ACNet: Approaching-and-Centralizing Network for Zero-Shot Sketch-Based Image Retrieval

Hao Ren, Ziqiang Zheng, Yang Wu, *Member, IEEE*, Hong Lu, *Member, IEEE*, Yang Yang, *Senior Member, IEEE*, Ying Shan, *Senior Member, IEEE*, and Sai-Kit Yeung

***Abstract*— The huge domain gap between sketches and photos poses huge challenges for Sketch-Based Image Retrieval (SBIR). The Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR) is more generic and practical but brings an even greater challenge: the additional knowledge gap between the seen and unseen categories. In order to simultaneously mitigate both gaps, we propose an Approaching-and-Centralizing Network (termed "ACNet") to jointly optimize sketch-to-photo synthesis and image retrieval. The retrieval module guides the synthesis module to generate large amounts of diverse photo-like images that help the sketch domain gradually approach the photo domain to eliminate the domain gap, and thus better serves retrieval. Meanwhile, the retrieval module itself centralizes the embeddings of training samples for learning a similarity measurement to eliminate the knowledge gap. Our approach is simple yet effective, which achieves state-of-the-art performance on two widely used ZS-SBIR datasets and surpasses previous methods by a large margin (*e.g.*, 8.2% improvement in terms of mAP@all on TU-Berlin Extended dataset).**

***Index Terms*— Sketch-based image retrieval, zero-shot learning, metric learning, deep learning.**

I. Introduction

SKETCH-BASED Image Retrieval (SBIR) [1], [2], [3], [4], [5], [6] aims to perform cross-domain image retrieval among the photos and sketches drawn by humans. Touch-screen devices (*e.g.*, smartphones and iPads) enable us to draw free-hand sketches conveniently. The drawn sketches are regarded as queries and the retrieval system is expected to return some relevant photos according to the user's intent. Considering the lack of colors, textures and detailed structural information, the sketches are highly *iconic*, *succinct* and *abstract*. The huge domain gap and the asymmetrical information between sketches and photos pose great challenges for SBIR. Category labels provide supervision to the retrieval model for overcoming the gap, but the model may take shortcuts [7] to overfit the category labels by paying too much attention to category-specific samples, feature representations, and their specific distributions, turning the retrieval problem into a classification problem. The follow-up Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR) is introduced in [8] in a more practical and realistic setting, where the test data are from unseen categories. The knowledge gap between seen categories and unseen categories makes ZS-SBIR more intractable. The domain gap and knowledge gap are the two biggest challenges for ZS-SBIR.

To simultaneously mitigate both gaps, we design a novel, simple, and effective **approaching and centralizing** network (**ACNet**), which jointly trains sketch-to-photo synthesis and image retrieval, as shown in Fig. 1. The sketch-to-photo synthesis module encourages the retrieval module to focus more on domain-agnostic information for proper similarity measurement. This is done by constantly refining and feeding the synthesized photo-like images into the retrieval module during the training phase. Even though there are some noise and uncertainty introduced along with the synthesis, the continuously generated and refined images are of high diversity, which gradually **approach** the photo domain and thus benefit training a robust retrieval module. Meanwhile, the retrieval module is designed to **centralize** the embeddings of images by using a centralized proxy rather than the specific training samples to represent each category. Therefore, it can learn a cross-domain similarity measurement that is aware of the overall category differences yet insensitive to the specific sample distribution in each category.

We choose CycleGAN [9] as our synthesis module due to its simplicity and effectiveness, and other image-to-image translation networks can also be applied. For the retrieval module, we utilize the NormSoftmax [10] loss to centralize the embeddings of both sketches and photos belonging to the same category. These two modules are jointly trained to ensure that both the domain gap and knowledge gap can be mitigated as much as possible. Differently, previous pairwise

Manuscript received 18 July 2022; revised 25 November 2022, 19 December 2022, and 4 February 2023; accepted 13 February 2023. Date of publication 24 February 2023; date of current version 6 September 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62072112 and in part by the National Key Research and Development Program of China under Grant 2020AAA0108301. This article was recommended by Associate Editor H.-C. Shih. *(Hao Ren and Ziqiang Zheng contributed equally to this work.) (Corresponding author: Hong Lu.)*

Hao Ren and Hong Lu are with the Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai 200438, China (e-mail: hren17@fudan.edu.cn; honglu@fudan.edu.cn).

Ziqiang Zheng is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China (e-mail: zhengziqiang1@gmail.com).

Yang Wu is with Tencent AI Lab, Shenzhen 518100, China (e-mail: dylanywu@tencent.com).

Yang Yang is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: dlyyang@gmail.com).

Ying Shan is with the ARC Lab, Tencent PCG, Shenzhen 518000, China (e-mail: yingsshan@tencent.com).

Sai-Kit Yeung is the Division of Integrative System and Design (ISD), and the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China (e-mail: saikit@ust.hk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3248646.

Digital Object Identifier 10.1109/TCSVT.2023.3248646

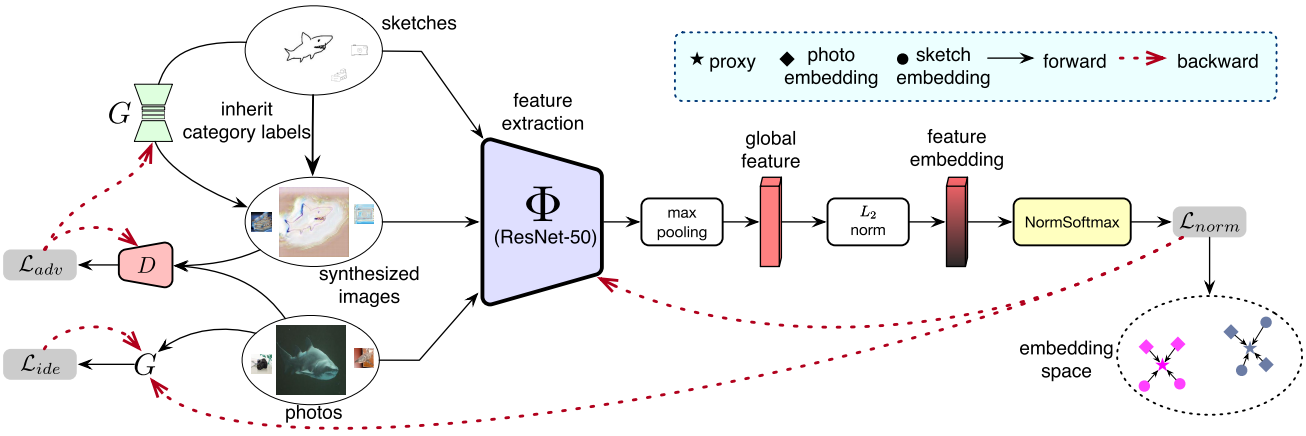


Fig. 1. Overview of the proposed ACNet. The sketch-to-photo generator G aims to translate the sketches to photo-like images while D is to distinguish whether the generated images are real photos and compute $\mathcal{L}_{adv}$. The identity loss $\mathcal{L}_{ide}$ helps generate more photo-like images by forcing G to reconstruct the real photos. The main pipeline Φ serves as the standard process for embedding learning. The embeddings of the sketches, synthesized images, and photos from the same category are enforced to be close to the assigned proxy, and far away from other proxies under the constraint of $\mathcal{L}_{norm}$. Different colors indicate the different categories. We design a joint training manner to integrate G and Φ. We refer the readers to check the forward and backward procedures to better understand our joint training scheme.

losses (*e.g.*, Contrastive [11] and Triplet [12]) aim to learn the similarity measurement by sampling the informative pairs or triplets in one batch. It is easy for these methods to overfit the specific training samples, and some may distract the training by bringing high gradients. The NormSoftmax loss assigns one proxy as the anchor for each category and learns the similarity measurement with the gradients from all the samples belonging to the assigned proxy. This loss function can better coordinate the relationship between all samples from each category through proxy-based optimization [13], [14], and can also better alleviate the influence of noise and uncertainty introduced by the synthesis module than other existing losses.

The proposed ACNet achieves new state-of-the-art performances on two widely used ZS-SBIR datasets. Extensive ablation studies have been conducted to dissect our method. Our main contributions can be summarized as follows:

- We propose an "Approaching-and-Centralizing Network (ACNet)" to integrate sketch-to-photo synthesis and image retrieval through a joint training manner, which mitigates both the domain gap and the knowledge gap.
- We adopt the NormSoftmax loss to stabilize our joint training and promote the generalization ability under the zero-shot setting, thanks to its centralizing effect driven by proxy-based optimization.
- Comprehensive ZS-SBIR experiments and ablation studies on Sketchy Extended [8] and TU-Berlin Extended [15] datasets demonstrating the superiority of the proposed ACNet.

II. Related Work

A. Sketch-Based Image Retrieval (SBIR)

SBIR [1], [16] has been studied for decades due to its commercial and realistic applications [17], [18]. Attempts for solving the SBIR task mostly focus on bridging the domain gap between the sketches and photos, which can roughly be grouped into hand-crafted [1], [2] and cross-domain deep learning-based methods [11], [12], [19], [20], [21], [22], [23]. Hand-crafted methods [1], [2] mostly work by extracting the edge map from natural photo images and then matching them with sketches using a Bag-of-Words model. Due to the great successes of deep learning methods, various specifically designed neural networks [3], [4], [5], [6], [24], [25] have been proposed for SBIR. Classical ranking losses, like contrastive loss [11], [20], triplet loss [12], [19] or classification loss [10] have also been introduced.

B. Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR)

SBIR requires all test categories to be seen during training, which cannot be guaranteed or practical in real-world applications. The more challenging, generic, and practical ZS-SBIR [15] task has attracted the attention of the computer vision community due to its real-world applications, in which the test categories do not appear in the training stage. Recent research [8], [26], [27], [28], [29], [30], [31] is exploring solutions for projecting sketches and photos into a shared semantic space to perform accurate cross-domain image retrieval. However, the huge domain gap and the highly abstract sketch representations make it very difficult to perform ideal feature-level content-style disentanglement [7], [26], [27], [28] or bi-direction synthesis [32].

To bridge the knowledge gap between seen and unseen categories, existing methods [28], [29], [33], [34], [35], [36], [37], [38] introduced the semantic embeddings from the extra annotations as class prototypes to present the relationships between semantic categories in the common space. However, there is no explicit connection between the embeddings extracted from visual images and the semantic embeddings borrowed from the extra annotations. Furthermore, the semantic embeddings from the language models [39], [40] are computed based on word co-occurrence frequency. Sometimes, these embeddings are not reliable [31] and cannot express intra-class visual variation. In contrast to previous methods, we assign one proxy for each category and update them **adaptively** based on the given training data with the synthesis module.

C. Sketch-to-Photo Synthesis and Joint Training

Sketch-to-photo synthesis [9], [32], [41], [42], [43], [44], [45], [46] is a notably challenging task in the field of computer vision, which aims to generate photo-realistic images from the given abstract and exaggerated sketches. Sketch2Photo [17] proposed to compose new photo images using the retrieved photo images from the given sketch. The semantic segmentation [47] and image blending [48] techniques were introduced to achieve photo editing according to the user's goal. The first deep learning-based free-hand sketch-to-photo synthesis is SketchyGAN [44], which aims to optimize an encoder-decoder model based on the aligned sketch-photo pairs. Ghosh et al. [49] proposed multi-class photo generation based on incomplete edges or sketches. Sketchformer [50] designed a sequential sketch-to-photo generation model to promote the naturalness of the photo images. Liu et al. [32] conducted unsupervised sketch-to-photo synthesis and further analyzed the potential of adopting the synthesized images for retrieval.

Ideally, a perfect sketch-to-photo generator could synthesize the desired photo images with distinguishable and reliable feature representations for more accurate image retrieval, whilst preserving the intra-class and inter-class distribution after translation. However, the unsupervised domain translation performance is plagued by the huge domain gap as well as the highly abstract sketch representations. The translated photo images still suffer from visual artifacts [51] and noise, even utilizing the semantic priors [3], [51], [52] and laborious generative networks [50], [51], [53], [54]. Since the two modules are optimized separately in existing methods [32], [54], the noise and uncertainty introduced in the synthesis module could be propagated to the retrieval module and the error accumulation could heavily restrict the retrieval performance. Even if the generated images look realistic to humans, their benefits may not be able to surpass the harm to the downstream retrieval task.

Differently, our proposed ACNet jointly optimizes synthesis and retrieval, and thus ensures a significant performance boost. First, previous works (*e.g.*, [32], [54]) optimize the sketch-to-photo module and further retrieval module in the **two-stage training** manner. The two modules are optimized separately. Differently, the proposed method optimizes the two modules in the **joint training** manner. The gradient of the retrieval module is directly propagated to the sketch-to-photo synthesis module and guides the generator on how to synthesize photo-like images with discriminated feature representations. Secondly, we only have the forward sketch-to-photo synthesis and the reconstruction of the photo images compared with the normal GAN structure, our goal is not to generate images with good image quality rather than boosting the overall retrieval performance. Furthermore, the parallel sketch-to-photo synthesis module could be regarded as an effective data augmentation, which could promote the robustness of the retrieval module and also the generalization ability to unseen images. The synthesis module can help alleviate the model overfit to training data. Bhunia et al. [55] also designed joint training, but combined photo-to-sketch synthesis and fine-grained SBIR through a semi-supervised manner (with some photo-sketch pairs). We work on ZS-SBIR, and we argue that some important information will be lost after the photo-to-sketch synthesis. Our ACNet does not require any photo-sketch pairs and the gradient of the retrieval module is directly propagated to the sketch-to-photo synthesis module to help the photo generation. Besides, we have a more in-depth dissection of the joint training of synthesis and retrieval.

III. Method

A. Problem Formulation

Consider n photos and m sketches denoted as $\mathcal{P} = \{(p_i, y_{p_i})|y_{p_i} \in \mathcal{Y}\}_{i=1}^{n}$, and $\mathcal{S} = \{(s_i, y_{s_i})|y_{s_i} \in \mathcal{Y}\}_{i=1}^{m}$ respectively. Under the SBIR setting, $\mathcal{S}$ and $\mathcal{P}$ are divided into the training set and test set with the same category label set $\mathcal{Y}$. SBIR aims to retrieve the best matched $p_j \in \mathcal{P}$ based on a query sketch s_i in $\mathcal{S}$, such that $y_{s_i} = y_{p_j}$. Under the ZS-SBIR setting, $\mathcal{Y}$ is split into $\mathcal{Y}_{tra}$ and $\mathcal{Y}_{test}$, in which there is no category overlap between $\mathcal{Y}_{tra}$ and $\mathcal{Y}_{test}$ ($\mathcal{Y}_{tra} \cap \mathcal{Y}_{test} = \emptyset$). The training data are $\mathcal{S}_{tra} = \{(s_i, y_{s_i})|y_{s_i} \in \mathcal{Y}_{tra}\}$, $\mathcal{P}_{tra} = \{(p_i, y_{p_i})|y_{p_i} \in \mathcal{Y}_{tra}\}$, and the test data are $\mathcal{S}_{test} = \{(s_i, y_{s_i})|y_{s_i} \in \mathcal{Y}_{test}\}$, $\mathcal{P}_{test} = \{(p_i, y_{p_i})|y_{p_i} \in \mathcal{Y}_{test}\}$. The ZS-SBIR model is trained on data $(\mathcal{S}_{tra}, \mathcal{P}_{tra})$, and tested on $(\mathcal{S}_{test}, \mathcal{P}_{test})$.

B. Main Pipeline

The overall architecture of the proposed method is illustrated in Fig. 1.

1) Approaching by Sketch-to-Photo Synthesis: Suppose the sketch s_i from $\mathcal{S}_{tra}$ and the photo p_j from $\mathcal{P}_{tra}$, we first aim to generate a photo-like image $s_i^* = G(s_i)$ based on s_i through a generator $G : \mathcal{S}_{tra} \rightarrow \mathcal{P}_{tra}$. The adversarial loss of GAN architecture can be expressed as:

$$\begin{aligned}\mathcal{L}_{adv} = &\mathbb{E}_{s_i, p_j \sim P_{data}(\mathcal{S}_{tra}, \mathcal{P}_{tra})}\left[\log D(p_j)\right] \\ &+\mathbb{E}_{s_i \sim P_{data}(\mathcal{S}_{tra})}\left[\log(1 - D(G(s_i)))\right],\end{aligned} \tag{1}$$

where D is the discriminator to distinguish whether the image is from the real photo domain. The goal is to learn a mapping function, which could generate photo-like images that match the real photo distribution $P_{data}(\mathcal{P}_{tra})$. After the sketch-to-photo synthesis, we assign the category label of s_i to s_i^*. It is non-trivial to define such label-preserving synthesis-based transformations, especially when uncertainty and noise have been introduced with image synthesis. The synthesized images possess more texture and RGB information and thus gradually approach the photo domain, which can better serve cross-domain image retrieval. Considering there is no pixel-level constraint for G, G would tend to generate images with visual artifacts. In order to alleviate this problem, identity loss $\mathcal{L}_{ide}$ between p_j and $G(p_j)$ is adopted as an additional constraint, which is firstly proposed in CycleGAN [9], expressed as:

$$\mathcal{L}_{ide} = \mathbb{E}_{\mathcal{P}_{tra}} \|G(p_j) - p_j\|_1. \tag{2}$$

Since s_i and p_j share similar semantic contents (*e.g.*, the category and structure information), we can boost the synthesis

performance by reconstructing the photos at the same time. With the full supervision of $\mathcal{L}_{ide}$, we could generate more photo-like images.

2) Feature Extraction: The sketch s_i, generated image s_i^*, and real photo p_j are fed into the same backbone network to extract features. Like previous methods [29], [31], [34], we adopt ResNet-50 [56] (denoted as Φ) as backbone. The outputs after max-pooling are transformed into the desirable embedding dimension through a fully connected layer. L_2-norm is adopted to obtain the final embedding for the retrieval task.

3) Centralizing With NormSoftmax: The NormSoftmax loss [10] is used as our objective function to increase the inter-class distance and reduce the intra-class distance over the sketch and photo set. Each category is assigned a learnable proxy, the learnable proxies are shared between sketches and photos. They are initialized with values sampled from the normal distribution $\mathcal{N}(0, 1)$. The final embedding x is enforced to be close to the proxy of its category, and far away from other proxies, as shown in Fig. 1. This property ensures that it is learning the similarity measurement rather than the category itself, so as to ensure that it also has a certain generalization ability in the category that has not been seen before, and will not be overfitted to the training data. It potentially solves the problem of the knowledge gap with the help of the synthesis module. The objective function for x is expressed as:

$$\mathcal{L}_{norm}(x) = -log(\frac{exp(\frac{x^T p_y}{t})}{\sum_{z \in Z} exp(\frac{x^T p_z}{t})}), \tag{3}$$

where Z is the set of all proxies, p_y is the proxy of x, t is temperature scale. We set $t = 0.05$ following the default setting in [10]. Based on the three inputs (s_i, p_j and s_i^*) of our backbone network, we can get three losses described as:

$$\mathcal{L}_{norm}^{s_i} = \mathcal{L}_{norm}(\Phi(s_i)), \tag{4}$$

$$\mathcal{L}_{norm}^{p_j} = \mathcal{L}_{norm}(\Phi(p_j)), \tag{5}$$

$$\mathcal{L}_{norm}^{s_i^*} = \mathcal{L}_{norm}(\Phi(s_i^*)), \tag{6}$$

with $\mathcal{L}_{norm}^{s_i^*}$ we can better reduce the domain gap between the sketch and photo domain through the intermediate synthesized images. The final NormSoftmax loss is expressed by aggregating these three losses as:

$$\mathcal{L}_{norm} = \mathcal{L}_{norm}^{s_i} + \mathcal{L}_{norm}^{p_j} + \mathcal{L}_{norm}^{s_i^*}. \tag{7}$$

4) Final Objective Function: The final objective function for G, D and Φ is described as:

$$\mathcal{L}(G, D, \Phi) = \mathcal{L}_{adv} + \lambda \mathcal{L}_{norm} + \gamma \mathcal{L}_{ide}, \tag{8}$$

where λ and γ are hyper-parameters to balance the contribution of each component. We set $\lambda = 10$ and $\gamma = 0.1$ in our experiments and provide comprehensive experiments using different values of λ and γ in Section IV-E.

5) Joint Approaching and Centralizing: We optimize G and Φ through a **joint training** manner and the synthesized images are constantly fed into Φ. Through the sketch-to-photo synthesis, we could generate sufficient photo-like examples with high data diversity and force Φ to extract more reliable and effective features under the constraint of $\mathcal{L}_{norm}^{s_i^*}$. Besides, by sufficiently generating samples in the latent space and enforcing them to be centralized through the proxies in the embedding space, we could promote the generalization ability of our backbone network. Our framework is **model-agnostic** and we can choose various GAN architectures for synthesis and backbone networks for feature extraction. We provide more experiments about different GANs and backbone networks in Section IV-E.

6) Inference: In the test phase, G first generates one photo-like image based on the sketch query, and the generated image is fed into Φ to obtain the embedding. The cosine similarity between obtained embedding and photo embeddings from the gallery is used as the criterion to perform the retrieval task. We do **not** conduct reverse photo-to-sketch synthesis since we are performing sketch-based image retrieval and the photos are not available at the inference time.

IV. Experiments

A. Experimental Setup

1) Datasets: Two widely used public datasets are tested in our experiments. The **Sketchy Extended** dataset contains 75,481 sketches, 73,002 photos (12,500 images from [57] and 60,502 images from ImageNet [58] organized by Liu et al. [3]) from 125 categories. We follow the same zero-shot data partitioning as [8], in which 21 unseen classes from ImageNet for testing and other classes for training. The **TU-Berlin Extended** [59] dataset contains 20,000 sketches evenly distributed over 250 object categories. 204,070 photo images collected by Liu et al. [3] are included. The partition protocol introduced in [15] is used to create zero-shot training and test sets. 30 randomly picked classes, each of which includes at least 400 photo images, are used for testing, and other classes are used for training.

2) Implementation Details: Following the previous methods [8], [26], [27], [29], [31], [33], [34], ResNet-50 [56] is adopted as the backbone network, and the embedding dimension is 512. The image resolution is set to 224×224 and the batch size is 64. Only random horizontal flipping is conducted for data augmentation. We follow the generator architecture of the vanilla CycleGAN [9] and design our sketch-to-photo generator G. The PatchGAN discriminator [60] architecture is adopted for designing D. We refer the readers to check Table I and Table III for more details. Please note that we only design the forward sketch-to-photo generator. Furthermore, to reduce the inference time and computational cost, we modify the channel number of the first convolutional layer to 8, and the number of residual blocks to 8, so that our generator is very lightweight. Similarly, we set the channel number of the first convolutional layer in D to 8. All the models are warm-up with 1 epoch and have been optimized with 10 epochs. We choose Adam [61] with learning rate of $1e^{-3}$ for optimization. Our

TABLE I
THE NETWORK CONFIGURATION OF G. c IS THE NUMBER OF CHANNELS OF THE FIRST CONVOLUTION LAYER. **CIR**, **RB** AND **DIR** INDICATE THE CONV-INSTANCENORM-RELU BLOCK, RESIDUAL BLOCK AND DECONV-INSTANCENORM-RELU BLOCK, RESPECTIVELY

G (sketch-to-photo generator)				
Module Type	Kernel Size	Stride Size	Channel Number	Padding Type
CIR	7	1	$c = 8$	reflect
	3	2	16 ($2 \times c$)	zero
	3	2	32 ($4 \times c$)	zero
RB	3	1	32 ($4 \times c$)	reflect
	3	1	32 ($4 \times c$)	reflect
	3	1	32 ($4 \times c$)	reflect
	3	1	32 ($4 \times c$)	reflect
	3	1	32 ($4 \times c$)	reflect
	3	1	32 ($4 \times c$)	reflect
	3	1	32 ($4 \times c$)	reflect
	3	1	32 ($4 \times c$)	reflect
DIR	3	2	16 ($2 \times c$)	zero
	3	2	8	zero
	7	1	3	reflect
Tanh	-	-	3	reflect

code is implemented with PyTorch [62] library and the experiments are conducted on the Geforce RTX 3090 GPU. The code of our work is available on https://github.com/leftthomas/ACNet. The ZS-SBIR experimental results of using different architectures of G are included in Section IV-E.

3) Evaluation Metrics: Precision ($Prec$) and mean Average Precision (mAP) are two main metrics for the evaluation of ZS-SBIR task [15]. For a fair comparison, we follow the standard evaluation method [34]. $Prec$ is calculated for top k (*i.e.*, 100, 200) ranked results, and mAP is computed for top k or all ranked results. Higher $Prec$ and mAP indicate better retrieval performance.

B. Triplet vs. NormSoftmax

We first aim to demonstrate that NormSoftmax loss [10] is more effective than Triplet loss [12] for the ZS-SBIR task. We conduct the ZS-SBIR experiments on both Sketchy Extended [8] and TU-Berlin Extended [15] datasets. The quantitative results are reported in Table II. To make a fair comparison, all the hyper-parameters are set to the same. The Triplet loss can only achieve 35.5% while the NormSoftmax loss has achieved 45.2% without sketch-to-photo synthesis in terms of $mAP@200$ on Sketchy Extended dataset, which has gained a large performance improvement. The proxy-based optimization framework could centralize all the samples that belong to the same proxy and prevent the model from remembering some category-specific samples, which is catastrophic for the generalization ability to unseen categories. The centralizing effect of NormSoftmax loss can make the retrieval model optimized better under category supervision. We can also observe that the NormSoftmax loss outperforms the Triplet loss by a large margin on the TU-Berlin Extended dataset, which indicates that the proxy-based loss has a significant priority over the triplets-based loss on the ZS-SBIR task.

Triplet loss requires careful hard negative mining among the mini-batches (*locally*) and weighting strategies to obtain the most informative pairs. In contrast to the pair-based loss, it is substantially easier to optimize NormSoftmax since it heavily reduces the sampling complexity (from $O(N^3)$ to $O(NC)$, where N, C are the number of samples and the number of proxies (*globally*), respectively, in a mini-batch). The non-convex Triplet loss can easily lead to local optima [63] whilst the convex NormSoftmax loss has a globally optimal solution.

C. Two-Stage Training vs. Joint Training

In this section, we aim to demonstrate the limitations of the previous two-stage training, which is adopted to mitigate the domain gap between the sketches and photos for the ZS-SBIR. For the first synthesis part, we adopt the vanilla CycleGAN [9] to perform the unpaired image-to-image (I2I) translation between the sketches and photos.[1] The train/test split strictly follows the ZS-SBIR setting. After the unpaired I2I model converges, we adopt the trained sketch-to-photo generator for inference and translate all the sketch images into photo-like images. Please note, both training and test sketch images have been translated to the photo domain for further training and testing. We then perform the ZS-SBIR experiments based on the synthesized photo-like images and the real photo images. The original category labels of the sketches are inherited for training. We choose both Triplet loss and the adopted NormSoftmax loss for optimization and the quantitative results are also reported in Table II.

Compared with the counterpart results achieved on the vanilla setting (between the original sketches and photos), there is a slight performance drop regardless of using Triplet loss or NormSoftmax loss on Sketchy Extended dataset. As for TU-Berlin Extended dataset, we can obtain better results based on NormSoftmax loss by approaching the sketch domain to the photo domain. In contrast, there is a performance drop while using Triplet loss for optimization, which also indicates that NormSoftmax [10] loss has a stronger ability to coordinate the real samples and the synthesized samples than Triplet [12] loss. The NormSoftmax loss is more compatible with the two-stage sketch-to-photo synthesis for data augmentation to mitigate the domain gap. This also reflects that if the two-stage sketch-to-photo synthesis module is used as a data augmentation, it will not consistently improve the retrieval performance. And compared with our joint training strategy (Exp 3&6), we can find that joint training has greatly improved the retrieval performance (*i.e.*, 43.5% $\rightarrow$ 51.7% in terms of $mAP@200$ on Sketchy Extended dataset with NormSoftmax loss). Additionally, both in Sketchy Extended and TU-Berlin Extended datasets, we can see that the performance improvement is consistent with the usage of NormSoftmax loss. Though the Triplet loss performs worse on TU-Berlin Extended dataset, this supports our claim in Section IV-B that NormSoftmax loss is more robust than Triplet loss.

To better understand the two-stage synthesis results, referring to [70], we compute the L_1 pixel-level distances between 50 randomly sampled instances from each training category

[1] We adopt the official implementation https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix following the default setting.

TABLE II

THE ZS-SBIR PERFORMANCE COMPARISON UNDER VARIOUS SETTINGS ON SKETCHY EXTENDED [8] AND TU-BERLIN EXTENDED [15] DATASETS: 1) TRIPLET AND NORMSOFTMAX FOR OPTIMIZATION; AND 2) WITH OR WITHOUT SKETCH-TO-PHOTO SYNTHESIS THROUGH TWO-STAGE TRAINING OR JOINT TRAINING TO MITIGATE THE DOMAIN GAP. THE BEST RESULTS FOR EACH SETTING ARE BOLD

Exp	Loss Type	Synthesis		Sketchy Extended				TU-Berlin Extended			
		Two-stage	Joint	mAP @200	mAP @all	Prec @100	Prec @200	mAP @200	mAP @all	Prec @100	Prec @200
1	$\mathcal{L}_{triplet}$	-	-	35.5	40.8	51.2	47.3	**38.1**	**36.8**	**49.8**	**47.1**
2		✓	-	33.1	38.2	48.1	44.4	36.3	35.2	47.7	45.4
3		-	✓	**39.1**	**44.5**	**53.2**	**49.4**	26.5	27.5	36.7	34.8
4	$\mathcal{L}_{norm}$	-	-	45.2	48.6	60.2	55.7	47.9	46.5	57.7	55.1
5		✓	-	43.5	47.5	58.0	53.9	48.4	46.6	58.0	55.6
6		-	✓	**51.7**	**55.9**	**64.3**	**60.8**	**57.7**	**57.7**	**65.8**	**64.4**

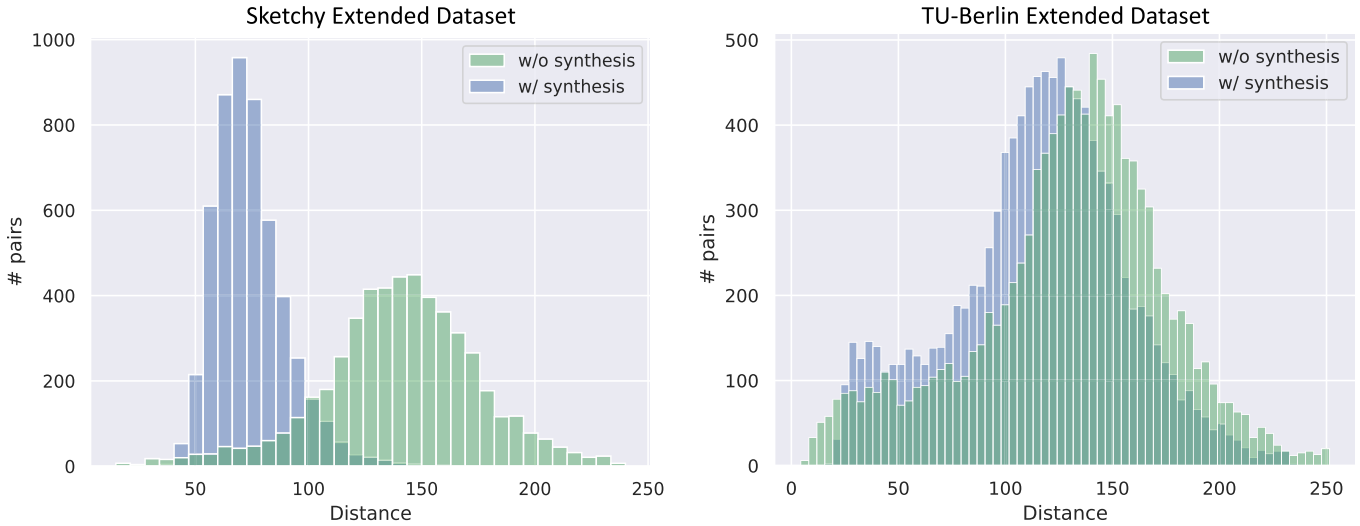


Fig. 2. Histograms of sample distances under two settings on Sketchy Extended [8] and TU-Berlin Extended [15] datasets: 1) between original sketches and real photos ("w/o synthesis") and 2) between synthesized images and real photos ("w/ synthesis").

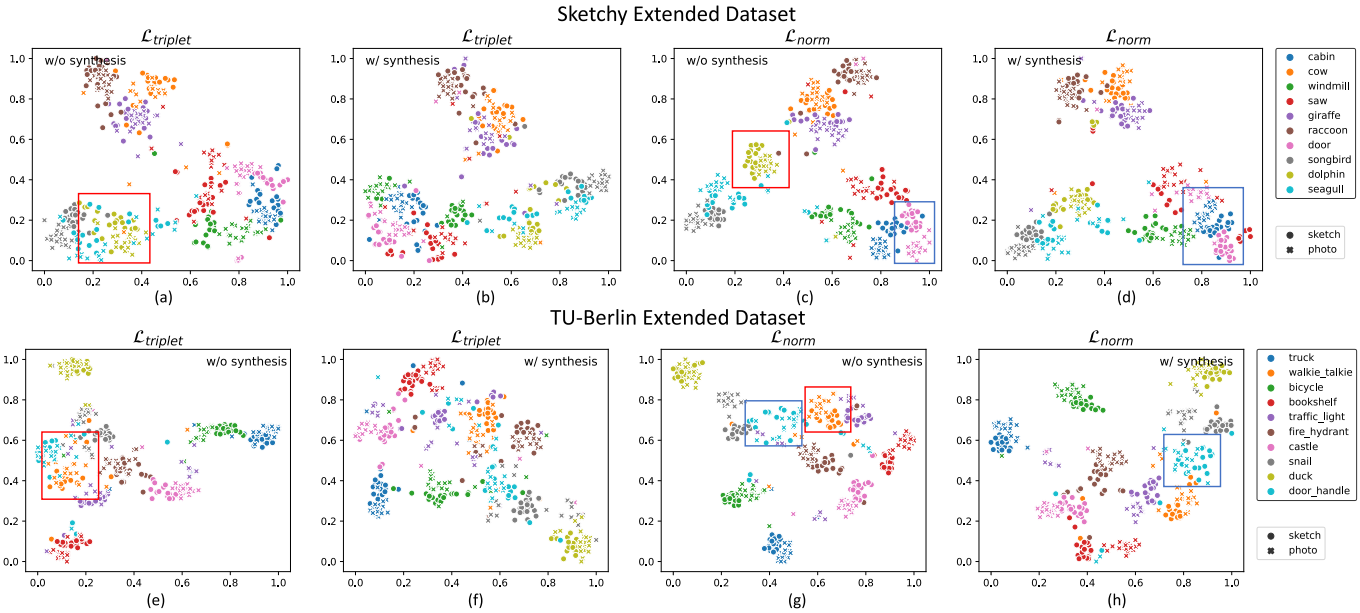


Fig. 3. T-SNE visualization of sketch and photo embeddings on Sketchy Extended [8] and TU-Berlin Extended [15] datasets. We randomly choose 20 samples from each of the 10 unseen test categories for visualization. Different colors refer to different categories. The two-stage training strategy cannot obtain more separable embeddings than the vanilla setting, regardless of using Triplet loss or NormSoftmax loss. We refer the readers to pay more attention to the regions covered by the same color boxes for better comparison.

under two settings: 1) between the original sketches and the real photos and 2) between the synthesized images and the real photos. We provide the distance histogram in Fig. 2 to illustrate the domain distance. With the two-stage synthesis, the distances have been reduced a lot on the Sketchy Extended dataset, which demonstrates that sketch-to-photo synthesis has effectively mitigated the domain gap. The sketches from the TU-Berlin Extended dataset are highly abstract, succinct, and

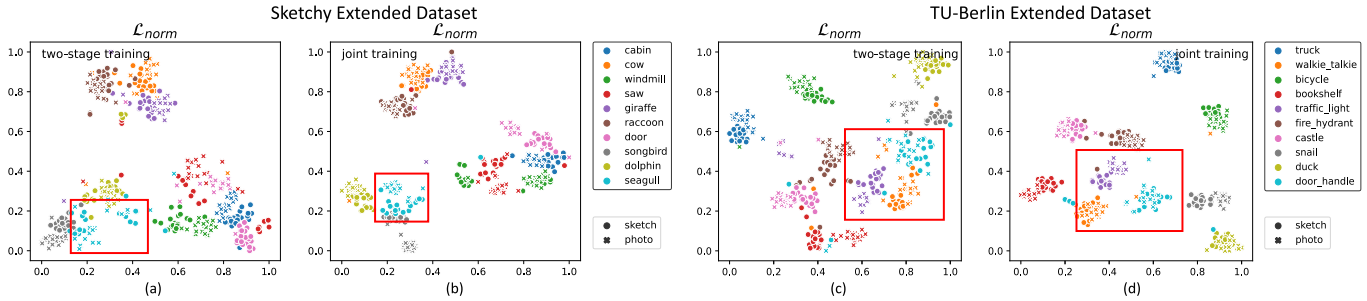


Fig. 4. T-SNE visualization of sketch and photo embeddings on Sketchy Extended [8] and TU-Berlin Extended [15] datasets under two-stage training and joint training. We refer the readers to pay more attention to the regions covered by the same color boxes for better comparison.

TABLE III

THE NETWORK CONFIGURATION OF D. c IS THE NUMBER OF CHANNELS OF THE FIRST CONVOLUTION LAYER. **CLR** AND **CILR** INDICATE THE CONV-LEAKYRELU LAYER AND CONV-INSTANCENORM-LEAKYRELU LAYER. THE SLOPE FOR THE LEAKYRELU IS 0.2

D (Discriminator)			
Module Type	Kernel Size	Stride Size	Channel Number
CLR	4	2	$c = 8$
CILR	4	2	16 ($2 \times c$)
	4	2	32 ($4 \times c$)
	4	1	64 ($8 \times c$)
Conv	4	1	1

exaggerated. With the two-stage synthesis, we can still reduce the domain gap.

Furthermore, we provide the T-SNE visualization of sketch and photo embeddings in Fig. 3 under the four settings (Exp 1,2,4&5) of Table II. The two-stage training strategy cannot obtain much more separable embeddings to achieve a large ZS-SBIR performance gain even though the synthesis can reduce the domain gap. We attribute this failure to the reason that the gradients of the retrieval module cannot be directly utilized for the optimization of the synthesis module. Thus, the synthesis module has no idea how to generate images, which can better serve image retrieval. We provide a direct comparison between the distribution of the embeddings from the two-stage training and the proposed joint training in Fig. 4. As illustrated, our joint training has a strong ability to split the embeddings from different classes, which leads to better ZS-SBIR performance.

D. Comparison With SOTA

We compare the proposed method against existing state-of-the-art ZS-SBIR methods on both Sketchy Extended [8] and TU-Berlin Extended [15] datasets. The quantitative results of different methods are reported in Table IV. The backbone of the proposed method is ResNet-50. Our method outperforms existing state-of-the-art methods by a large margin even without any specially designed backbones [65] or semantic guidance [35]. Our method has achieved 57.7% $mAP@all$ and 65.8% $Prec@100$ on the TU-Berlin Extended dataset. The highest results of the other methods are only 49.5% $mAP@all$ and 61.6% $Prec@100$, and our method has achieved 8.2% improvement on $mAP@all$ and 4.2% improvement on $Prec@100$. Besides, our results of using the hashing codes even exceed the previous highest performance. For the Sketchy Extended [8] datasets, the $Prec@200$ reaches 60.8%, which is comparable to the previous best results.

The proposed ACNet has achieved a larger performance gain on TU-Berlin Extended dataset than Sketchy Extended dataset. We attribute this to the fact that the sketches from the TU-Berlin Extended dataset have more abstract sketch representations and fewer sketch details. Thus, the proposed ACNet can achieve a large performance gain by making the sketch domain approach to the photo domain. We provide more experimental results of using various embedding dimensions and backbone networks (VGG-16 [71] and ResNet-50 [56]) in Section IV-E.

The qualitative results are shown in Fig. 5. We selected three instances from the two datasets to provide an intuitive comparison. The proposed ACNet could effectively return the correct photos given a query sketch. We provide two similar instances from the same category "cow" and the two instances have the same orientation and similar shape representation except for the fine-grained representations on the head part. The proposed method could distinguish these fine-grained representations and provide corresponding desired photos rather than the same photos with prominent feature representations, which demonstrates that the proposed ACNet has extracted effective representations on the unseen categories. Under a more challenging case: the "couch" sketch on the third row of the TU-Berlin Extended dataset, the fifth retrieved photo belongs to the "purse" even though the two categories are conceptually different. This failure is caused by that the retrieved wrong photo sharing very similar structural representations with the real couch photos.

We also provide some failure results on the Sketchy Extended and TU-Berlin Extended datasets in Fig. 6. We selected three instances from the two datasets to provide an intuitive illustration. Take the case of "seagull" as an example, the top-2 wrongly retrieved photos are belonging to "songbird", which is a very similar category to "seagull". The distinction between these two categories is typically classified as a fine-grained problem. We do not effectively capture the differences in local details because our method concentrates on the extraction of global features. Even so, we can still easily see that there are many similarities between the query sketch and the retrieved photos, including shape, posture, etc, which also shows that our method can extract meaningful features to some extent. It is not difficult to see that there is some degree of similarity and correlation between the retrieved

TABLE IV

OVERALL ZS-SBIR COMPARISON OF OUR METHOD AND OTHER APPROACHES ON SKETCHY EXTENDED [8] AND TU-BERLIN EXTENDED [15] DATASETS. "†"DENOTES RESULTS OBTAINED BY HASHING CODES, AND "-" MEANS THAT CORRESPONDING RESULTS ARE NOT REPORTED IN THE ORIGINAL PAPERS. THE BEST AND SECOND-BEST RESULTS ARE BOLD AND UNDERLINED, RESPECTIVELY

Methods	Venue	Semantic	Dim	Sketchy Extended		TU-Berlin Extended	
				mAP@200	Prec@200	mAP@all	Prec@100
GN-Triplet [57]	TOG'16	×	1024	8.3	16.9	18.9	24.1
DSH [3]	CVPR'17	×	64†	5.9	15.3	12.2	19.8
CAAE [8]	ECCV'18	×	4096	15.6	26.0	-	-
CVAE [8]	ECCV'18	×	4096	22.5	33.3	0.5	0.1
DSN [64]	IJCAI'21	×	512	-	-	48.1	58.6
NAVE [31]	IJCAI'21	×	512	-	-	49.3	60.7
RPKD [65]	ACM MM'21	×	64†	37.1	48.5	36.1	49.1
			512	50.2	59.8	48.6	61.2
SBTKNet [30]	PR'22	×	512	50.2	59.6	48.0	60.8
ZSH [66]	ACM MM'16	✓	64†	-	-	13.9	17.4
SAE [67]	CVPR'17	✓	300	13.6	23.8	16.1	21.0
ZSIH [15]	CVPR'18	✓	64†	-	-	22.0	29.1
SEM-PCYC [26]	CVPR'19	✓	64†	-	-	29.3	39.2
			64	47.0	43.7	29.7	42.6
Doodle [33]	CVPR'19	✓	4096	46.1	37.0	10.9	-
SAKE [34]	ICCV'19	✓	64†	35.6	47.7	35.9	48.1
			512	49.7	59.8	47.5	59.9
StyleGuide [68]	TMM'20	✓	4096	35.8	40.0	25.4	35.5
AMDReg [69]	ECCV'20	✓	512	-	-	44.7	57.4
OCEAN [35]	ICME'20	✓	512	-	-	33.3	46.7
PDFD [36]	IJCAI'20	✓	512	-	-	48.3	60.0
AMF [38]	TIP'22	✓	512	**52.9**	47.7	19.6	28.6
TCN [37]	TPAMI'22	✓	64†	40.1	51.4	38.1	50.6
			512	51.6	**60.8**	49.5	61.6
ACNet (Ours)	-	×	64†	48.2	58.1	53.3	63.8
			512	51.7	**60.8**	**57.7**	**65.8**

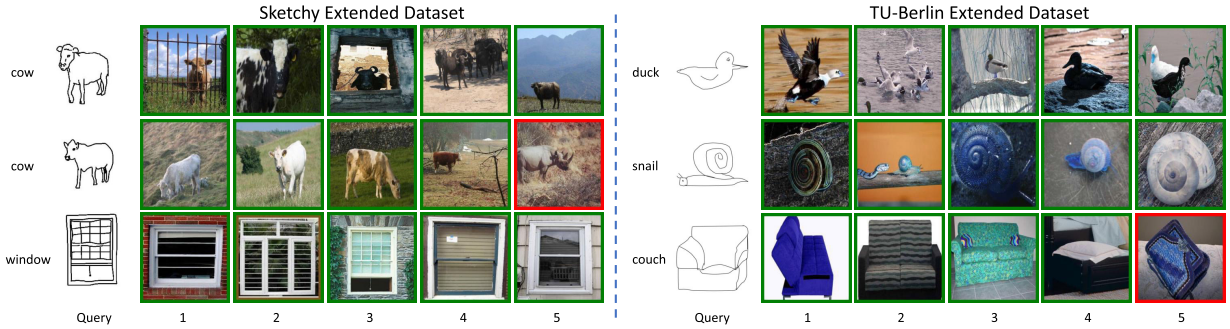


Fig. 5. Top-5 ZS-SBIR retrieval results (successful cases) from the proposed model (ResNet-50 backbone with 512 embedding dimension) on Sketchy Extended [8] and TU-Berlin Extended [15] datasets. Correct results are shown with a green border, while incorrect results are shown with a red border.

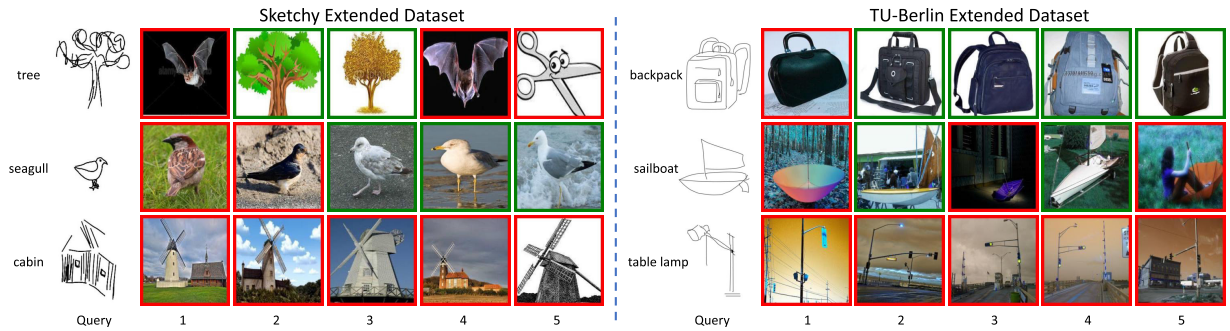


Fig. 6. Top-5 ZS-SBIR retrieval results (failure cases) from the proposed model (ResNet-50 backbone with 512 embedding dimension) on Sketchy Extended [8] and TU-Berlin Extended [15] datasets. Correct results are shown with a green border, while incorrect results are shown with a red border.

photos and the given sketches from the other failure cases too, which further demonstrates the effectiveness of our method in extracting features.

To verify the generality of our method, we also report the experimental results of the Generalized ZS-SBIR [35] (GZS-SBIR) task in Table VI, in which both seen and unseen categories are included in the gallery. It can be seen from the table that although our method does not aim at addressing the GZS-SBIR task, it still achieves comparable performance, and even obtains the best results on TU-Berlin Extended dataset. Different from StyleGuide [68], the proposed method does not require the class prior as the constraint for sketch-to-image

TABLE V

ABLATION STUDIES FOR THE PROPOSED METHOD ON SKETCHY EXTENDED [8] AND TU-BERLIN EXTENDED [15] DATASETS. WE ADOPT RESNET-50 AS OUR BACKBONE AND THE EMBEDDING DIMENSION IS 512. THE MODEL WITHOUT SYNTHESIS MODULE IS USED AS THE BASELINE (EXP 1). THE BEST RESULTS ARE BOLD

Exp	$\mathcal{L}_{norm}^{s_i}$	$\mathcal{L}_{norm}^{p_j}$	$\mathcal{L}_{adv}$	$\mathcal{L}_{ide}$	$\mathcal{L}_{norm}^{s_i^*}$	Sketchy Extended mAP @200	Sketchy Extended mAP @all	Sketchy Extended Prec @100	Sketchy Extended Prec @200	TU-Berlin Extended mAP @200	TU-Berlin Extended mAP @all	TU-Berlin Extended Prec @100	TU-Berlin Extended Prec @200
1	✓	✓	-	-	-	45.2	48.6	60.2	55.7	47.9	46.5	57.7	55.1
2	-	✓	✓	-	✓	47.8	52.7	60.9	57.1	53.4	54.3	61.9	60.3
3	-	✓	✓	✓	✓	48.4	53.0	61.4	57.6	54.3	54.9	62.9	61.3
4	✓	✓	✓	-	✓	49.7	53.7	62.2	58.7	56.9	56.6	65.2	63.5
5	✓	✓	✓	✓	✓	**51.7**	**55.9**	**64.3**	**60.8**	**57.7**	**57.7**	**65.8**	**64.4**

TABLE VI

OVERALL GZS-SBIR COMPARISON OF OUR METHOD AND OTHER APPROACHES ON SKETCHY EXTENDED [8] AND TU-BERLIN EXTENDED [15] DATASETS. "†"DENOTES RESULTS OBTAINED BY HASHING CODES, AND "-" MEANS THAT CORRESPONDING RESULTS ARE NOT REPORTED IN THE ORIGINAL PAPERS. THE BEST AND SECOND-BEST RESULTS ARE BOLD AND UNDERLINED, RESPECTIVELY

Methods	Venue	Semantic	Dim	Sketchy Extended mAP@all	Sketchy Extended Prec@100	TU-Berlin Extended mAP@all	TU-Berlin Extended Prec@100
ZSIH [15]	CVPR'18	✓	64†	21.9	29.6	14.2	21.8
SEM-PCYC [26]	CVPR'19	✓	64	30.7	36.4	19.2	29.8
StyleGuide [68]	TMM'20	✓	4096	33.1	38.1	21.5	29.1
OCEAN [35]	ICME'20	✓	512	-	-	31.2	34.1
AMF [38]	TIP'22	✓	512	**38.0**	**43.8**	16.2	23.8
ACNet (Ours)	-	×	512	28.2	38.4	**32.0**	**40.1**

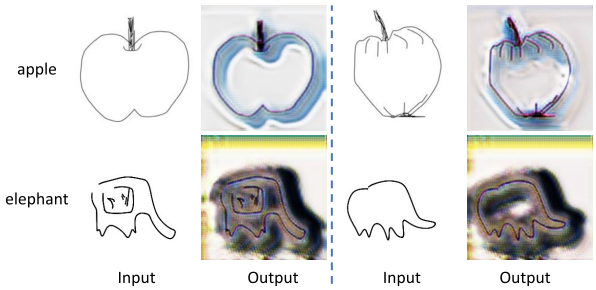


Fig. 7. The generated image of our sketch-to-photo generator does not look realistic, but it is made most beneficial to retrieve the corresponding photo images.

synthesis. Moreover, the proposed method designs the joint training for better utilizing the sketch-to-image synthesis to reduce the domain gap. We do not focus on generating high quality images and instead promote the ZS-SBIR performance. It is worth noting that our method does not use any semantic information as previous works. This also illustrates the advantage of our method.

E. Ablation Study

1) Effectiveness of $\mathcal{L}_{adv}$ and $\mathcal{L}_{norm}^{s_i^*}$: Considering the fact that $\mathcal{L}_{adv}$ and $\mathcal{L}_{norm}^{s_i^*}$ are inseparable, we explore the effectiveness of their integration. The joint training with $\mathcal{L}_{norm}^{s_i^*}$ could enable the gradients of the retrieval module to back-propagate to the synthesis module, which makes the synthesis better serve the retrieval module. Also, the adversarial loss $\mathcal{L}_{adv}$ could help generate photo-like images with high diversity and better align the sketch domain and the photo domain. We have achieved significant retrieval performance improvement as shown in Table V by comparing Exp 1&2 (from 46.5% to 54.3% in terms of $mAP@all$ on the Tu-Berlin Extended dataset) under the joint training manner with the GAN-based synthesis module.

2) Effectiveness of $\mathcal{L}_{ide}$: The generator G is designed mainly for mapping the sketch domain distribution to the photo domain. However, with only the adversarial loss constraint and the large domain shift between sketch images and photo images, G may generate some meaningless photo images randomly without the content constraint. The identity loss L_{ide} is inspired by CycleGAN, which designs the identity loss to better preserve the domain-agnostic feature representations. Please note that we only have the forward sketch-to-photo synthesis and the reconstruction of the photo images compared with CycleGAN. With the constraint of the pixel-level supervision from reconstructing the photo images, we observe marginal improvements by comparing Exp 2&3 (also Exp 4&5). The $mAP@all$ score increases from 54.3% to 54.9% (from 56.6% to 57.7%) and the $Prec@100$ score increases by 1% on TU-Berlin Extended dataset as shown in Table V.

3) Effectiveness of $\mathcal{L}_{norm}^{s_i}$: We removed $\mathcal{L}_{norm}^{s_i}$ to explore its influence. Without $\mathcal{L}_{norm}^{s_i}$, there is a slight retrieval performance drop by comparing Exp 3&5 (from 57.7% to 54.9% in terms of $mAP@all$ on TU-Berlin Extended dataset) as reported in Table V. The same phenomenon can be observed by comparing Exp 2&4. Since our goal is to better retrieve the photo images from the same category as the given sketch query, we set the optimization objective of the task to minimize the loss of retrieval ($\mathcal{L}_{norm}^{p_j}$ and $\mathcal{L}_{norm}^{s_i^*}$). However, due to the unstable training of the generator itself and the poor quality of the images generated in the early stage, the supervision signal provided by $\mathcal{L}_{norm}^{s_i^*}$ is not reliable. It makes the final model move towards the sub-optimal solution. By adding $\mathcal{L}_{norm}^{s_i}$, it can ensure that more stable supervisions could be provide during the optimization process, and reduce the impact of poor quality images on the retrieval performance.

4) Selection of λ and γ: In our final objective function (Equ. 8), there are two hyper-parameters for balancing the contributions of loss functions. To explore the sensitiveness of the proposed ACNet to the two hyper-parameters, we have designed three experiments using different combinations of λ and γ: ($\lambda = 0.1, \gamma = 10$), ($\lambda = 1.0, \gamma = 1.0$) and ($\lambda = 10, \gamma = 0.1$). The quantitative comparison is reported in Table VII. As observed, the proposed ACNet achieves the highest scores when $\lambda = 10$ and $\gamma = 0.1$.

5) Comparison With Other Loss Functions: In this paper, we have chosen the NormSoftmax loss to perform the challenging ZS-SBIR task considering its simplicity and powerful

TABLE VII

OVERALL COMPARISON OF OUR METHOD WITH DIFFERENT HYPER-PARAMETERS ON SKETCHY EXTENDED [8] AND TU-BERLIN EXTENDED [15] DATASETS. THE BEST RESULTS ARE BOLD

Exp	λ	γ	Sketchy Extended				TU-Berlin Extended			
			mAP @200	mAP @all	Prec @100	Prec @200	mAP @200	mAP @all	Prec @100	Prec @200
1	0.1	10	46.9	51.7	60.4	56.7	55.0	54.9	63.7	61.9
2	1.0	1.0	49.9	54.1	63.0	59.2	56.7	56.5	65.2	63.5
3	10	0.1	**51.7**	**55.9**	**64.3**	**60.8**	**57.7**	**57.7**	**65.8**	**64.4**

TABLE VIII

THE ZS-SBIR PERFORMANCE COMPARISON UNDER DIFFERENT LOSS FUNCTIONS ON SKETCHY EXTENDED [8] DATASET. THE BEST RESULTS ARE BOLD

Loss Name	mAP@200	mAP@all	Prec@100	Prec@200
Cross-Entropy	39.4	43.5	54.8	50.1
Margin	38.1	46.7	51.4	49.0
ProxyAnchor	49.6	52.7	62.1	58.7
NormSoftmax	**51.7**	**55.9**	**64.3**	**60.8**

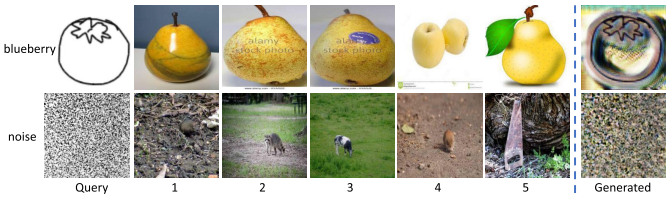


Fig. 8. Top-5 retrieval results from the proposed model (ResNet-50 backbone with 512 embedding dimension) on Sketchy Extended [8] dataset by given a query sketch that is not from the search set and a noise image. The generated images of our sketch-to-photo generator to the corresponding query images are shown in the last column.

ability. We conducted more experiments of selecting three different loss functions: one normal Cross-Entropy loss, one more pair-based loss (Margin loss [72]) and one more proxy-based loss (ProxyAnchor [73] loss). We report the experimental results in Table VIII. The experimental results show that the retrieval performance by using ProxyAnchor loss is much closer to NormSoftmax than Cross-Entropy loss and Margin loss, though they are inferior compared with NormSoftmax loss.

6) Effectiveness of G: Here we discuss the effectiveness of G. In this paper, we adopted the generator architecture of CycleGAN for generating photo-like images from the sketches. The CycleGAN adopts the symmetric structure for each domain, but our method only adopts a one-way generator (sketch-to-photo generator), because our goal is to use G as a feature intensifier to generate intermediate features. The essence of the designed generator is to generate more photo-like information for sketches so that the generated outputs can serve the subsequent retrieval, rather than making the generated image itself visually realistic. In fact, it does look a bit odd (not human-friendly, but algorithm-friendly) as shown in Fig. 7. This figure shows both the generated images of the well-drawn (the cases of "apple") and ill-drawn (the cases of "elephant") sketches. The goal is NOT to optimize both synthesis and retrieval. The synthesis module is optimized by the supervision from the retrieval module, which does not lead to visually realistic images.

For better understanding the behavior of our synthesis module and the whole system, we explore the scenarios when

TABLE IX

THE ZS-SBIR PERFORMANCE COMPARISON UNDER DIFFERENT GANS ON SKETCHY EXTENDED [8] DATASET. THE BEST RESULTS ARE BOLD

GAN Name	mAP@200	mAP@all	Prec@100	Prec@200
CyCleGAN	**51.7**	**55.9**	**64.3**	**60.8**
StarGAN v2	50.2	53.1	64.2	59.8

TABLE X

THE ZS-SBIR PERFORMANCE COMPARISON UNDER DIFFERENT IMAGE AUGMENTATION TECHNIQUES ON SKETCHY EXTENDED [8] DATASET. THE BEST RESULTS ARE BOLD

Exp	Color Jittering	Solarization	mAP @200	mAP @all	Prec @100	Prec @200
1	-	-	45.2	48.6	60.2	55.7
2	✓	-	46.1	50.4	60.5	56.3
3	-	✓	43.9	48.6	58.5	54.5
4	✓	✓	43.8	48.3	58.4	54.3
Ours			**51.7**	**55.9**	**64.3**	**60.8**

a sketch that is not from the search set and a noise image are provided as the input. The retrieval results and the generated images are shown in Fig. 8. For the given "blueberry" sketch (taken from the Quick, Draw! dataset [74]), which is not part of the Sketchy Extended [8] dataset, we can find the retrieved photos are all belonging to the category of "pear", which share the similar shape as the query sketch. This also reflects that our method can give intuitive and comprehensible retrieval results even when the given query images are not belonging to the datasets. And for the noise image, as the query image itself does not show any obvious features, there is no correlation between the retrieved photos. The retrieved photos are belonging to "mouse", "raccoon", "cow", "mouse" and "saw", respectively.

G is designed to be very lightweight in this paper. Designing it this way can better demonstrate that the proposed method does not depend on a sophisticated and heavy synthesis network. We are not trading the model complexity for better accuracy. We conducted more experiments on GANs by replacing CycleGAN with StarGAN v2 [75], a recent and more sophisticated network. The experimental results in Table IX show that using a new and complicated GAN does not really improve. From the results, we can find that the performance obtained by replacing CycleGAN with StarGAN v2 is similar, which also verifies that our method can be applied to different synthesis networks and also not so sensitive to the synthesis network architecture. Finally, we also argue that our goal is to promote retrieval performance rather than improve the naturalness and aesthetic quality of the synthesized samples.

7) Comparison With Data Augmentation: To further verify the effectiveness of our synthesis module, we conduct experiments on the Sketchy Extended dataset using different image augmentation techniques (*e.g.*, solarization and color jittering) rather than the synthesis augmentation. The results are reported in Table X. Applying color jittering could improve performance slightly, but it is much less effective than using the proposed sketch-to-photo synthesis. Solarization augmentation can even result in worse performance. The experimental results further confirm the effectiveness of our synthesis module.

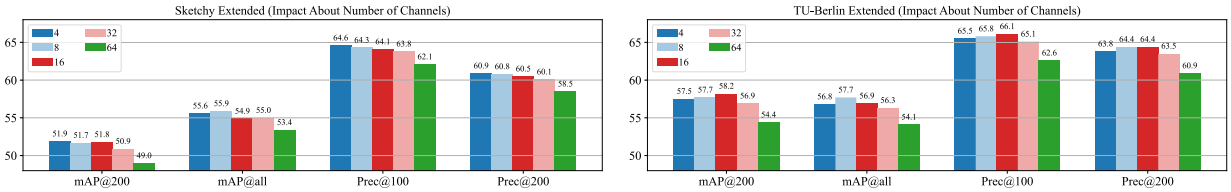


Fig. 9. Overall comparison of our method with different number of channels on Sketchy Extended [8] and TU-Berlin Extended [15] datasets.

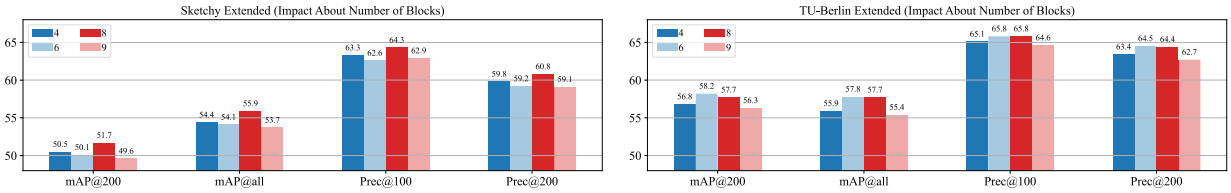


Fig. 10. Overall comparison of our method with different number of blocks on Sketchy Extended [8] and TU-Berlin Extended [15] datasets.

TABLE XI

OVERALL COMPARISON OF OUR METHOD WITH DIFFERENT BACKBONE NETWORKS AND EMBEDDING SIZES ON SKETCHY EXTENDED [8] AND TU-BERLIN EXTENDED [15] DATASETS. THE BEST RESULTS FOR EACH BACKBONE ARE BOLD

Backbone	Dim	Sketchy Extended				TU-Berlin Extended			
		mAP @200	mAP @all	Prec @100	Prec @200	mAP @200	mAP @all	Prec @100	Prec @200
VGG-16	64	32.6	38.0	48.7	44.7	39.8	37.1	50.6	48.0
	512	38.3	42.2	53.3	49.3	47.2	43.9	58.1	55.3
	4096	**40.0**	**43.2**	**54.6**	**50.8**	**51.7**	**47.9**	**62.3**	**59.3**
ResNet-50	64	43.0	46.0	56.8	52.7	47.5	44.9	57.2	54.9
	512	**51.7**	**55.9**	**64.3**	**60.8**	**57.7**	57.7	**65.8**	**64.4**
	4096	51.1	55.7	63.8	60.0	57.3	**58.6**	64.6	63.5

8) Different Architectures of G: We also explore the influences of choosing different architectures for G and D on the final ZS-SBIR results. We conduct the corresponding experiments in two ways. We first set the number of channels c to different values for both G and D. The quantitative results are reported in Fig. 9. We observe that we can achieve better results when using a lightweight architecture. We guess that it is more possible for the model with a bigger network to introduce noise and uncertainty to the downstream retrieval module. Later, we design different architectures for G by choosing different number of residual blocks and the results are reported in Fig. 10. An appropriate number (*e.g.*, 6 or 8) of residual blocks could achieve the best ZS-SBIR results.

9) Various Backbones and Embedding Sizes: In this section, we aim to explore the effectiveness and sensitivity of choosing various backbone networks: ResNet-50 [56] and VGG-16 [71]. We conduct the ZS-SBIR experiments on both the Sketchy Extended [8] and TU-Berlin Extended [15] datasets. All the quantitative results under various settings are reported in Table XI. With the same embedding size, we can obtain better results based on ResNet-50 network than VGG-16 network. The proposed ACNet could achieve a very impressive 58.6% $mAP@all$ score on TU-Berlin Extended dataset by choosing the ResNet-50 with a 4096 embedding size as the backbone network.

V. DISCUSSION

A. Limitations

When given a sketch query that is highly succinct and abstract in Fig. 12, it is extremely challenging to judge whether it is "window" or "door" due to the class ambiguity. Since the sketch-to-photo synthesis module of our method may still fail to generate desired reasonable images with high confidence, our method cannot handle this case well, either. We think that this ambiguous case could be resolved by introducing semantic information, *e.g.*, adding embedding information such as CLIP [78] and BERT [79] to guide image generation, which can be our future work.

B. Comparison With Unsupervised Domain Adaptation

1) Overview Comparison: We provide the setting comparison between the two Unsupervised Domain Adaptation (UDA) methods (UMDA [76] and CAPQ [77]) and our proposed method in Fig. 11. UMDA and CAPQ conduct the experiments under similar setting, in which the distribution shift comes from the difference between source dataset and target dataset. In detail, the experimental setting of UMDA and CAPQ is a less restrictive zero-shot setting, in which the training model has access to part of unlabelled data on the target domain. Different from UMDA and CAPQ, the proposed method mainly concerns about the ZS-SBIR task as described in Fig. 11, where the target data are not available during the whole training procedure. The knowledge gap (also a significant distribution shift) between the training and testing demands that the model should has a strong generalization ability to **unseen** testing data. Thus, the ZS-SBIR task is very different from the domain adaptation tasks mentioned in UMDA and CAPQ. And we cannot perform a fair comparison between these methods under the same experimental setting. These UDA methods **cannot** be directly introduced into the ZS-SBIR task.

2) Discussion About Information Difference Between Domains: Compared with UMDA, the information difference between the shop images and consumer images mainly rely on the background changes and thus is less than the information difference between sketch images and photo images in our ZS-SBIR task. As for CAPQ, the text description is a more abstract and semantic representation compared with the visual perception. The information difference presented in CAPQ is larger than the information difference in ZS-SBIR. However, we argue that the inter-class variation of ZS-SBIR is less

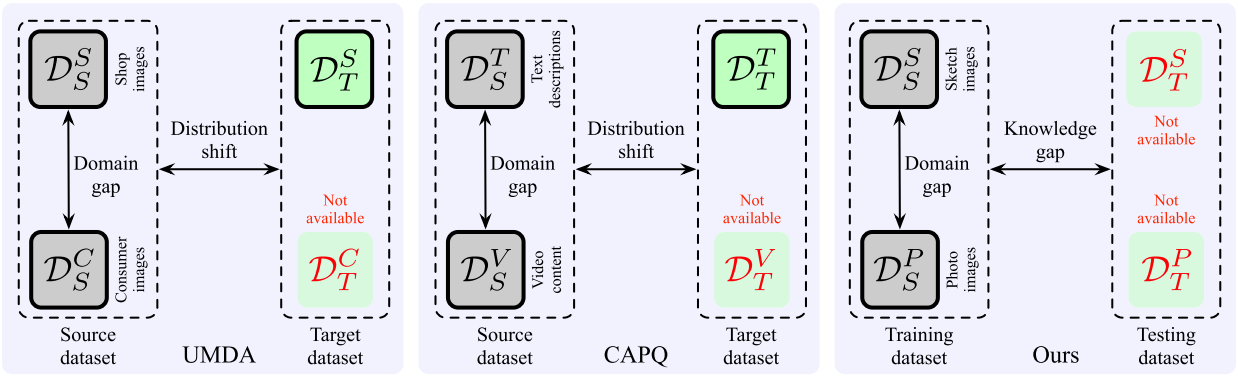


Fig. 11. The experimental setting comparison between UMDA [76], CAPQ [77] and our method.

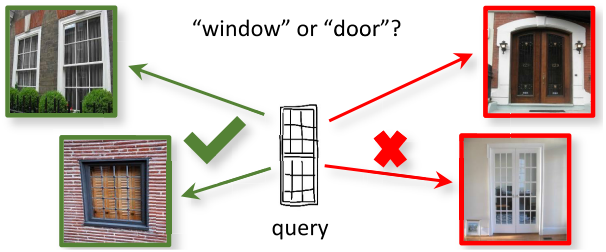


Fig. 12. Ambiguous case: the given sketch has the similar structure as "window" and "door."

than the difference among the activity events in the text-video retrieval. This requires that the trained model for ZS-SBIR should focus on the more fine-grained feature representations between classes. Furthermore, the zero-shot setting will also introduce further challenge for effective ZS-SBIR. How to transfer these methods to ZS-SBIR task can be an useful direction for future research.

VI. Conclusion

In this work, we proposed a novel, simple and effective joint synthesis-and-retrieval network called Approaching-and-Centralizing Network (ACNet) for Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR) and achieved state-of-the-art performance on Sketchy Extended [8] and TU-Berlin Extended [15] datasets. The proposed ACNet could effectively reduce both the domain gap and the knowledge gap by constantly generating samples with high diversity and centralizing the embeddings of both sketches and photos belonging to the same category. Our joint training framework provides valuable insight into how to integrate synthesis and other vision tasks.

References

[1] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "Sketch-based image retrieval: Benchmark and bag-of-features descriptors," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 11, pp. 1624–1636, Nov. 2011.

[2] R. Hu and J. Collomosse, "A performance evaluation of gradient field HOG descriptor for sketch based image retrieval," *Comput. Vis. Image Understand.*, vol. 117, no. 7, pp. 790–806, Jul. 2013.

[3] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao, "Deep sketch hashing: Fast free-hand sketch-based image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2862–2871.

[4] L. Wang, X. Qian, X. Zhang, and X. Hou, "Sketch-based image retrieval with multi-clustering re-ranking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4929–4943, Dec. 2020.

[5] L. Wang, X. Qian, Y. Zhang, J. Shen, and X. Cao, "Enhancing sketch-based image retrieval by CNN semantic re-ranking," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3330–3342, May 2020.

[6] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, "CrossATNet—A novel cross-attention based framework for sketch-based image retrieval," *Image Vis. Comput.*, vol. 104, Dec. 2020, Art. no. 104003.

[7] J. Li, Z. Ling, L. Niu, and L. Zhang, "Zero-shot sketch-based image retrieval with structure-aware asymmetric disentanglement," *Comput. Vis. Image Understand.*, vol. 218, Apr. 2022, Art. no. 103412.

[8] S. K. Yelamarthi, S. K. Reddy, A. Mishra, and A. Mittal, "A zero-shot framework for sketch based image retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 300–317.

[9] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[10] A. Zhai and H.-Y. Wu, "Classification is a strong baseline for deep metric learning," in *Proc. Brit. Mach. Vis. Conf.*, 2019, p. 91.

[11] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 1735–1742.

[12] W. Ge, "Deep metric learning with hierarchical triplet loss," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 269–285.

[13] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh, "No fuss distance metric learning using proxies," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 360–368.

[14] E. W. Teh, T. DeVries, and G. W. Taylor, "ProxyNCA++: Revisiting and revitalizing proxy neighborhood component analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 448–464.

[15] Y. Shen, L. Liu, F. Shen, and L. Shao, "Zero-shot sketch-image hashing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3598–3607.

[16] J. Lei, Y. Song, B. Peng, Z. Ma, L. Shao, and Y.-Z. Song, "Semi-heterogeneous three-way joint embedding network for sketch-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 3226–3237, Sep. 2020.

[17] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2Photo: Internet image montage," *ACM Trans. Graph.*, vol. 28, no. 5, pp. 1–10, Dec. 2006.

[18] H. Sun, J. Xu, J. Wang, Q. Qi, C. Ge, and J. Liao, "DLI-Net: Dual local interaction network for fine-grained sketch-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 7177–7189, Oct. 2022.

[19] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*.

[20] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5022–5030.

[21] P. Torres and J. M. Saavedra, "Compact and effective representations for sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2115–2123.

[22] A. Fuentes and J. M. Saavedra, "Sketch-QNet: A quadruplet ConvNet for color sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2134–2141.

[23] C. Hu and G. H. Lee, "Feature representation learning for unsupervised cross-domain image retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 529–544.

[24] P. Lu, G. Huang, H. Lin, W. Yang, G. Guo, and Y. Fu, "Domain-aware SE network for sketch-based image retrieval with multiplicative Euclidean margin softmax," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3418–3426.

[25] A. Sain, A. K. Bhunia, Y. Yang, T. Xiang, and Y.-Z. Song, "StyleMeUp: Towards style-agnostic sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8504–8513.

[26] A. Dutta and Z. Akata, "Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5089–5098.

[27] T. Dutta and S. Biswas, "Style-guided zero-shot sketch-based image retrieval," in *Proc. Brit. Mach. Vis. Conf.*, 2019, p. 9.

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

[29] Z. Zhang, Y. Zhang, R. Feng, T. Zhang, and W. Fan, "Zero-shot sketch-based image retrieval via graph convolution network," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12943–12950.

[30] O. Tursun, S. Denman, S. Sridharan, E. Goan, and C. Fookes, "An efficient framework for zero-shot sketch-based image retrieval," *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108528.

[31] W. Wang, Y. Shi, S. Chen, Q. Peng, F. Zheng, and X. You, "Norm-guided adaptive visual embedding for zero-shot sketch-based image retrieval," in *Proc. Thirtieth Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1106–1112.

[32] R. Liu, Q. Yu, and S. X. Yu, "Unsupervised sketch to photo synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 36–52.

[33] S. Dey, P. Riba, A. Dutta, J. L. Llados, and Y.-Z. Song, "Doodle to search: Practical zero-shot sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2179–2188.

[34] Q. Liu, L. Xie, H. Wang, and A. Yuille, "Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3662–3671.

[35] J. Zhu, X. Xu, F. Shen, R. K.-W. Lee, Z. Wang, and H. T. Shen, "Ocean: A dual learning approach for generalized zero-shot sketch-based image retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.

[36] X. Xu, M. Yang, Y. Yang, and H. Wang, "Progressive domain-independent feature decomposition network for zero-shot sketch-based image retrieval," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 984–990.

[37] H. Wang, C. Deng, T. Liu, and D. Tao, "Transferable coupled network for zero-shot sketch-based image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9181–9194, Dec. 2022.

[38] T. Jing, H. Xia, J. Hamm, and Z. Ding, "Augmented multimodality fusion for generalized zero-shot sketch-based visual retrieval," *IEEE Trans. Image Process.*, vol. 31, pp. 3657–3668, 2022.

[39] Y. Goldberg and O. Levy, "Word2Vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method," 2014, *arXiv:1402.3722*.

[40] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[41] M. Eitz, R. Richter, K. Hildebrand, T. Boubekeur, and M. Alexa, "Photosketcher: Interactive sketch-based image synthesis," *IEEE Comput. Graph. Appl.*, vol. 31, no. 6, pp. 56–66, Nov./Dec. 2011.

[42] X. Gao, N. Wang, D. Tao, and X. Li, "Face sketch–photo synthesis and retrieval using sparse representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 8, pp. 1213–1226, Aug. 2012.

[43] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 700–708.

[44] W. Chen and J. Hays, "SketchyGAN: Towards diverse and realistic sketch to image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9416–9425.

[45] Y. Zhong, Y. Qi, Y. Gryaditskaya, H. Zhang, and Y.-Z. Song, "Towards practical sketch-based 3D shape generation: The role of professional sketches," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 9, pp. 3518–3528, Sep. 2021.

[46] S.-Y. Wang, D. Bau, and J.-Y. Zhu, "Sketch your own GAN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14050–14060.

[47] P. Arbelaez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik, "Semantic segmentation using regions and parts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3378–3385.

[48] N. Gracias, M. Mahoor, S. Negahdaripour, and A. Gleason, "Fast image blending using watersheds and graph cuts," *Image Vis. Comput.*, vol. 27, no. 5, pp. 597–607, Apr. 2009.

[49] A. Ghosh et al., "Interactive sketch & fill: Multiclass sketch-to-image translation," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 1171–1180.

[50] L. S. F. Ribeiro, T. Bui, J. Collomosse, and M. Ponti, "Sketchformer: Transformer-based representation for sketched structure," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14153–14162.

[51] Z. Li, C. Deng, K. Wei, W. Liu, and D. Tao, "Learning semantic priors for texture-realistic sketch-to-image synthesis," *Neurocomputing*, vol. 464, pp. 130–140, Nov. 2021.

[52] J. Zhang et al., "Generative domain-migration hashing for sketch-to-image retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 297–314.

[53] V. K. Verma, A. Mishra, A. Mishra, and P. Rai, "Generative model for zero-shot sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 704–713.

[54] Z. Zheng, Y. Wu, X. Han, and J. Shi, "ForkGAN: Seeing into the rainy night," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 155–170.

[55] A. K. Bhunia, P. N. Chowdhury, A. Sain, Y. Yang, T. Xiang, and Y.-Z. Song, "More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4247–4256.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[57] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: Learning to retrieve badly drawn bunnies," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–12, Jul. 2016.

[58] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[59] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–10, 2012.

[60] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[62] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.

[63] M. Boudiaf et al., "A unifying mutual information view of metric learning: Cross-entropy vs. pairwise losses," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 548–564.

[64] Z. Wang, H. Wang, J. Yan, A. Wu, and C. Deng, "Domain-smoothing network for zero-shot sketch-based image retrieval," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1143–1149.

[65] J. Tian, X. Xu, Z. Wang, F. Shen, and X. Liu, "Relationship-preserving knowledge distillation for zero-shot sketch based image retrieval," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 5473–5481.

[66] Y. Yang, Y. Luo, W. Chen, F. Shen, J. Shao, and H. T. Shen, "Zero-shot hashing via transferring supervised knowledge," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 1286–1295.

[67] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3174–3183.

[68] T. Dutta, A. Singh, and S. Biswas, "StyleGuide: Zero-shot sketch-based image retrieval using style-guided image generation," *IEEE Trans. Multimedia*, vol. 23, pp. 2833–2842, 2021.

[69] T. Dutta, A. Singh, and S. Biswas, "Adaptive margin diversity regularizer for handling data imbalance in zero-shot SBIR," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 349–364.

[70] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 172–189.

[71] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–13.

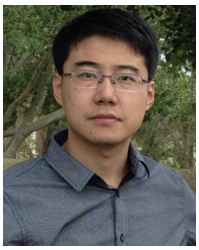
[72] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, "Sampling matters in deep embedding learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2840–2848.
[73] S. Kim, D. Kim, M. Cho, and S. Kwak, "Proxy anchor loss for deep metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3238–3247.
[74] D. Ha and D. Eck, "A neural representation of sketch drawings," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–12.
[75] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN V2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8188–8197.
[76] V. Sharma, N. Murray, D. Larlus, M. S. Sarfraz, R. Stiefelhagen, and G. Csurka, "Unsupervised meta-domain adaptation for fashion retrieval," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1348–1357.
[77] Q. Chen, Y. Liu, and S. Albanie, "Mind-the-gap! Unsupervised domain adaptation for text-video retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 2, pp. 1072–1080.
[78] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[79] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

Hao Ren received the B.Eng. degree in software engineering from the Zhejiang University of Technology in 2017. He is currently with the Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai, China. His research interests include pattern recognition and computer vision.

Ziqiang Zheng received the B.Eng. degree in communication engineering from the Ocean University of China in 2019. He is currently with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology (HKUST), Hong Kong, China. His research interests include generative adversarial networks and computer vision.

Yang Wu (Member, IEEE) received the B.S. and Ph.D. degrees from Xi'an Jiaotong University in 2004 and 2010, respectively. From 2011 to 2014, he was a Program-Specific Researcher with the Academic Center for Computing and Media Studies, Kyoto University. From December 2014 to June 2019, he was an Assistant Professor with the NAIST International Collaborative Laboratory for Robotics Vision, Nara Institute of Science and Technology (NAIST). From July 2019 to May 2021, he was a Program-Specific Senior Lecturer with the Department of Intelligence Science and Technology, Kyoto University. He is currently a Principal Researcher with Tencent AI Lab. His research interests include computer vision, pattern recognition, and multimedia content analysis, enhancement, and generation.

Hong Lu (Member, IEEE) received the B.Eng. and M.Eng. degrees in computer science and technology from Xidian University, Xi'an, China, in 1993 and 1998, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2005. From 1993 to 2000, she was a Lecturer and a Researcher with the School of Computer Science and Technology, Xidian University. From 2000 to 2003, she was a Research Student with the School of Electrical and Electronic Engineering, Nanyang Technological University. Since 2004, she has been with the School of Computer Science, Fudan University, Shanghai, China, where she is currently a Professor. Her current research interests include computer vision, machine learning, pattern recognition, and robotic tasks.



Yang Yang (Senior Member, IEEE) received the bachelor's degree in computer science from Jilin University, Changchun, China, in 2006, the master's degree in computer science from Peking University, Beijing, China, in 2009, and the Ph.D. degree in computer science from The University of Queensland, Brisbane, Australia, in 2012. He is currently with the University of Electronic Science and Technology of China, Chengdu, China. His current research interests include multimedia content analysis, computer vision, and social media analytics.

Ying Shan (Senior Member, IEEE) is currently a Distinguished Scientist with Tencent and the Director of the ARC Lab, Tencent PCG. Before joining Tencent, he was a Post-Doctoral Researcher with Microsoft Research, a Senior MTS with SRI International (Sarnoff Subsidiary), and a Principal Scientist Manager with Microsoft Bing Ads. He is leading the research and development efforts in web search and content AI for a suite of social media and content distribution products. He has published more than 70 papers in top conferences and journals in the areas of computer vision, machine learning, and data mining. He holds a number of U.S./international patents. He has served as a Senior PC Member for KDD. He has served as the Area Chair for CVPR.

Sai-Kit Yeung received the B.Eng. degree (Hons.), the M.Phil. degree in bioengineering, and the Ph.D. degree in electronic and computer engineering from The Hong Kong University of Science and Technology (HKUST), Hong Kong, China, in 2003, 2005, and 2009, respectively. He is currently an Associate Professor with the Division of Integrative Systems and Design (ISD), HKUST. Before joining HKUST, he was an Assistant Professor with the Singapore University of Technology and Design (SUTD) and founded the Vision, Graphics and Computational Design (VGD) Group. During his time with SUTD, he was also a Visiting Assistant Professor with Stanford University and MIT. Prior to that, he was a Post-Doctoral Scholar with the Department of Mathematics, University of California at Los Angeles (UCLA), Los Angeles, CA, USA. He was a Visiting Student with the Image Processing Research Group, UCLA, in 2008, and the Image Sciences Institute, University Medical Center Utrecht, The Netherlands, in 2007.