Active Learning for Fine-Grained Sketch-Based Image Retrieval

Himanshu Thakur* [1]
hthakur@andrew.cmu.edu
Soumitri Chattopadhyay* [2]
soumitri@cs.unc.edu

[1] Carnegie Mellon University
[2] UNC Chapel Hill

Abstract

The ability to retrieve a photo by mere free-hand sketching highlights the immense potential of Fine-grained sketch-based image retrieval (FG-SBIR). However, its rapid practical adoption, as well as scalability, is limited by the expense of acquiring faithful sketches for easily available photo counterparts. A solution to this problem is Active Learning, which could minimise the need for labeled sketches while maximising performance. Despite extensive studies in the field, there exists no work that utilises it for reducing sketching effort in FG-SBIR tasks. To this end, we propose a novel active learning sampling technique that drastically minimises the need for drawing photo sketches. Our proposed approach tackles the trade-off between uncertainty and diversity by utilising the relationship between the existing photo-sketch pair to a photo that does not have its sketch and augmenting this relation with its intermediate representations. Since our approach relies only on the underlying data distribution, it is agnostic of the modelling approach and hence is applicable to other cross-modal instance-level retrieval tasks as well. With experimentation over two publicly available fine-grained SBIR datasets ChairV2 and ShoeV2, we validate our approach and reveal its superiority over adapted baselines.

1 Introduction

The success of computer vision applications can be largely attributed to deep learning architectures [14, 39], which, in turn, have yielded favourable results due to their access to large-scale labelled databases [12, 20] for training. Being in the age of Big Data, enormous volumes of data is easily available; however, proper annotation of the same is a painstakingly cumbersome as well as an expensive process, often requiring specialized qualifications if the task at hand demands for domain expertise, such as handling medical images [43]. To alleviate this bottleneck, researchers have proposed various annotation-efficient methods [3, 19, 21, 47] to standard computer vision tasks like classification and segmentation. A commonly used technique is active learning [6, 34, 36], which seeks to find the most "useful" unlabelled data samples to be annotated for learning, so as to reduce annotation cost as well as increase overall generalisability on the supervised learning task to be performed.

*Both authors contributed equally to the paper.

Apart from conventional visual tasks, active learning has been applied to other domains such as video captioning [9], hand pose estimation [8] and single-image super-resolution [42].

In this paper, we embrace a paradigm shift to tackle the aforementioned challenges in a domain which is fundamentally very much different from traditional vision tasks – fine-grained sketch-based image retrieval (FG-SBIR) [4, 5, 30], a relatively newer direction of research from traditional category-level SBIR [11, 41]. As the name suggests, FG-SBIR aims at exploiting the finer sketch representations for cross-modal instance-level retrieval, achieved by learning an embedding space where sketch-photo pairs lie close to each other. The most common approach in several works [4, 23, 46] has been to train a supervised triplet loss-based model [13] that learns feature similarities between an image and its corresponding sketch and hence requires a large number of sketch-photo pairs. However, drawing full sketches is both time-consuming and difficult, since it requires artistic expertise and amateurish sketching can only lead to learning degradation. Thus, there is a need to develop a robust, annotation-efficient pipeline for FG-SBIR. Very few recent works have been proposed treading on this motivation, such as a generalisable zero-shot [23] and semi-supervised learning [5] FG-SBIR framework. While the latter involves training of two networks which makes it computationally expensive, the performance of [23] is far from fully-supervised alternatives.

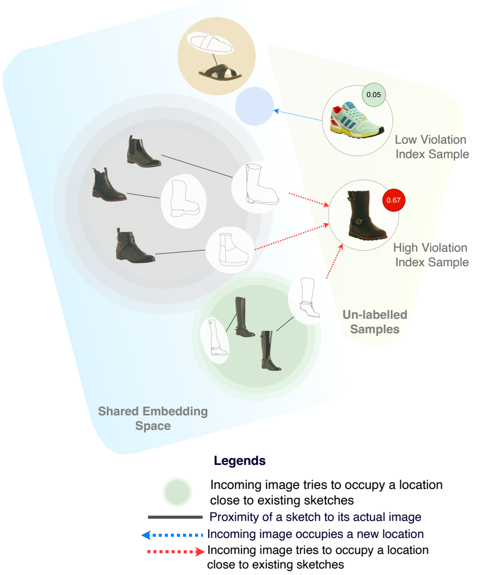
0.05
Low Violation Index Sample
0.67
High Violation Index Sample
Un-labelled Samples
Shared Embedding Space
Legends
Incoming image tries to occupy a location close to existing sketches
Proximity of a sketch to its actual image
Incoming image occupies a new location
Incoming image tries to occupy a location close to existing sketches

Figure 1: Intuition behind our proposed AL framework. The *violation index* quantifies the "disturbance" introduced into the learned embedding space by the incoming photo sample, due to having a greater similarity with a previously paired sketch present in the latent space. More details are provided in section 4.

Developing an Active Learning pipeline for FG-SBIR imposes some unique challenges. Considering a traditional FG-SBIR model which lacks a probability distribution of the samples in its output, there is an absence of a direct way to measure the uncertainty of such a model. Hence, it becomes difficult to select a photo for labelling without having an estimate of its uncertainty from the model. Moreover, off-the-shelf active learning methods [18, 36] that were primarily proposed for classification tasks are not suitable for FG-SBIR, since for classification the learning mechanism *draws firm discriminatory boundaries* among samples, whereas in a cross-modal instance-level retrieval setup the objective is to draw *softer decision boundaries*, as the samples belong to the same category and only differ in minute fine-grained details. Additionally, the FG-SBIR model holds the photo and sketch embeddings in a joint space and thus, selection based on only photo or sketch might yield previously unseen results, compared to selection from a single modality. Hence, developing a sampling technique for the FG-SBIR requires the handling of both modalities - photo and its sketch.

The effectiveness of an active learning pipeline depends highly upon the *technique* of selecting samples from the unlabelled pool. As a result, a good technique for sampling photos would lead to the maximum increase in the model's performance. To this end, we propose a novel sampling strategy for active learning that utilises the evolving relations between photos and their sketches to approximate the influence of a new photo from the unlabelled pool, on the model's existing knowledge. Our sampling technique is informed by the embedding space learnt using the labelled pool of photo-sketch pairs and two major aspects of an unlabelled photo – *its predicted representation* and *its approximate potential of influence* in the existing embedding space. While the former is the basis of our sampling technique, the latter helps in tackling the classic trade-off between uncertainty and diversity-based sampling techniques. Specifically, to model an unlabelled photo's potential of influencing the existing embedding space, we formulate a quantity, *violation index*. The violation index of a photo acts as a proxy for measuring the amount of confusion it could create in the existing sketch-photo pair embedding space (refer to Figure 1 for an intuitive understanding). Further, for diversity sampling, we choose *k*-MEANS++ due to its ability to converge faster, as highlighted in [1]. The overall advantage is two-fold: since we adopt a relative method of approximating the influence using the task network itself, it does not need an auxiliary learner or source of knowledge. Moreover, our technique relies on the underlying distribution of the data itself and hence is agnostic of the model used. The number of samples to be chosen depends on the permissible budget of annotation; the chosen samples are then queried for their sketches to be drawn, which are then paired and put into the training examples.

To sum up, the primary contributions of the presented study are as follows: (1) For the first time, we propose an *active learning* pipeline for annotation-efficient fine-grained SBIR; (2) To this end, we formulate a novel sampling strategy that incorporates uncertainty as well as diversity to quantify the “usefulness" of an unlabelled sample for querying its label (here, sketch); (3) With suitable experimentation and comparison with adopted baselines on two publicly available FG-SBIR datasets, as well as conducting ablations on the proposed framework, we demonstrate the usefulness of our approach.

2 Related Works

Fine-grained SBIR: Although SBIR was originally proposed and studied as a category-level retrieval task [7, 31, 41], recently there has been significant interest among researchers towards exploiting the *fine-grained* information that sketches provide [4, 5, 23, 33, 37] for enhanced cross-domain matching, something other query mediums (e.g. text) fail to do. The first deep learning-based approach by Yu *et al.* [46] was further enhanced using cross-domain generative-discriminative learning [22] and attention mechanisms [37]. More recent studies include zero-shot-like cross-category FG-SBIR [23, 32]; cross-modal co-attention-based hierarchical model [29]; on-the-fly SBIR setup for early retrieval [4]; style-agnostic meta-learning setup [30] and a semi-supervised framework [5] to tackle data scarcity in FG-SBIR [33]. However, none of these works seeks to address the need for a smart data labelling system for cross-modal instance-level retrieval problems like FG-SBIR so that we can achieve optimal performance using the minimum possible labelling budget.

Active Learning: Active learning (AL) [6, 26] has been extensively studied for over two decades, the primary goal being to develop an effective strategy to reduce annotation effort by “actively" selecting representative samples and improve learning. Apart from conventional image classification [3, 36], AL has been widely used for various computer vision tasks such as medical imaging [6], image and video segmentation [2, 35, 44], among others.

Broadly speaking, AL approaches may be categorized as: (1) Uncertainty-based methods [3, 10, 40, 45], which aim to construct a so-called acquisition function that quantifies the uncertainty of the model on unlabelled data points, based on which the points are sampled and queried for annotation; and (2) Representation-based methods [1, 16, 34, 36], which seek to learn a common embedding space for the labelled and unlabelled data items so as to sample the unlabelled data points that capture the most diverse regions of the embedding and thereby better represent the overall data distribution. Existing AL methods mostly deal with classification and thus cannot be directly adopted for FG-SBIR where paired photo and sketch need to be aligned in the vicinity in the joint embedding space. This cross-modal instance-wise matching brings an exclusively different set of challenges for employing AL in FG-SBIR. We intend to address and tackle these through this work, which, to the best of our knowledge, is the first to introduce AL to a cross-modal instance-level retrieval problem.

3 Background and Problem Formulation

Baseline FG-SBIR: Instead of complicated pre-training [24] or joint-training [5], we use a three-branch state-of-the-art siamese network [23] as our baseline retrieval model, which is considered to be a strong baseline to date [23, 37]. Each branch starts from ImageNet pre-trained VGG-16 [15], sharing equal weights. Given an input image $I \in \mathbb{R}^{H \times W \times 3}$, we extract the convolutional feature-map $\mathcal{F}(I)$, which upon global average pooling followed by l_2 normalisation generates a d dimensional feature embedding. This model has been trained with an anchor sketch (a), a positive (p) photo, and a negative (n) photo triplets $\{a, p, n\}$ using *triplet-loss*. Triplet-loss aims at increasing the distance between anchor sketch and negative photo $\delta^- = ||\mathcal{F}(a) - \mathcal{F}(n)||_2$, while simultaneously decreasing the same between anchor sketch and positive photo $\delta^+ = ||\mathcal{F}(a) - \mathcal{F}(p)||_2$. Therefore, the triplet-loss $\mathcal{L}$ with the margin hyperparameter $\mu > 0$ can be written as:

$$\mathcal{L} = max\{0, \delta^+ - \delta^- + \mu\} \tag{1}$$

During inference, given a gallery of M photos $\{P_i\}_{i=1}^{M}$, we can compute a list of d dimensional vectors as $G = \{\mathcal{F}(P_i)\}_{i=1}^{M}$. Now, given a query sketch S, and pair-wise distance metric, we obtain a top-q retrieved list as $Ret_q(F(S), G)$. If the paired (ground-truth) photo appears in the top-q list, we consider accuracy to be true for that sketch sample.

Active Learning for FG-SBIR: Following the principle of active learning shown in Equation 2, we aim to minimise the number of rounds $\mathcal{R}$ so that the fewest possible sketches are needed to be acquired. At the end of each of round r, the performance measures $\mathcal{P}$ are recorded to quantify and understand the effectiveness of our sampling technique $\mathcal{X}$.

$$\min_{\mathcal{R}} \min_{\mathcal{L}} [\mathcal{X}(\mathcal{L}|\mathcal{P}_0 \subset \cdots \mathcal{P}_k \subset \mathcal{P}_U)]_{r=1}^{\mathcal{R}} \tag{2}$$

4 Proposed Method

Overview: We aim to design an active learning framework specially suited for fine-grained SBIR so as to make it label-efficient. To keep our framework fairly simple, we have used the baseline triplet network (described in Section 3) as the retrieval model. During the training process, we leverage active learning to select the most informative examples for querying its labels, the details of which are given in the subsequent sections. The AL mechanism is

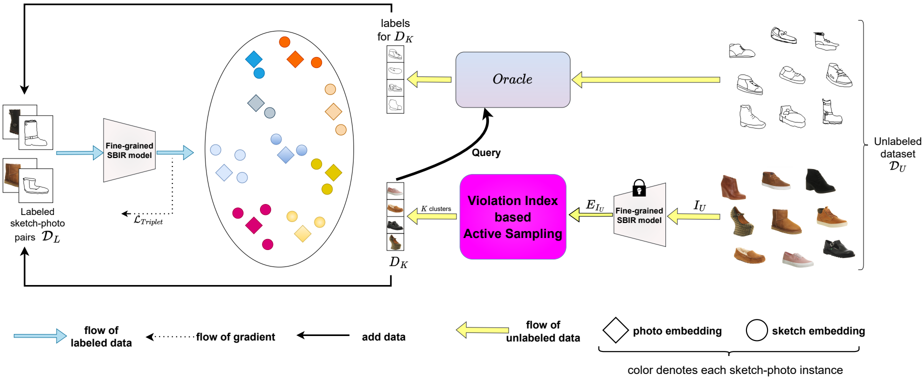

Figure 2: Overall workflow of the proposed active learning framework for fine-grained SBIR. The method starts with a subset of sketch-photo pairs for learning the FG-SBIR model, following which it is used to compute violation index of unlabeled photos with respect to the embedding space. Our sampling technique selects a suitable query set of photos which are passed to the Oracle (the ground truth sketch provider for queried photos) to obtain their sketch counterparts. These pairs are subsequently added to the labeled subset.

employed in training cycles; at the end of each cycle the sampling technique selects photos from the unlabelled data and queries for their paired sketches, which are provided by the Oracle (ground truth annotator) and added to the labelled training pool.

4.1 Introducing Violation Index

In this section, we introduce the core idea of our contribution – the **violation index**. Our approach is inspired by the learning process of cross-modal retrieval frameworks. When an FG-SBIR model learns, it tries to put an image and its sketch nearby in the embedding space, while pushing away other images. Formally, consider a model $\mathcal{M}$, the embedding of an image obtained from final layer of the model as E_I, and that of the corresponding sketch as E_S. Also, let $E_{I'}$ be the embedding of an image other than E_I, N_L be the total size of the labelled dataset $\mathcal{D}_L$ and N_U be the total size of the unlabeled dataset $\mathcal{D}_U$. To learn the similarities, following Section 3 let us consider a triplet loss function defined as follows:

$$\mathcal{L} = \max\left(||E_I - E_S||^2 - ||E_I - E_{I'}||^2 + \mu, 0\right) \tag{3}$$

The objective of the learning process is to minimise $\mathcal{L}$ over all images in the dataset (Eq 4). Hence, an ideal embedding space would comprise images lying closest to its actual sketch. However, a non-ideal embedding space is more realistic and introduces some new challenges.

$$minimize \sum_{i=1}^{N_L} \mathcal{L}\left(E_I^{(i)}, E_S^{(i)}, E_{I'}^{(i)}\right) \tag{4}$$

Now, when a new image I' from the unlabelled set $\mathcal{D}_U$ is introduced during training, its corresponding sketch might not lie closest to it in the embedding space, a condition defined in Equation 5. Hence, in the model's view, the image now seems closer to one of the existing sketches rather than its own. This phenomenon leads to perturbation in the existing embedding space. We quantify the degree of disturbance an image from the unlabelled pool of images produces in the existing embedding space and call it its *violation index*.

$$||E_{I'} - E_S||^2 \leq ||E_I - E_S||^2 \tag{5}$$

Formally, we define the violation index (VI) of the photo embedding $E_{I'}$ as:

$$ViolationIndex(E_{I'}) = \frac{1}{N_L} \times \sum_{i=1}^{N_L} \frac{||E_I^{(i)} - E_S^{(i)}||}{||E_{I'} - E_S^{(i)}||} \tag{6}$$

Thus, the violation index is an improvement over a simpler distance-based sampling technique since it accounts for the fact that the inherent imperfections in sketches cause the cross-modal embedding space to be sensitive to new image-sketch pairs. Moreover, since the metric is built on relative distances instead, computing the relative similarities between a new image and existing pairs help surface novel samples that still do not violate existing sketches or images.

4.2 VI-based Active Sampling

Given the violation index of an image shows its average relative distance from existing image-sketch pairs and hints towards how many of them it violates in the learned embedding space. Intuitively, images with a low violation index would be relatively unseen in the training data, whereas ones with a high violation index might closely resemble one or more images from the training set.

Images with a low violation index or **min violating samples** are relatively unseen in the training data, which means that they may be novel or unique compared to other images. These images may contain features or characteristics that are not commonly found in the training set. On the other hand, images with a high violation index or **max violating samples** are more likely to closely resemble one or more images from the training set. This suggests that these images may contain familiar features that are present in the training data.

When considering the selection strategy in Active Learning, a naive solution could be to select the samples with minimum violation indices,i.e., the sampling technique X gives us the set $\{x \in D_U :| VISet_{D_U} \cap (-\infty, VI(E_x) |< K\}$ where $VISet_{D_U}$ is the set of violation indices for the images in the unlabelled set and $VI(E_x)$ is the violation index for image x (Eq 6). Although with this approach the perturbation in the existing embedding space is minimized, it loses out on the opportunity to reduce uncertainty by learning through closely resembling images. As such, there exists a tradeoff between reducing existing uncertainty and learning novel instances. Hence, a better selection strategy would be an ensemble of minimum and maximum violating image samples, i.e, a better X gives us the set

$$\{x \in D_U :| VISet_{D_U} \cap (-\infty, VI(E_x)) |< p\} \;\cup$$
$$\{x \in D_U :| VISet_{D_U} \cap (VI(E_x), \infty) |> (K - p)\}$$

where $p \in I$, is a hyper-parameter such that $0 \leq p \leq N_U$.

Since the violation index does not inherently capture the diversity of images, a further enhancement of our selection strategy includes diversity sampling to maximize novelty in the selected subset. To do this, we adopt a *kmeans*++ based clustering of the image embeddings followed by violation index-based selection inside each cluster. We select *kmeans*++ due to its ability to converge faster [1]. The overall workflow has been depicted in Figure 2.

5 Experiments and Results

Datasets: We use QMUL-Shoe-V2 [23, 27] and QMUL-Chair-V2 [38] datasets that have been specifically designed for FG-SBIR. QMUL-Shoe-V2 contains a total of 6,730 sketches and 2,000 photos, of which we use 6,051 and 1,800 respectively for training and the rest for testing. For QMUL-Chair-V2, we split it as 1,275/725 sketches and 300/100 photos for training/testing respectively. For each photo, one of its possible sketches is selected randomly, and is considered as its label. Initially, we consider 300 photo-sketch pairs as the labelled set and the rest as unlabelled (i.e. absence of their corresponding sketches).

Implementation: We implemented our framework in PyTorch [25] accelerated by an 11 GB Nvidia RTX 2080-Ti GPU. ImageNet [28] pre-trained VGG-16 [15] network (embedding dimension $D = 256$) is used as the backbone network for both sketch and photo branches. In all experiments, we use Adam optimizer [17] with learning rate of $1e-4$, batch size 16 and train the base model with a triplet objective. In the active learning setup, we conduct 5 cycles of complete training, where at the end of each round, we employ our sampling technique to add K samples to the labelled pool after the provision of its actual sketch.

In each active learning round, we obtain the embeddings for the photos and sketches from the labelled set and the photos from the unlabelled set. Following this, our sampling technique utilises these embeddings to select K photos from the unlabelled set, whose corresponding sketches from the dataset are then used to label them. Once labelled, these photos and their sketches are added to the labelled set.

Evaluation Metrics: Following the standard FG-SBIR setting [4, 5], we quantify the performance of the sampling technique using acc.@q metric ($q = 1, 10$), i.e. the percentage of sketches having true-match photos appearing in the top-q list, after each AL round.

5.1 Baselines

To the best of our knowledge, there has been no previous work on active learning for fine-grained sketch-based image retrievals. Thus, we compare ours with a few SoTA active learning techniques adapted suitably for fine-grained SBIR. We choose three widely used baseline sampling techniques, namely: random sampling, kmeans sampling, and coreset sampling. As evident from our choice of baseline methods, we do not consider uncertainty-based sampling techniques due to the limitations discussed in Section 3.3. Random sampling is a widely used baseline for evaluating active sampling techniques. Here, we randomly select photos for labelling based on our-predefined budget. In K-means sampling [48], we cluster photos into as many clusters as the labeling budget and select the photos closest to the cluster centroid. In case of a tie, we randomly select any one of the closest photos. Finally, Coresets sampling [34] is another diversity sampling technique in which we utilise "coresets" to find the most representative photos for sampling. To find the coresets, we use a farthest-first approach to select maximally distant photos in the embedding space.

5.2 Performance Analysis

To compare our approach with baselines, we report the mean and one standard deviation of acc.@1 values across 5 different runs. We also report the initial accuracy of the model. Figure 3 compares our violation index-based approach and baseline techniques. Before any

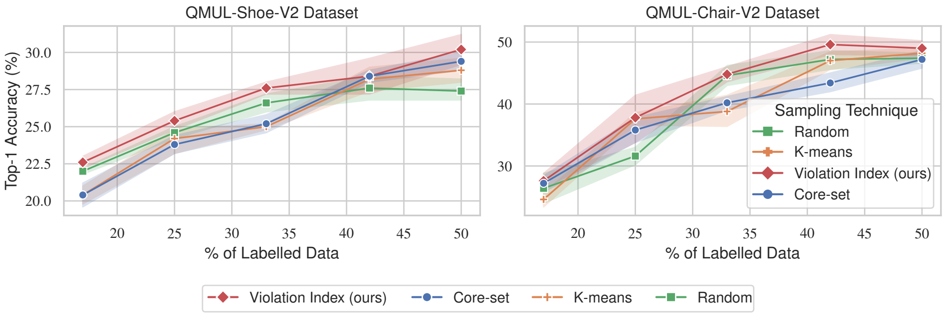

Figure 3: Comparing VI-based active sampling (ours) with SoTA AL baselines for fine-grained SBIR. All results are reported as mean of 5 runs and an error band of $\sigma = 1.0$.

active learning, the acc.@1 after training it on 8% labelled data were 13% (49.8%) and 11.7% (42.8%) on the Shoe and Chair datasets respectively.

Our method outperforms the baselines on both QMUL-Shoe-V2 and QMUL-Chair-V2 datasets and obtains consistently higher mean accuracies. We consider the mean accuracy differences between our approach and the baselines to further substantiate the results shown in Figure 3. On the QMUL-Chair-V2 dataset, we achieve a mean gain of 2.6% acc.@1 compared to the 3 baseline techniques. On the QMUL-Shoe-V2 dataset, we observe a mean increase of 1.1% acc.@1.

Our proposed approach achieves comparable performance by utilizing only 40-50% of the dataset. Compared to the state-of-the-art acc.@1 of 36.47% on 100% of the QMUL-Shoe-V2 dataset [23], we obtain a mean acc.@1 of 29.8% by only utilizing 40% of the dataset. This is achieved when we use α=0 as our hyperparameter. We also see consistent results on the QMUL-Chair-V2 dataset (using α=0.7), where we obtain 49.6% acc.@1 by only using 50% of the dataset.

5.3 Ablation Study

To better understand the effects of violation index-based sampling, we perform ablations on violation index-based selection and the choice of hyper-parameter.

Significance of violation index and diversity-based sampling: For this, we consider selecting only the unlabelled images that have the smallest and largest violation indices in active learning rounds performed on the QMUL-Shoe-V2 dataset.

As per Figure 4, an interesting observation is made: In early active learning cycles, selecting minimum VI yields higher acc.@1 compared to maximum VI. As the training dataset grows, this relationship is inverted, with higher acc.@1 achieved through Maximum VI-based selection. Empirical results support our hypothesis on violation indices, indicating that with less training data, increased retrieval accuracy primarily stems from selecting diverse instances. Conversely, with larger training data, reducing model uncertainty on similar instances contributes more to acc.@1. Thus, selecting maximum VI samples yields optimal results.

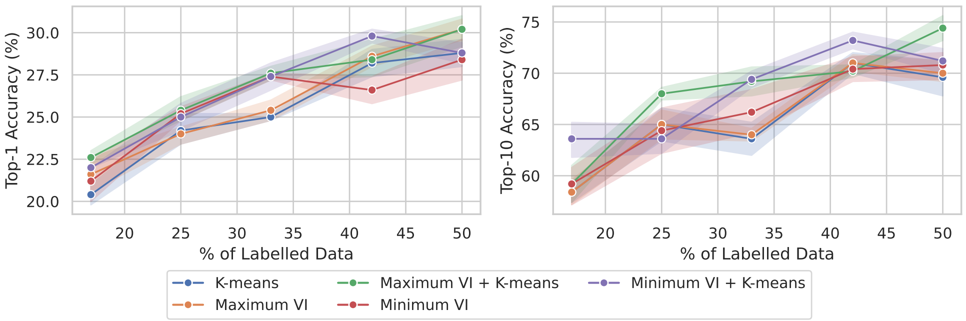

Figure 4: Ablations on diversity-based sampling and violation index. We compare only VI-based sampling, diversity-based sampling, and a combination of both.

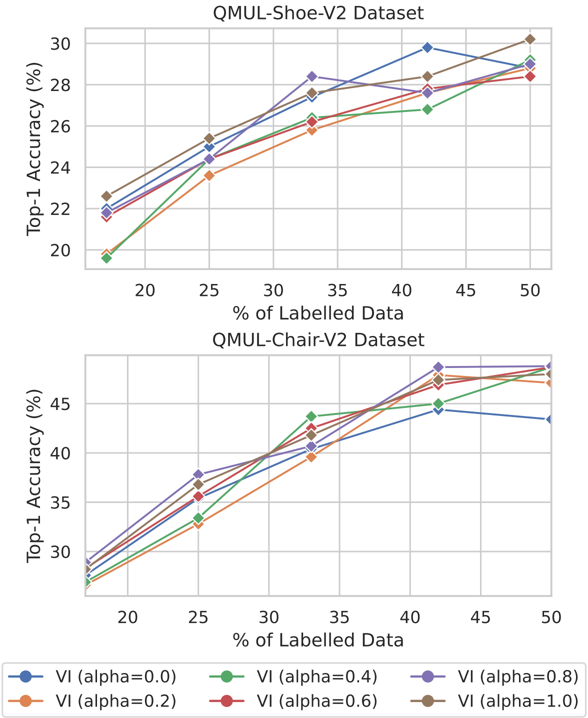

Figure 5: Ablations on value of hyper-parameter α on the ShoeV2 and ChairV2 datasets.

Another significant aspect of our experiment involves the need for diversity-based sampling, not just violation index-based sampling. As shown in Figure 4, violation index-based selection within diverse clusters obtained from kmeans++ consistently outperforms vanilla violation-based approach and kmeans++. Thus, the violation index captures semantic relations between labeled and unlabeled datasets but does not explicitly utilize relations between unlabeled images for diverse selection.

Sensitivity to hyper-parameter: We analyze the sensitivity of our method to the hyperparameter α, the results shown in Figure 5. We vary the value of α from 0 to 1 with a gap of 0.1 and report the mean acc.@1 on both datasets. On the QMUL-Shoe-V2 dataset, we observe a steady increase in acc.@1 as we increase α. On this dataset, α=0 produces the best results. On the QMUL-Chair-V2 dataset, we observe a steady increase in initial and final acc.@1 as we increase α. As per Figure 6 and on this dataset, α=0.7 produces the best results. On both datasets, there is a sudden drop in performance as we increase α from 0.0 to 0.1 and 0.2. We also present the same results compared with the baseline method in Figure 6. We believe that studying this interesting behavior is an open research avenue.

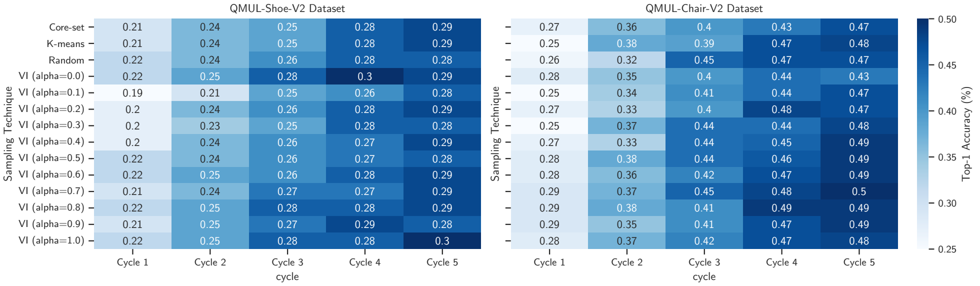

Figure 6: Ablations on value of hyper-parameter α and its comparison with baselines on the QMUL-Shoe-V2 and QMUL-Chair-V2 datasets.

6 Conclusion

We have proposed an active learning framework to tackle the annotation bottleneck in fine-grained SBIR. To this end, we brought forth a quantifiable metric, *violation index*, that measures the latent space displacements due to addition of a new instance. With suitable experiments and ablations, we have shown the robustness of our model compared to classification-specific AL works, especially in the low-data regimes. Our model is modality agnostic and thus can be leveraged for any cross-modal retrieval task. In future, we plan to extend our studies by investigating its applicability to such other tasks.

References

[1] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *ICLR*, 2019.

[2] Soufiane Belharbi, Ismail Ben Ayed, Luke McCaffrey, and Eric Granger. Deep active learning for joint classification & segmentation with weak annotator. In *WACV*, 2021.

[3] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *CVPR*, 2018.

[4] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In *CVPR*, 2020.

[5] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval. In *CVPR*, 2021.

[6] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 2021.

[7] Tu Bui, Leonardo Ribeiro, Moacir Ponti, and John Collomosse. Deep manifold alignment for mid-grain sketch based image retrieval. In *ACCV*, 2018.

[8] Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Active learning for bayesian 3d hand pose estimation. In *WACV*, 2021.

[9] David M Chan, Sudheendra Vijayanarasimhan, David A Ross, and John F Canny. Active learning for video description with cluster-regularized ensemble ranking. In *ACCV*, 2020.

[10] Jongwon Choi, Kwang Moo Yi, Jihoon Kim, Jinho Choo, Byoungjip Kim, Jinyeop Chang, Youngjune Gwon, and Hyung Jin Chang. VaB-AL: Incorporating class imbalance and difficulty with variational bayes for active learning. In *CVPR*, 2021.

[11] John Collomosse, Tu Bui, Michael J Wilber, Chen Fang, and Hailin Jin. Sketching with style: Visual search with sketches and aesthetic context. In *ICCV*, 2017.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[13] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *SIMBAD*, 2015.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[15] Andrew Zisserman Karen Simonyan. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[16] Kwanyoung Kim, Dongwon Park, Kwang In Kim, and Se Young Chun. Task-aware variational adversarial active learning. In *CVPR*, 2021.

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

[18] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *NeurIPS*, 2019.

[19] Qimai Li, Xiao-Ming Wu, Han Liu, Xiaotong Zhang, and Zhichao Guan. Label efficient semi-supervised learning via graph filtering. In *CVPR*, 2019.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[21] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *CVPR*, 2021.

[22] Kaiyue Pang, Yi-Zhe Song, Tony Xiang, and Timothy M Hospedales. Cross-domain generative learning for fine-grained sketch-based image retrieval. In *BMVC*, 2017.

[23] Kaiyue Pang, Ke Li, Yongxin Yang, Honggang Zhang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Generalising fine-grained sketch-based image retrieval. In *CVPR*, 2019.

[24] Kaiyue Pang, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval. In *CVPR*, 2020.

[25] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[26] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *arXiv preprint arXiv:2009.00236*, 2020.

[27] Umar Riaz Muhammad, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Learning deep sketch abstraction. In *CVPR*, 2018.

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[29] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Cross-modal hierarchical modelling for fine-grained sketch based image retrieval. In *BMVC*, 2020.

[30] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *CVPR*, 2021.

[31] Aneeshan Sain, Ayan Kumar Bhunia, Vaishnav Potlapalli, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketch3t: Test-time training for zero-shot sbir. In *CVPR*, 2022.

[32] 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*, 2023.

[33] Aneeshan Sain, Ayan Kumar Bhunia, Subhadeep Koley, Pinaki Nath Chowdhury, Soumitri Chattopadhyay, Tao Xiang, and Yi-Zhe Song. Exploiting unlabelled photos for stronger fine-grained sbir. In *CVPR*, 2023.

[34] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018.

[35] Yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *CVPR*, 2020.

[36] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *ICCV*, 2019.

[37] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV*, 2017.

[38] Jifei Song, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Learning to sketch with shortcut cycle consistency. In *CVPR*, 2018.

[39] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[40] Dan Wang and Yi Shang. A new active labeling method for deep learning. In *IJCNN*, 2014.

[41] Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *CVPR*, 2015.

[42] Haijun Wang, Xinbo Gao, Kaibing Zhang, and Jie Li. Single-image super-resolution using active-sampling gaussian process regression. *IEEE TIP*, 2016.

[43] Zihan Wu, Rongbo Shen, Junzhou Huang, Liansheng Wang, and Jianhua Yao. Strongly supervised mitosis detection in breast histopathology images using weak labels. In *ISBI*, 2021.

[44] Shuai Xie, Zunlei Feng, Ying Chen, Songtao Sun, Chao Ma, and Mingli Song. Deal: Difficulty-aware active learning for semantic segmentation. In *ACCV*, 2020.

[45] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *CVPR*, 2019.

[46] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *CVPR*, 2016.

[47] Beichen Zhang, Liang Li, Shijie Yang, Shuhui Wang, Zheng-Jun Zha, and Qingming Huang. State-relabeling adversarial active learning. In *CVPR*, 2020.

[48] Fedor Zhdanov. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*, 2019.