Zero-Shot Sketch Based Image Retrieval via Modality Capacity Guidance

Yanghong Zhou[3,1] , **Dawei Liu**[1] , **P. Y. Mok**[1,2,3,*]

[1]School of Fashion and Textiles, The Hong Kong Polytechnic University
[2]Research Institute for Intelligent Wearable Systems, The Hong Kong Polytechnic University
[3]Research Centre of Textiles for Future Fashion, The Hong Kong Polytechnic University
yanghong.zhou@connect.polyu.hk, {dawei.liu, tracy.mok}@polyu.edu.hk

Abstract

Zero-shot sketch-based image retrieval (ZS-SBIR), aiming to recognize and retrieve relevant photos based on freehand sketch queries that belong to unseen categories in the search set, has sparked considerable interest, benefiting from the rapid advancements in multimodal learning and feature representation research. Despite the recent improvements in performance, there are still rooms for refining feature representation and thus enhancing the generalization capabilities of the models. Most of the existing research efforts have primarily focused on learning the feature distribution of modalities within specific datasets, without considering the broader dataset-agnostic 'population distribution' of relevant modalities. In this paper, we investigate the modality population distribution and apply such knowledge to guide feature learning. Specifically, we propose a modality capacity constraint loss to control the learning of population distribution for sketches and photos. This loss can be effectively combined with retrieval loss (e.g., triplet loss) or classification loss (e.g., InfoNCE loss) to enhance the performance of ZS-SBIR, through the fine-tuning process of pre-trained models like CLIP and DINO. Extensive experiment results have demonstrated our significant performance improvements, achieving an increase of 7.3%/3.2% and 19.9%/10.3% in terms of mAP@200/P@200 compared to the state-of-the-art models on CLIP and DINO, respectively, on the Sketchy-ext dataset (split 2). Data, code, and supplementary information are available at https://github.com/YHdian0716/ZS-SBIR-MCC.git

1 Introduction

Zero-shot learning (ZSL) aims to recognize and categorize objects or classes that have never been seen before in the training process. As a specific application of ZSL, zero-shot sketch-based image retrieval (ZS-SBIR) involves a combination of ZSL and SBIR, aiming to retrieve photos by inputting query sketches of categories not seen in the training set. Benefit from the rapid advancements of self-supervised models (e.g., DINO) and large language-vision models (e.g., CLIP), the performance of ZSL-SBIR has improved by large margins. These models have been all trained on large-scale datasets, equiping them with generalized vision features, which play a vital role in ZSL-SBIR. For instance, based on DINO, some studies [Tian *et al.*, 2022; Wang *et al.*, 2022; Tian *et al.*, 2023] leveraged knowledge distillation to enhance the model's ability to generalize to unseen classes and improve its discriminability to distinguish instances and categories. Other studies attemp to adapt CLIP for SBIR, leveraging its vision-semantic alignment capability for effective knowledge transfer. For example, [Sain *et al.*, 2023a] and [Dong *et al.*, 2023] proposed to use prompt learning and adapters to adapt CLIP text encoder to ZSL-SBIR while preserving its generalizability.

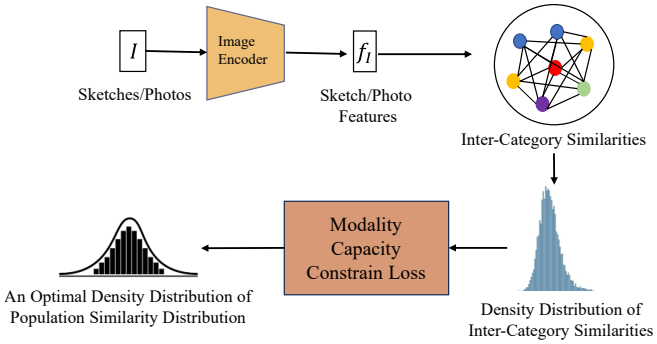

Figure 1: Illustration of our proposed method.

In the early work of image retrieval, statistical knowledge, such as distribution characteristics, is often utilized to design hand-crafted features [Stricker and Swain, 1994; Brunelli and Mich, 2001]. Such overall distribution characteristics, however, have been largely overlooked in recent deep learning-based model training. We argue that a good understanding of representation statistics is essential for model learning too. In the context of ZS-SBIR, existing loss functions, such as triplet loss and InfoNCE loss, focus solely on the relationship between individual samples without considering the overall distribution characteristics. To address this issue, we introduce in this paper a new concept of *modality*

*Corresponding author: tracy.mok@polyu.edu.hk

capacity and design a new loss according to the law of large numbers (LLN) to guide the feature learning process. More specifically, the modality capacity examines the effectiveness of a representation space (see Figure 1) by calculating the averaged inter-class similarity among samples within the same modality. With the modality capacity constraint loss defined, we can drive the modality capacity towards an optimal state, resulting in more balanced decision boundaries. This assists the model to self adjusted to learning more robust and generalizes, in particularly when inter-class similarity is either very small or too large, where existing models fail to learn discriminative representations or the learning is biased towards a specific category. Therefore, our method can facilitate representation learning for different modalities and, finally, improve the performance of ZS-SBIR. The main contribution of this paper can be summarized:

- To our best knowledge, our work is the first to consider the effectiveness of population during the ZS-SBIR learning process.
- A modality capacity is introduced to measure the effectiveness of a representation space for retrieval.
- A novel modality capacity constraint loss is designed to guide the model to learn more robust and generalized representations by constraining the modality capacity.
- Without introducing extra parameters, our proposed method, when combined with triplet/InfoNCE loss, obtains an increase of $2.4\%/1.8\%$ and $24.0\%/2.5\%$ respectively in terms of mAP@200 compared to the baselines of CLIP and DINO.

2 Related Work

Zero shot learning. Zero-shot learning (ZSL) is a method that enables a model to classify data samples, whose classes are not present during training. Most ZSL approaches leverage semantic information (e.g., manually defined attributes [Luo *et al.*, 2020] or word vectors [Akata *et al.*, 2016]) to bridge the gap between both seen and unseen classes. These works usually employ an embedding or mapping function to establish a connection between the low-level visual features and their respective semantic vectors. They can be presented as 'visual $\rightarrow$ semantics' (e.g., [Chen *et al.*, 2018]), 'semantics $\rightarrow$ visual' (e.g., [Shigeto *et al.*, 2015]) or 'visual $\rightarrow$ latent $\leftarrow$ semantics' (e.g., [Zhang *et al.*, 2017]). Instead, we believe that models like CLIP and DINO, which are pre-trained on large-scale datasets, possess a comprehensive understanding of vision and can achieve ZS-SBIR when appropriate guidance is provided. In short, we focus on the relationship between 'vision (sketch)' and 'vision (photo)'.

ZS-SBIR. When combining ZSL and SIBR, research studies can be classified as vision-centric, vision-semantic-alignment, or a combination of both. Vision-centric works often focus on properties within the vision domain. [Wang *et al.*, 2021b] proposed using an image bank to bridge the domain gap. [Wang *et al.*, 2021a] utilized vision feature norm to guide the sketch and photo alignment. Visual image generation is also a promising and widely-considered sub-direction due to the success of GANs and their variants. [Ren *et al.*, 2023] utilized CycleGAN to enhance the knowledge of feature extractor to further improve the generalization ability of ZS-SBIR. [Dutta and Akata, 2019] proposed a cycle-consistent generator leveraging semantic information for better knowledge about the unseen classes. Recently, several studies have demonstrated the superior performance of combining vision-related semantics with ViT models pre-trained on large-scale datasets. For example, [Dong *et al.*, 2023] leveraged the CLIP text encoder and incorporated adaptors into DINO, resulting in competitive performance. [Sain *et al.*, 2023a] proposed an approach that fine-tunes CLIP using a prompt learning mechanism. This enables the integration of vision-semantic knowledge into the ZS-SBIR problem. However, none of these methods considered high-level population distribution information, potentially resulting in a bised learning that is unsuitable for unexplored categories.

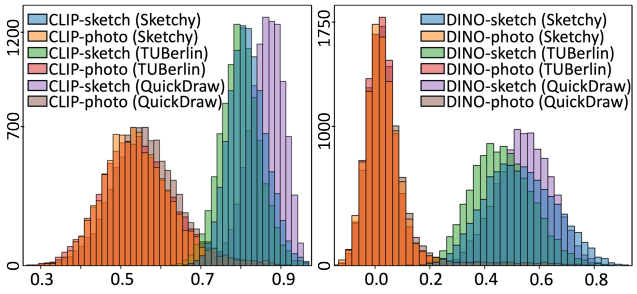

Figure 2: Inter-category similarity histogram investigation (according to Table 1). Left: CLIP-Based. Right: DINO-based. The results are obtained through 10k samplings.

Dataset \ Capacity	$\mathcal{D}_p$	$\mathcal{D}_s$
Sketchy-Ext	0.73 $\pm$ 0.08	0.91 $\pm$ 0.03
TUBerlin-Ext	0.77 $\pm$ 0.07	0.89 $\pm$ 0.04
QuickDraw-Ext	0.81 $\pm$ 0.07	0.94 $\pm$ 0.02
Sketchy-Ext	0.07 $\pm$ 0.12	0.77 $\pm$ 0.09
TUBerlin-Ext	0.09 $\pm$ 0.11	0.67 $\pm$ 0.09
QuickDraw-Ext	0.15 $\pm$ 0.13	0.80 $\pm$ 0.08

Table 1: 'mean $\pm$ std' of the inter-category similarities obtained from the pre-trained CLIP (top) and DINO (bottom). The results are calculated using category representations (mean feature value).

3 Method

In this section, we first describe our problem formulation and the baseline. By investigating population distribution of the learned features of the baseline model, a novel metric – modality capacity contraint loss – is then introduced and integrated with existing retrieval losses for a more effective representation learning.

3.1 Problem Formulation

In ZS-SBIR, the dataset consists of two types of samples: sketches P and photos S, and is denoted as $D = \{(img_i, y_i)|y_i \in C\}_{i=1}^{N}$, where each sketch/photo img_i is associated with a corresponding label y_i. The sketches and photos are divided into two subsets $D^{seen} = \{P^{seen}, S^{seen}\}$

and $D^{unseen} = \{P^{unseen}, S^{unseen}\}$ for training and testing. More specifically, $D^{seen} = \{(img_i, y_i)|y_i \in C^{seen}\}_{i=1}^{N^s}$, where $N^s = |P^{seen}| + |S^{seen}|$, $img_i \in P^{seen} \bigcup S^{seen}$ and C^{seen} is the seen category set. Similarly, $D^{unseen} = \{(img_i, y_i)|y_i \in C^{unseen}\}_{i=1}^{N^u}$, where $N^u = |P^{unseen}| + |S^{unseen}|$, $img_i \in P^{unseen} \bigcup S^{unseen}$ and C^{unseen} is the unseen category set. There is no overlap between C^{seen} and C^{unseen}, that is $C^{seen} \cap C^{unseen} = \emptyset$. Our goal is to train a model using the seen data D^{seen} to retrieve photos of the same category for a query sketch from the unseen data S^{unseen}.

3.2 Baseline ZS-SBIR Framework

We here provide an overview of our baseline framework for ZS-SBIR. Given a sketch/photo pair, a sketch/photo feature extractor is utilized to obtain the vision feature $f_s = \mathcal{F}_s(s)$ and $f_p = \mathcal{F}_p(p)$, where f_s, $f_p \in \mathbf{R^c}$ are the sketch and photo features, respectively, and c is the feature dimension. To establish our baseline, we employ either DINO or CLIP as the vision extractor, which is trained with different loss functions based on the respective characteristics of the extractor.

DINO-based. DINO is trained on large-scale image datasets in a self-supervised manner, to learn discriminative visual representations. Recent studies have shown that using the visual representations directly as a common space yields good results in ZS-SBIR. As a result, we employ the triplet loss to train the model based on DINO in the visual space, without leveraging semantic information.

During training, we sample a triplet consisting of a sketch anchor s, a positive photo p^+ and a negative photo p^-. The objective of the triplet loss $\mathcal{L}^{tri}$ is to minimize the distance $d(s, p^+)$ between s and p^+, while simultaneously maximizing the distance $d(s, p^-)$ between s and p^+. Here, $d(\cdot, \cdot)$ is measured using the cosine distance. The calculation of the triplet loss can be represented as:

$$\mathcal{L}^{tri} = max\{0, \mu + d(s, p^+) - d(s, p^-)\} \tag{1}$$

where μ is a hyperparameter that defines the minimum desired margin between the positive and negative samples.

CLIP-based. CLIP learns joint representations of image and text, which in turn facilitates zero-shot learning. We adopt CLIP-AT [Sain *et al.*, 2023b] and leverage text embedding to guide the feature learning by incorporating a classification loss as follows:

$$\mathcal{L}^{cls} = \frac{1}{N}\sum_{i=1}^{N} -\log P(y_i|p_i) - \log P(y_i|s_i) \tag{2}$$

$$P(y_i|I) = \frac{\exp(sim(f_I, f_t^{y_i})/\tau)}{\sum_{j=1}^{|C^{seen}|} exp(sim(f_I, f_t^j)/\tau)} \tag{3}$$

where $I \in \{s, p\}$ represents a sketch or photo, $sim(\cdot)$ denotes the cosine similarity function, and τ is the temperature hyperparameter. In addition, several recent studies (e.g., [Sain *et al.*, 2023a]) have demonstrated the effectiveness of combining $\mathcal{L}^{tri}$ (Eq. 1) with CLIP. Therefore, we also adopt this setting as one of our baselines. In sum, a total of three baselines are adopted based on DINO and CLIP with triplet or classification loss.

	mAP@200	P@200	$\mathcal{D}_p$	$\mathcal{D}_s$
CLIP	0.372	0.320	0.91	0.73
CLIP w/ $\mathcal{L}^{mcc}$	**0.506**	**0.433**	**0.42**	**0.22**

Table 2: Effectiveness of D_I in mAP@200 and P@200.

3.3 Population Similarity Investigation

To evaluate the effectiveness of representations in the image retrieval task, the commonly used metrics are P@k and mAP@k. P@k is the proportion of retrieved top-k images that are from the same category. The mAP@k is the mean AP@k of all queris where the AP@k of each query is computed by $AP@k = \sum_{i=1}^{k} \frac{P@i \times \gamma(i)}{N}$ with $\gamma(i) = 1$, if the i-th ranked image is from the same category as that of the input query sketch, otherwise, $\gamma(i) = 0$, and N is the total number of relevant images. These metrics provide a comprehensive performance assessment for retrieval system. However, they are non-differentiable and can not be directly interpreted as a loss function to guide the representation learning through training. To address this issue, we propose a modality capacity to measure the effectiveness of representations. Taking into account inter-category dissimilarities, we calculate the modality capacity as the averaged similarity value of all negative pairs belonging to different categories:

$$\mathcal{D}_I = (\frac{1}{m_I}\sum_{j=1}^{N}\sum_{k\neq j}^{N} \mathbb{1}_{c(I_j)\neq c(I_k)} \cdot d(f_j, f_k)) \tag{4}$$

where $I \in \{s, p\}$, s and p represent sketch and photo modality, N is the data size and $c(\cdot)$ represents a category of a sketch or photo. m_I is a normalization factor, counting the valid number of (s, s) (or (p, p)) pairs with different categories:

$$m_I = \sum_{j=1}^{N}\sum_{k\neq j}^{N} \mathbb{1}_{c(I_j)\neq c(I_k)} \tag{5}$$

Table 1 shows the calculated modality capacities of sketch and photo in three datasets using CLIP and DINO, respectively, while Figure 2 illustrated the corresponding histograms. It can be shown that the mean of inter-category similarity on sketches is higher than that on photos. The mean of inter-category similarity based on CLIP is higher than that based on DINO while the standard deviation is larger, indicating that DINO learns more discriminative visual representations than CLIP but less stable. It shows that the distributions based on the same baseline architecture are quite similar across different datasets. Hence, we can find out a optimal estimation of population similarity to guide the model learning regardless the datasets used for a modality. Table 2 shows the values of mAP@all and P@200 in the sketch to image retrieval task. As shown, the lower the value of modality capacity indicates the better the image retrieval performance.

3.4 Modality Capacity Constraint Loss

The modality capacity constraint loss is defined, aiming to utilize the proposed modality capacity to guide the feature

Methods	S	Vis-Enc	Sketchy-Ext Split 1		Sketchy-Ext Split 2		TU-Berlin		QuickDraw	
			mAP@all	P@100	mAP@200	P@200	mAP@all	P@100	mAP@all	P@200
ZS-CAAE (ECCV'18)	-	VGG-16	-	-	0.156	0.260	0.005	0.003	-	-
ZS-CVAE (ECCV'18)	-	VGG-16	-	-	0.225	0.333	0.005	0.001	0.003	0.003
ZS-CCGAN (CVPR'19)	✓	VGG-16	0.349	0.463	-	-	0.297	0.426	-	-
ZS-GRL (CVPR'19)	✓	VGG-16	-	-	0.369	0.370	0.110	0.121	0.075	0.068
ZS-IIAE (NIPS'20)	-	VGG-16	0.573	0.659	0.373	0.485	0.412	0.503	-	-
ZS-Sketch3T (CVPR'22)	-	VGG-16	-	-	0.579	0.648	0.507	0.671	-	-
ZS-SAKE (ICCV'19)	✓	ResNet50	0.547	0.692	0.497	0.598	0.475	0.599	-	-
ZS-GCN (AAAI'20)	✓	ResNet50	0.382	0.538	0.568	0.487	0.110	0.121	-	-
ZS-TCN (TPAMI'21)	✓	ResNet50	0.616	0.763	0.516	0.608	0.495	0.616	0.140	0.231
ACNet (TCSVT'23)	-	ResNet50	-	-	0.517	0.608	0.577	0.658	-	-
SBTKNet (PR'22)	-	ResNet50	0.553	0.698	0.502	0.596	0.480	0.608	0.119	0.167
NAVE (IJCAI'21)	-	ResNet50	0.613	0.725	-	-	0.493	0.607	-	-
ZSE (CVPR'23)	-	ViT	0.736	0.808	0.504	0.602	0.569	0.637	0.142	0.202
ZS-PSKD (MM'22)	-	DINO-ViT	0.671	0.762	0.535	0.630	0.495	0.608	0.148	0.212
ZS-TVT (AAAI'22)	-	DINO-ViT	0.648	0.796	0.531	0.618	0.484	0.662	0.149	0.293
Sherry‡ (arxiv'23)	✓	DINO-ViT	0.741	0.835	0.616	0.695	0.541	0.664	-	-
CLIP-AT (CVPR'23)	✓	CLIP-ViT	-	-	0.723	0.725	0.651	0.732	0.202	0.388
Ours w/ $\mathcal{L}^{cls+mcc}$	✓	CLIP-ViT	0.771	0.841	0.782	0.747	**0.668**	**0.773**	**0.332**	**0.436**
Ours w/ $\mathcal{L}^{tri+mcc}$	✓	CLIP-ViT	0.784	0.842	0.790	0.757	0.642	0.750	0.314	0.402
Ours w/ $\mathcal{L}^{tri+mcc}$	✓	DINO-ViT	**0.817**	**0.875**	**0.805**	**0.768**	0.636	0.722	0.222	0.317

Table 3: Quantitative comparison. 'S' and 'Vis-Enc' means 'Semantic' and 'Visual-Encoder', respectively.

learning of the model in the training process. To this end, we calculate the modality capacity for each batch of data and minimize this modality capacity to a certain value γ_I. The modality capacity constraint loss is given as:

$$\mathcal{L}_I^{mcc} = |(\frac{1}{m_I}\sum_{j=1}^{B}\sum_{k\neq j}^{B}\mathbb{1}_{c(I_j)\neq c(I_k)}\cdot d(I_j,I_k)) - \gamma_I| \quad (6)$$

where B is the batch size. According to the law of large numbers (LLN), when the sampling size is sufficiently large, the feature capacity on the whole dataset can also be minimized to the value γ_I.

γ_I is a hyperparameter controlling the target modality capacity value within the population I, i.e., sketch or photo. For the representation learning of different modalities, the γ_I should be set differently. However, as γ_I involves the estimation of the general modality capacity for different modalities, once this value has been found on a large dataset for a modality, it can be directly applied to other datasets of the same modality.

Total loss. Our proposed modality capacity constraint loss can be applied to different baselines for ZS-SBIR. The total loss based on DINO and CLIP is respectively as:

$$\mathcal{L}^{tri+mcc} = \lambda_1\cdot\mathcal{L}^{tri} + \lambda_2\cdot\mathcal{L}_s^{mcc} + \lambda_3\cdot\mathcal{L}_p^{mcc} \quad (7)$$

$$\mathcal{L}^{cls+mcc} = \lambda_4\cdot\mathcal{L}^{cls} + \lambda_5\cdot\mathcal{L}_s^{mcc} + \lambda_6\cdot\mathcal{L}_p^{mcc} \quad (8)$$

where $\lambda_{1,2,3,4,5,6}$ are the hyperparameters, $\mathcal{L}^{tri}$ is the triplet loss and $\mathcal{L}^{cls}$ is the classification (or InfoNCE) loss (Eq. 2).

	γ_s	γ_p
Ours CLIP-ViT w/ $\mathcal{L}^{cls+mcc}$	0.2	0.0
Ours CLIP-ViT w/ $\mathcal{L}^{tri+mcc}$	0.1	0.0
Ours DINO-ViT w/ $\mathcal{L}^{tri+mcc}$	0.0	0.0

Table 4: Hyperparameter settings.

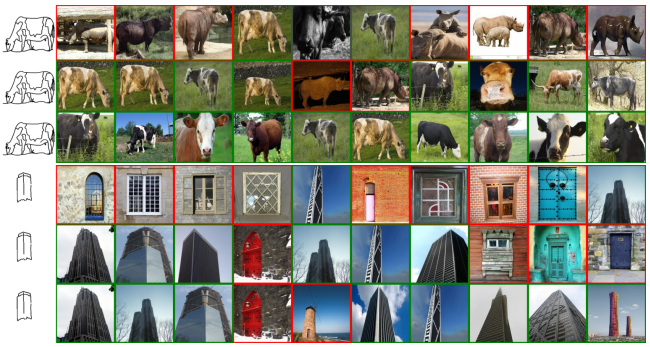

Figure 3: Qualitative comparison on the Sketchy-Ext dataset.

4 Experiments

4.1 Datasets and Experimental Settings

Dataset. We evaluated the effectiveness of our proposed modality capacity constraint loss on there widely-used benchmarks: Sketchy-Ext [Liu *et al.*, 2017], TUBerlin-Ext [Eitz *et al.*, 2012] and a subset of QuickDraw-Ext [Dey *et al.*, 2019a]. Sketch Ext. [Liu *et al.*, 2017] contains 75,471 sketches and 60,502 photos, with a total of 125 categories. The TUBer-

$\mathcal{L}^{cls}$	$\mathcal{L}_s^{mcc}$	$\mathcal{L}_p^{mcc}$	mAP@200	P@200
✓	✗	✗	0.764	0.730
✓	✓	✗	0.770	0.735
✓	✗	✓	0.745	0.711
✓	✓	✓	0.782	0.747

Table 5: Ablation study of $\mathcal{L}_s^{mcc}$ and $\mathcal{L}_p^{mcc}$ using 'Ours (CLIP-ViT) w/ $\mathcal{L}^{cls}$' on Sketchy-Ext dataset (split 2).

$\mathcal{L}^{tri}$	$\mathcal{L}_s^{mcc}$	$\mathcal{L}_p^{mcc}$	mAP@200	P@200
✓	✗	✗	0.766	0.734
✓	✓	✗	0.786	0.750
✓	✗	✓	0.749	0.716
✓	✓	✓	0.790	0.757

Table 6: Ablation study of $\mathcal{L}_s^{mcc}$ and $\mathcal{L}_p^{mcc}$ using 'Ours (CLIP-ViT) w/ $\mathcal{L}^{tri}$' on Sketchy-Ext dataset (split 2).

lin dataset [Eitz *et al.*, 2012] contains 204,489 photos and 40,000 free-hand sketches with 250 categories of 80 free-hand sketches each. QuickDraw Ext. [Dey *et al.*, 2019a] contains 50 million sketches over 345 categories. Following [Dey *et al.*, 2019b], we utilized a subset of QuickDraw Ext. [Dey *et al.*, 2019a] composing of 330,000 sketches, 204,000 photos and 110 categories for evaluation. For data partitioning, we also followed [Dey *et al.*, 2019b] to divide Sketch Ext. [Liu *et al.*, 2017] into 100/104 categories for training and 25/21 categories for testing, which are denoted as 'Sketchy Ext Split 1' and 'Sketchy Ext Split 2', respectively. We utilized 25 categories from Sketch Ext. [Liu *et al.*, 2017] and 30 categories from TUBerlin-Ext [Eitz *et al.*, 2012] for testing and utilized the rest 100/220 categories for training. Table 4 lists the values of our γ_s and γ_p for various proposed methods.

Implementation details. All the experiments were conducted on Pytorch with 11GB Nvidia RTX 3080-Ti GPU. We used Adam optimizer to train the models with learning rates of $lr = 1e-4$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The input size of the images was 224×224. The models were trained for 60 epochs with a batch size of 64. During the training stage, all the parameters of the models were frozen except for the layer normalization. The loss weights were set as $\lambda_1 = \lambda_4 = 1$, $\lambda_2 = \lambda_5 = 4$ and $\lambda_3 = \lambda_6 = 8$. The margin μ was set as 0.3.

4.2 Comparison with Existing Methods

We compared our proposed method with 17 state-of-the-art methods in Table 3. Our models based on CLIP and DINO both outperform the listed SOTA methods on all three datasets. Specifically, our model based on DINO ('Ours (DINO-ViT) w/ $\mathcal{L}^{tri}$' in Table 3) achieves the best performance on the Sketchy-Ext dataset and outperforms other DINO-based models by a large margin, in terms of all metrics on three datasets. It should be noted that all these SOTA methods utilized extra information such as text embeddings and the guidance from teacher models for model training, whereas our model only relies on the pre-trained vision model and fine-tunes it by combining the triplet loss and the proposed $\mathcal{L}^{mcc}$. This demonstrates that our method takes full advantage of the pre-trained DINO model knowledge and adapt it for ZSL-SBIR by our proposed $\mathcal{L}^{mcc}$.

ViT	$\mathcal{L}^{tri}$	$\mathcal{L}^{cls}$	$\mathcal{L}^{mcc}$	mAP@200	P@200
DINO	✓	✗	✗	0.775	0.734
DINO	✓	✗	✓	**0.805**	**0.768**
DINO	✗	✓	✗	0.375	0.345
DINO	✗	✓	✓	**0.615**	**0.551**
CLIP	✓	✗	✗	0.766	0.734
CLIP	✓	✗	✓	**0.790**	**0.757**
CLIP	✗	✓	✗	0.764	0.730
CLIP	✗	✓	✓	**0.782**	**0.747**

Table 7: Ablation study of $\mathcal{L}^{mcc}$ on Sketchy-ext dataset.

	mAP@200	P@200
CLIP	0.372	0.320
CLIP w/ $\mathcal{L}^{mcc}$	**0.506**	**0.433**
DINO	0.248	0.234
DINO w/ $\mathcal{L}^{mcc}$	**0.645**	**0.574**

Table 8: Fine-tuning without knowing the image categories.

Moreover, our method ('Ours (CLIP-ViT) w/ $\mathcal{L}^{cls}$') obtains better performance than the best CLIP-based method, i.e., CLIP-AT [Sain *et al.*, 2023a], which has the same architecture as our CLIP-based model and was trained exploiting classification loss and prompt learning. Without prompt learning, our model $\mathcal{L}^{mcc}$ surpasses CLIP-AT by 5.9% and 2.2%, in terms of mAP@200 and P@200, respectively, on the Sketchy-Ext dataset (Split 2), by 1.7% and 4.1% in mAP@all and P@100, respectievly, on TUBerlin-Ext dataset, and by 13% in mAP@all on QuickDraw-Ext dataset. This demonstrates the effectiveness of our proposed $\mathcal{L}^{mcc}$.

Figure 3 presents a qualitative comparison among three methods: our proposed method ('Ours (CLIP-ViT) w/ $\mathcal{L}^{cls+mcc}$'), CLIP-AT, and our baseline model ('Ours (CLIP-ViT) w/ $\mathcal{L}^{cls}$'). It demonstrates that our model outperforms the other methods in identifying visually similar photos against the input sketch. Specifically, when considering the cow sketch (first 3 rows in Figure 3), our method accurately retrieves the top 10 photos, while both CLIP-AT and the baseline occasionally retrieve rhinoceros photos.

4.3 Ablation Studies

To evaluate the effectiveness of our proposed loss, we conducted ablation studies to examine individual components of (i) the proposed modality capacity constraint loss $\mathcal{L}_s^{mcc}$ and $\mathcal{L}_p^{mcc}$; (ii) the effect of the proposed modality capacity constraint loss in conjunction with different retrieval losses based on CLIP and DINO; (iii) the effect of using only modality capacity constraint loss on the Sketchy-Ext dataset (Split 2).

The effect of $\mathcal{L}_s^{mcc}$ and $\mathcal{L}_p^{mcc}$. Table 5 and Table 6 show the ablation study results based on CLIP with $\mathcal{L}^{tri}$ and $\mathcal{L}^{cls}$, respectively. Our observation is that removing either $\mathcal{L}_s^{mcc}$

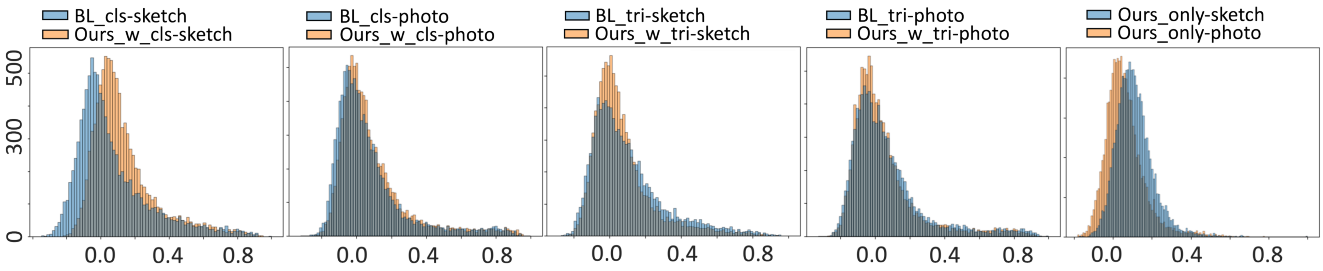

Figure 4: Inter-category similarity comparison (according to Table 11) on the Sketchy-Ext dataset.

	task	$\mathcal{L}^{cls}$	$\mathcal{L}^{mcc}$	mAP@200	P@200
CLIP	p2p	-	-	0.891	0.840
Baseline	p2p	✓	✗	0.941	**0.905**
Ours	p2p	✓	✓	**0.945**	**0.905**
CLIP	s2s	-	-	0.716	0.312
Baseline	s2s	✓	✗	0.835	0.776
Ours	s2s	✓	✓	**0.843**	**0.787**

Table 9: Sketch-to-sketch (s2s) and photo-to-photo (p2p) image retrieval comparison.

From	To
door, cabin, helicopter, pear, saw, scissors, skyscraper, sword, tree, wheelchair, windmill, window	object
dolphin	living thing that can swim
cow, giraffe, mouse, raccoon, rhinoceros	living thing that lives on land
bat, seagull, songbird	living thing that can fly

Table 10: Mapping table for super class based image retrieval.

or $\mathcal{L}_p^{mcc}$ leads to a decrease on both mAP@200 and P@200 on CLIP. Moreover, the impact of removing $\mathcal{L}_s^{mcc}$ is higher than $\mathcal{L}_p^{mcc}$. This indicates that improving the sketch representation is more important than photo representation. This is consistent with the observation in Table 1 that the modality capacity of sketch modality is higher than that of photo, requiring more improvement for the sketch representation.

The effect of the proposed modality capacity constraint loss in conjunction with different retrieval losses. We conducted an ablation study to investigate the effectiveness of our proposed loss function ($\mathcal{L}^{mcc}$ as a whole) when combined with various retrieval losses (such as triplet loss and InfoNCE loss) and different model architectures, as shown in Table 7. The results demonstrate that incorporating our proposed loss function significantly improves the performance of ZS-SBIR. For instance, when we introduce our loss function to the method that uses CLIP ViT with InfoNCE loss, we observe a notable increase of 3.4% in mAP@200. An interesting finding is that when the model architecture is DINO ViT, the $\mathcal{L}^{cls}$ results in a poor performance of only 0.375 in terms of mAP@200. The reason behind this could be attributed to the fact that the pre-trained DINO ViT is a fully self-supervised vision model, lacking exposure to the semantic embeddings, thereby posing extreme challenge to align the vision and semantic spaces through the supervision of the InfoNCE loss. However, after applying our proposed modality capacity constraint loss , the mAP is increased from 0.375 to 0.615, which demonstrates that our proposed loss is complementary with the InfoNCE loss and is able to improve the representation though the overall similarity guidance rather than the semantic alignment.

The effect of only using modality capacity constraint loss. Table 8 shows the comparisons results with the original pre-trained models and the model fined using only our proposed modality capacity constraint loss . As it can be seen, merely using our proposed modality capacity constraint loss has also significantly improved features for ZSL-SBIR. Hence, our proposed loss can also enable the model to learn discriminative representations to distinguish images like triplet loss and InfoNCE loss. In addition, the performance increase on DINO is higher than that on CLIP using our proposed loss, because the representation from CLIP contains semantic information, affecting the visual similarity distribution and confusing the guidance.

4.4 Hyperparameter Analysis

Figure 5 shows the sensitivity analysis on performance when hyperparameters are tuned across different values. Since the pre-trained models perform worse on sketches than photos (see Table 8), our our strategy is to find out the optimal setting of γ_s first and then search the optimal setting of γ_p. As shown, when γ is set in a range of 0 to 0.3, the model can be effectively improved, meaning that it does not need careful tuning of hyperparameter γ. Figure 5 also shows that it is advantageous to constrain the inter-category mean of sketches to approximately 0.2, while maintaining the mean of photos around 0.0. This finding aligns with the characteristics of the sketch and photo modalities. For images of different categories, variations in background, angle, and lighting, leading to higher similarity between images belonging to different categories. Therefore, it is reasonable to minimize the inter-similarity to zero. However, for sketches of different categories, even if the sketch shapes are different, they still share the same background and with black lines, exhibiting a certain level of similarity. Hence, it is unreasonable to minimize the inter-similarity to zero for sketches; the inter-similarity should be minimized to a value slightly greater than zero.

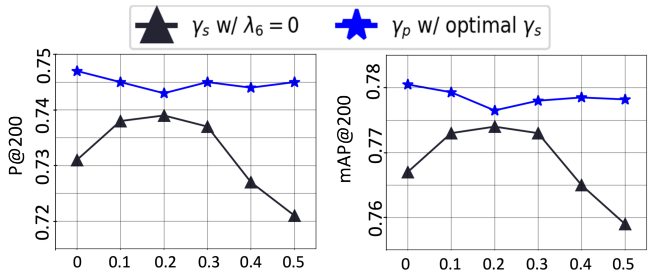

Figure 5: Performance from various settings of γ_s and γ_p using 'Ours (CLIP-ViT) w/ $\mathcal{L}^{cls+mcc}$ ' on the Sketchy-Ext dataset.

Input: 'A living thing that can fly'

Figure 6: Super Class Based Image Retrieval. Top: Baseline (CLIP-ViT w/ $\mathcal{L}^{cls}$). Bottom: Ours (CLIP-ViT w/ $\mathcal{L}^{cls}$ + $\mathcal{L}^{mcc}$).

4.5 Inter-Category Similarity Analysis

Table 11 displays the mean and standard deviation of inter-category similarities after fine-tuning CLIP and Figure 4 shows the density distribution comparisons of inter-category similarities of baselines and our models. Our results reveal distinct patterns for baselines utilizing respective $\mathcal{L}^{cls}$ and $\mathcal{L}^{tri}$. Specifically, the $\mathcal{L}^{cls}$ aligns both photos and sketches to a shared semantic space, making mean of the inter-category similarities of photos and sketches becomes extremely similar (0.11 vs 0.11). On the other hand, the $\mathcal{L}^{tri}$ involves the interaction between the sketch and photo branches during the training process, allowing for different mean values (0.18 vs 0.10) to be achieved for sketch and photo modalities. When $\mathcal{L}^{cls}$ is combined with our $\mathcal{L}^{mcc}$, the inter-category mean value of sketches increases. This is because that when the photo population distribution is constrained with a target mean, the sketch population can better preserve original knowledge from the pre-trained model due to the stabilization of the counterpart (from 0.91 to 0.35 vs from 0.91 to 0.18). The preservation of such knowledge would be advantageous for aligning photo and sketch spaces. However, when combined with $\mathcal{L}^{tri}$, we constrain the two modalities to have highly similar population characteristics ($\gamma_s = 0.1$ and $\gamma_p = 0.0$) to achieve better overlay of photos (0.09 ± 0.28) and sketches (0.14 ± 0.20) in a common space.

5 Other Applications

Image-to-image Retrieval. we investigate whether controlling γ_s or γ_p individually can benefit the tasks of sketch-to-sketch or photo-to-photo retrieval. To do so, we employ the setting of 'Ours (CLIP) w/ $\mathcal{L}^{cls}$' to fine-tune CLIP for retrieval . Table 9 demonstrates that using proper value of γ_I can yield better single-domain retrieval performance.

Loss	$\mathcal{D}_s$	$\mathcal{D}_p$
$\mathcal{L}^{cls}$	$0.11_{\pm 0.31}$	$0.11_{\pm 0.27}$
$\mathcal{L}^{tri}$	$0.18_{\pm 0.30}$	$0.10_{\pm 0.28}$
$\mathcal{L}^{cls+mcc}$	$0.35_{\pm 0.22}$	$0.07_{\pm 0.29}$
$\mathcal{L}^{tri+mcc}$	$0.14_{\pm 0.28}$	$0.09_{\pm 0.28}$
$\mathcal{L}^{mcc}$	$0.42_{\pm 0.20}$	$0.22_{\pm 0.23}$

Table 11: 'mean ± std' of the inter-category similarities after fine-tuning on the Sketchy-Ext dataset using the CLIP-based model.

	mAP@200	P@200	P@100
Baseline	0.909	0.920	0.917
Ours	**0.949**	**0.945**	**0.947**

Table 12: Quantitative comparison of super class based image retrieval (using Table 10) on Sketchy-Ext dataset. 'Baseline' and 'Ours' refers to '$\mathcal{L}^{cls}$' and '$\mathcal{L}^{cls+mcc}$ ', respectively. Both methods utilize the CLIP-ViT architecture.

Super class based Photo/Sketch retrieval. In addition to traditional explicit category/text-based image retrieval methods, there is a growing need to leverage learned features for performing 'fuzzy match', which is widely accepted and applied in modern text-based retrieval systems. When it comes to image based retrieval systems, we refer such 'fuzzy match' task as super class based photo/sketch retrieval, where a super class is a description less implicit than the category. Table 12 illustrates the superior performance using our method. Figure 6 further demonstrates that our method yields more accurate result in the super class based image retrieval task.

6 Conclusions

In this paper, we investigate the modality capacity to effectively guide the feature learning process for ZS-SBIR. We propose a modality capacity constraint loss that enables control over the inter-category similarity during the training process. Our loss function seamlessly integrates with existing retrieval loss functions, including triplet loss and InfoNCE loss. Extensive experiments have been conducted to demonstrate the effectiveness of our approach, which outperforms the state-of-the-art (SOTA) methods by a significant margin. Additionally, our proposed loss function has the potential to benefit other applications, such as sketch-to-sketch, photo-to-photo, and even challenging super class-based image retrieval tasks. We will explore applying the modality capacity to other multi-modality tasks as our future work. Hopefully, this work can inspire the research community on more effective and in-depth modality capacity utilization strategies.

Acknowledgements

The work described in this paper was supported, in part, by grant from the Innovation and Technology Fund (Grant number ITP/028/21TP) and The Hong Kong Polytechnic University (Project code: CD95).

References

[Akata *et al.*, 2016] Zeynep Akata, Mateusz Malinowski, Mario Fritz, and Bernt Schiele. Multi-cue zero-shot learning with strong supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 59–68, 2016.

[Brunelli and Mich, 2001] Roberto Brunelli and Ornella Mich. Histograms analysis for image retrieval. *Pattern Recognition*, 34(8):1625–1637, 2001.

[Chen *et al.*, 2018] Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1043–1052, 2018.

[Dey *et al.*, 2019a] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In CVPR [2019b], pages 2179–2188.

[Dey *et al.*, 2019b] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *CVPR*, pages 2179–2188, 2019.

[Dong *et al.*, 2023] Shiyin Dong, Mingrui Zhu, Nannan Wang, Heng Yang, and Xinbo Gao. Adapt and align to improve zero-shot sketch-based image retrieval. *arXiv*, 2023.

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

[Eitz *et al.*, 2012] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM TOG*, 31(4):1–10, 2012.

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

[Luo *et al.*, 2020] Yuxuan Luo, Xizhao Wang, and Weipeng Cao. A novel dataset-specific feature extractor for zero-shot learning. *Neurocomputing*, 391:74–82, 2020.

[Ren *et al.*, 2023] Hao Ren, Ziqiang Zheng, Yang Wu, Hong Lu, Yang Yang, Ying Shan, and Sai-Kit Yeung. Acnet: Approaching-and-centralizing network for zero-shot sketch-based image retrieval. *IEEE TCSVT*, 2023.

[Sain *et al.*, 2023a] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *CVPR*, pages 2765–2775, 2023.

[Sain *et al.*, 2023b] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2765–2775, 2023.

[Shigeto *et al.*, 2015] Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto. Ridge regression, hubness, and zero-shot learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I 15*, pages 135–151. Springer, 2015.

[Stricker and Swain, 1994] Stricker and Swain. The capacity of color histogram indexing. In *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 704–708. IEEE, 1994.

[Tian *et al.*, 2022] Jialin Tian, Xing Xu, Fumin Shen, Yang Yang, and Heng Tao Shen. Tvt: Three-way vision transformer through multi-modal hypersphere learning for zero-shot sketch-based image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2370–2378, 2022.

[Tian *et al.*, 2023] Jialin Tian, Xing Xu, Zuo Cao, Gong Zhang, Fumin Shen, and Yang Yang. Zero-shot sketch-based image retrieval with adaptive balanced discriminability and generalizability. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 407–415, 2023.

[Wang *et al.*, 2021a] Wenjie Wang, Yufeng Shi, Shiming Chen, Qinmu Peng, Feng Zheng, and Xinge You. Norm-guided adaptive visual embedding for zero-shot sketch-based image retrieval. In *IJCAI*, pages 1106–1112, 2021.

[Wang *et al.*, 2021b] Zhipeng Wang, Hao Wang, Jiexi Yan, Aming Wu, and Cheng Deng. Domain-smoothing network for zero-shot sketch-based image retrieval. *arXiv preprint arXiv:2106.11841*, 2021.

[Wang *et al.*, 2022] Kai Wang, Yifan Wang, Xing Xu, Xin Liu, Weihua Ou, and Huimin Lu. Prototype-based selective knowledge distillation for zero-shot sketch based image retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 601–609, 2022.

[Zhang *et al.*, 2017] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2021–2030, 2017.