-modal retrieval, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1419–1428.
[42] S.K. Yelamarthi, S.K. Reddy, A. Mishra, A. Mittal, A zero-shot framework for sketch based image retrieval, in: European Conference on Computer Vision, Springer, 2018, pp. 316–333.
[43] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: a deep quadruplet network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 403–412.
[44] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, in: International Conference on Machine Learning, 2015, pp. 2048–2057.
[45] A. Khatun, S. Denman, S. Sridharan, C. Fookes, Joint identification-verification for person re-identification: a four stream deep learning approach with improved quartet loss function, Comput. Vision Image Understanding (2020) 102989.
[46] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Rai-

son, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: an imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035.

[47] P. Lu, G. Huang, Y. Fu, G. Guo, H. Lin, Learning large euclidean margin for sketch-based image retrieval, arXiv:1812.04275(2018).

[48] J. Li, Z. Ling, L. Niu, L. Zhang, Zero-shot sketch-based image retrieval with structure-aware asymmetric disentanglement, arXiv:1911.13251(2019).

[49] Y. Gong, S. Lazebnik, A. Gordo, F. Perronnin, Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 35 (12) (2012) 2916–2929.

[50] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.

[51] Y. Fu, T. Xiang, Y.-G. Jiang, X. Xue, L. Sigal, S. Gong, Recent advances in zero-shot recognition: toward data-efficient understanding of visual content, IEEE Signal Process. Mag. 35 (1) (2018) 112–125 .

Osman Tursun is currently a PhD candidate at the School of Electrical Engineering and Robotics, Queensland University of Technology. His current research interests include large-scale image retrieval, image restoration, and weakly-supervised learning. He received a BSc degree in computer science from the University of Science and Technology of China (USTC), and an MSc degree in computer engineering from the Middle East Technical University (METU), Turkey.

Dr Simon Denman received a BEng (Electrical), BIT, and PhD in the area of object tracking from the Queensland University of Technology (QUT) in Brisbane, Australia. He is currently a Senior Lecturer within the School of Electrical Engineering and Robotics at QUT, and is co-leader of the Applied Data Science programme in the QUT Centre for Data Science. His active areas of research include deep machine learning and its applications to computer vision and signal processing.

Sridha Sridharan has a BSc (Electrical Engineering) degree and obtained a MSc (Communication Engineering) degree from the University of Manchester, UK and a PhD degree from University of New South Wales, Australia. He is currently with the Queensland University of Technology (QUT) where he is a Professor in the School Electrical Engineering and Robotics. Professor Sridharan is the Leader of the Research Program in Signal Processing, Artificial Intelligence, and Vision Technologies (SAIVT) at QUT, with strong focus in the areas of computer vision, pattern recognition and machine learning. He has published over 600 papers consisting of publications in journals and in refereed international conferences in the areas of Image and Speech technologies. Prof Sridharan has also received a number of research grants from various funding bodies including Commonwealth competitive funding schemes such as the Australian Research Council (ARC) and the National Security Science and Technology (NSST) unit.

Ethan Goan is a PhD candidate within the School of Electrical Engineering and Robotics at Queensland University of Technology, where he has also received BEng (Electrical). He is a member of the Signal Processing, Artificial Intelligence and Technology (SAIVT) Lab and the Australian Centre for Excellence in Mathematics and Statistics. His research interests include probabilistic inference for large scale computer vision models.

Clinton Fookes is a Professor in Vision and Signal Processing in the School of Electrical Engineering & Robotics at the Queensland University of Technology. He holds a BEng (Aerospace/ Avionics), an MBA, and a PhD in computer vision. He actively researches across computer vision, machine learning, signal processing and pattern recognition areas. He serves on the editorial boards for the Pattern Recognition Journal and the IEEE Transactions on Information Forensics and Security. He is a Senior Member of the IEEE, an Australian Institute of Policy and Science Young Tall Poppy, an Australian Museum Eureka Prize winner, and a Senior Fulbright Scholar.