Active Learning for Fine-grained Sketch-based Image Retrieval

Himanshu Thakur [a] Soumitri Chattopadhyay [b]

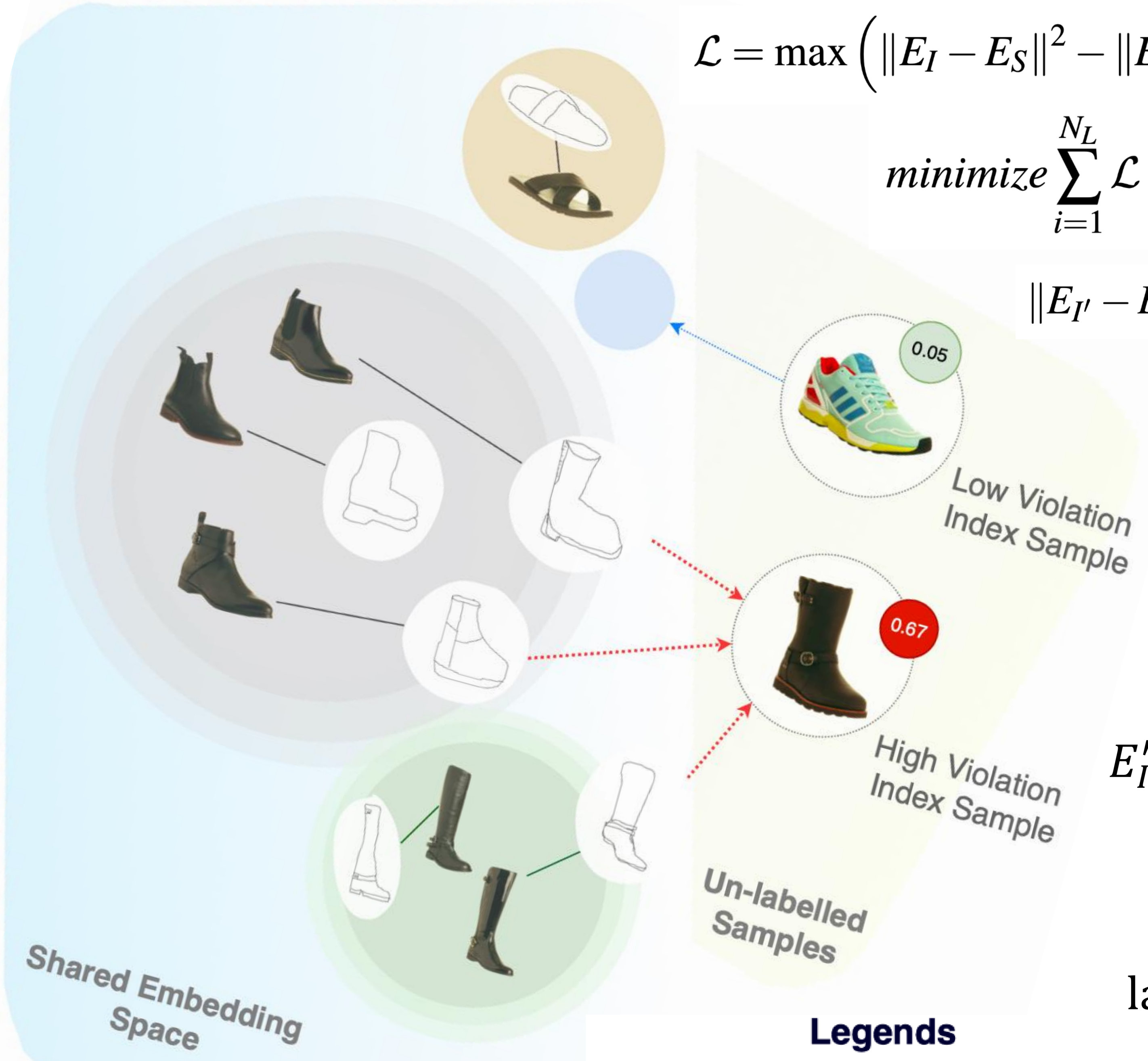
a. Language Technologies Institute, Carnegie Mellon University
b. Dept. of Computer Science, UNC Chapel Hill
(Both authors contributed equally to the paper)

BMVC 2023
The 34th British Machine Vision Conference
20th - 24th November 2023, Aberdeen, UK

Motivation and Challenges

- Drawing *faithful* sketches for sketch-photo pairing in fine-grained SBIR is an arduous task requiring great skills, and so, obtaining large-scale annotated sketch-photo pairs is a bottleneck.
- **Active Learning** is a possible solution to alleviate annotation bottleneck, which seeks to *find the smallest "most informative" subset* of data to be annotated which once added to the training dataset maximizes performance.
- However, off-the-shelf active learning methods are not suitable for FG-SBIR due to their intrinsic nature of drawing ***rigid*** decision boundaries, while the latter requires ***soft*** discrimination boundaries.
- Additionally, FG-SBIR learns a *joint sketch-photo embedding* space and hence, an. AL sampling would require handling of **both sketch and photo modalities.**

Understanding Violation Index



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

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

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

E_I = image embedding

E_S = sketch embedding

E_I' = new image embedding

N_L = size of labeled dataset

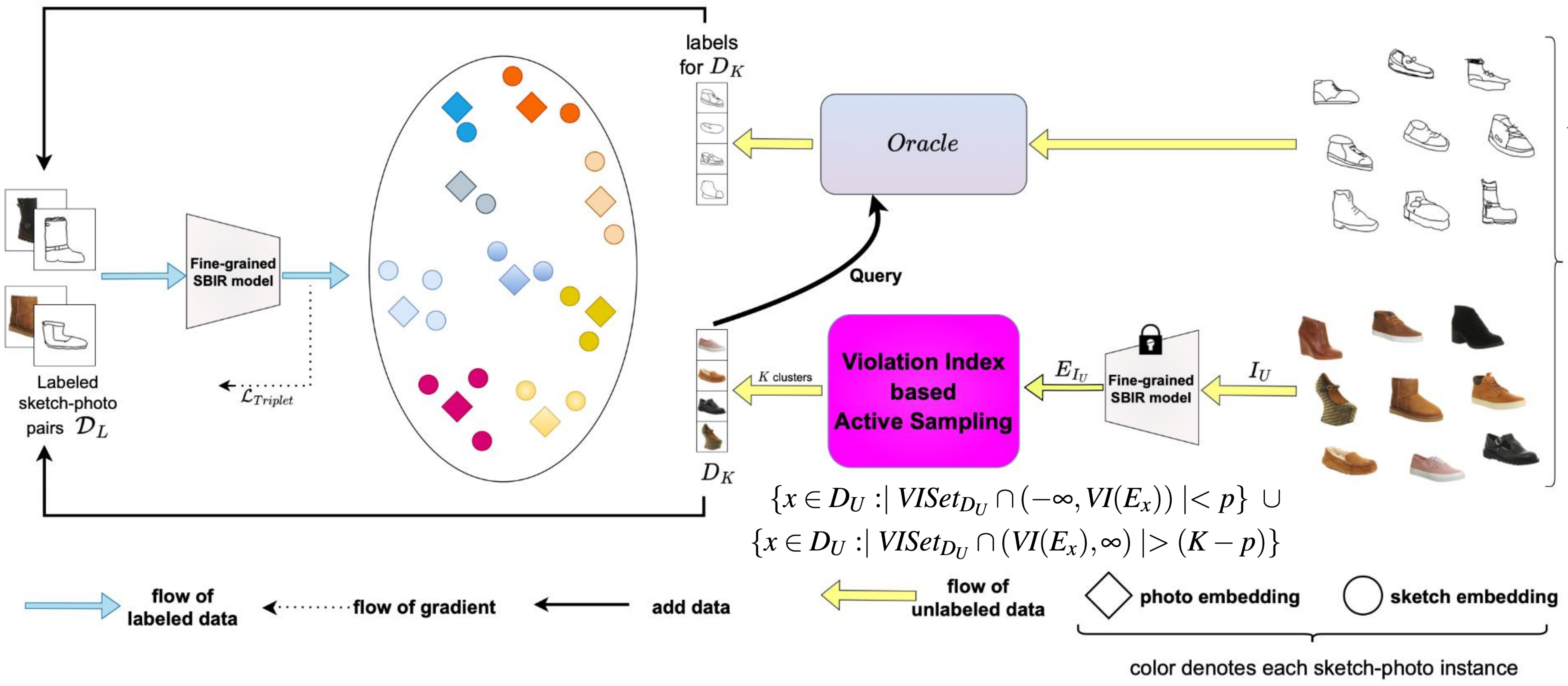
Legends

- 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

Proposed Method

We formulate a metric, **Violation Index** (VI), such that it quantifies the **degree of perturbation** an image from the unlabelled pool of images *produces* in the *existing embedding space* when introduced in an AL round.

- Images with **low violation index** (or **min violating** samples) => *relatively* ***unseen/novel*** *concept*; does not match with any *existing* sketch embeddings
- Images with **high violation index** (or **max violating** samples) => closely ***resemble one or more images*** from the *training set*; matches with multiple existing sketch embeddings *(uncertainty)*



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

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

- For selection strategy in AL, there always exists a **tradeoff** between *reducing existing uncertainty* and *learning novel instances*.
- To balance these counterintuitive effects, we consider a *combination of both* ***max-violating*** *(higher VI values) and* ***min-violating*** *(lower VI values)* using a hyperparameter α as the *mixing ratio*.

Experimental Setup

Datasets:

- QMUL-ShoeV2 and QMUL-ChairV2 (refer to paper for data splits)
- Initially, **300** sketch-photo pairs are considered as the training set

Implementation:

- Gold standard **Triplet** FG-SBIR model; VGG-16 encoder; N=5 cycles of AL

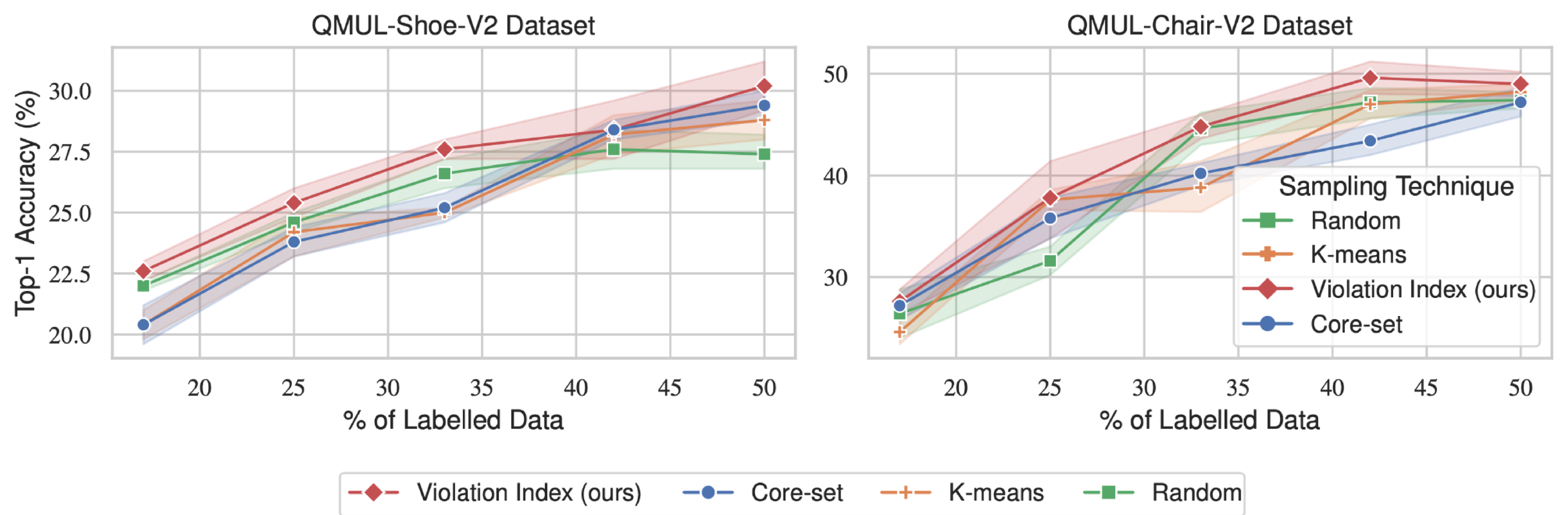
Evaluation metric:

- **Acc@q** i.e. percentage of sketches with true matched photo in the top-q list

SOTA Comparison

Comparison with adopted Active Learning baselines:

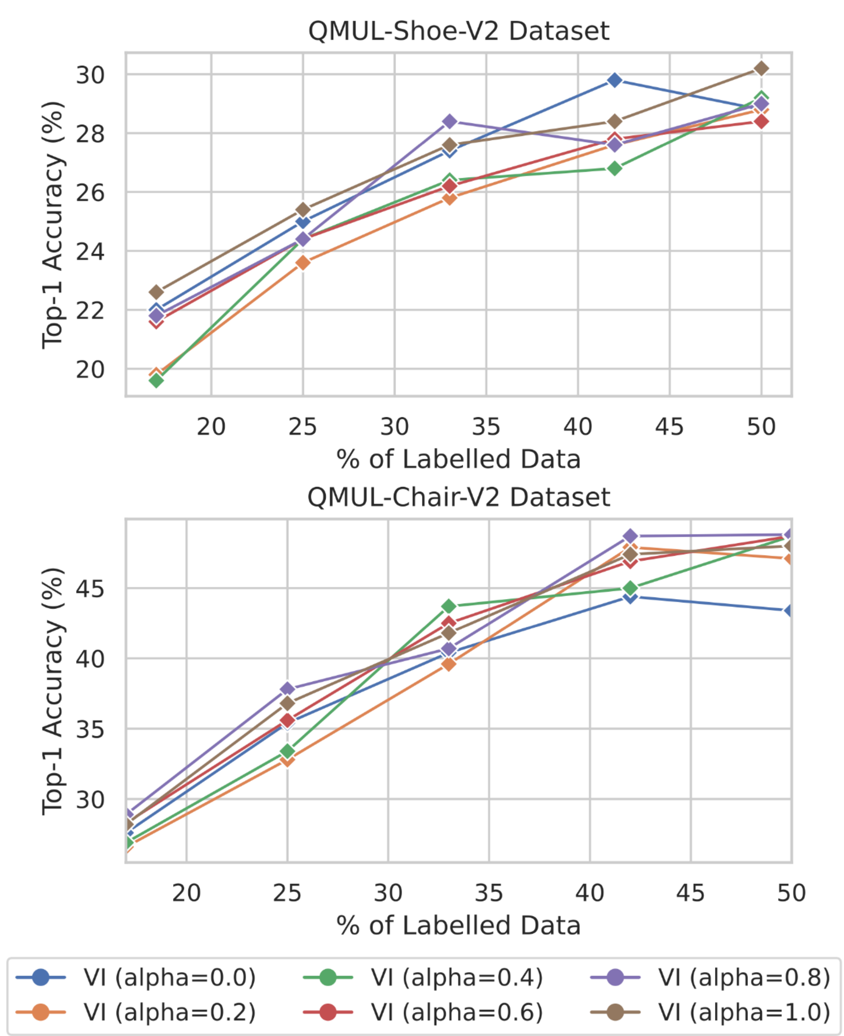
- Currently there exist no AL framework for FG-SBIR; we **adopt** some *classical AL sampling* strategies to our setup: Random, Core-set, K-means
- Our **VI**-based approach ***outperforms*** *these baselines* and performs well under **low-data regime** (fig. below; see paper for details)



Ablation Studies

Sensitivity to hyper-parameter α:

- We observe a **steady increase in acc.@1** as we increase α from 0 to 1
- There is a *sudden drop in performance* as we initially increase α from 0.0 to 0.1/0.2 (refer to adjacent figure)

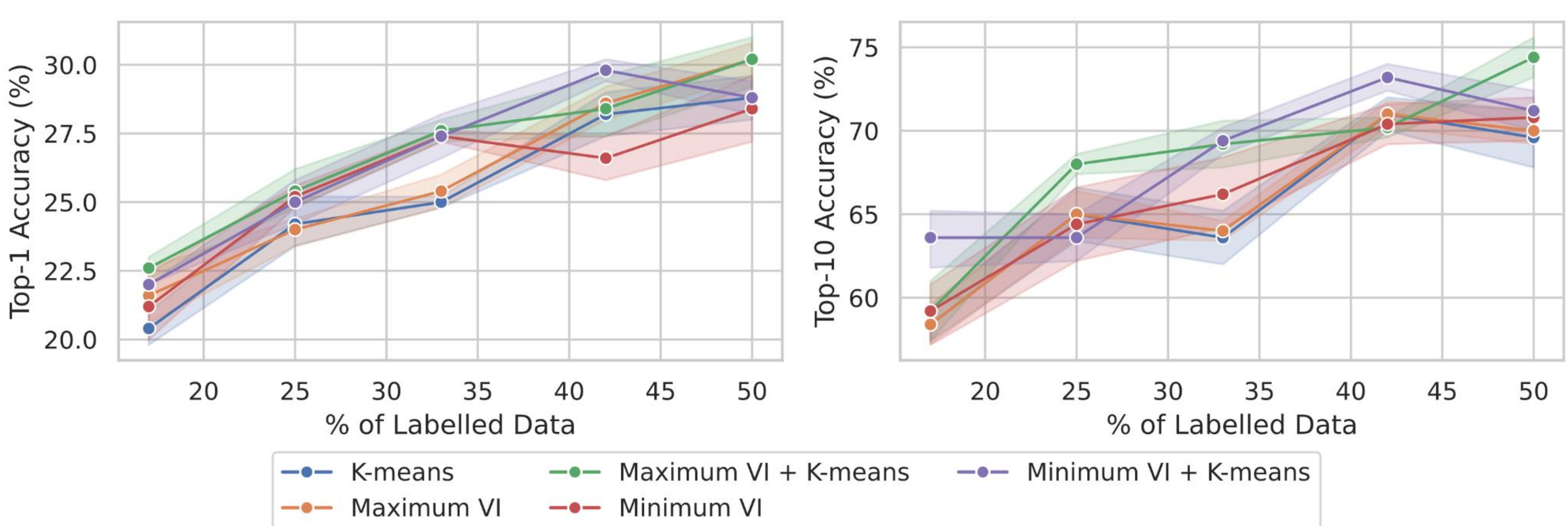


Significance of Violation Index:

- In early active learning cycles, selecting minimum VI performs *better* compared to maximum VI. This trend ***reverses*** with increase in training data (see fig. below)

Significance of Diversity sampling:

- Diverse clusters obtained by **kmeans++** consistently **outperforms** vanilla VI-baseline (see fig. below)



For more details, please refer to the arXiv version of our paper at: https://arxiv.org/abs/2309.08743 or email the authors at: hthakur@andrew.cmu.edu ; soumitri@cs.unc.edu | Thanks for visiting our poster!