Supplementary material for
CLIP for All Things Zero-Shot Sketch-Based Image Retrieval, Fine-Grained or Not

Aneeshan Sain[1,2] Ayan Kumar Bhunia[1] Pinaki Nath Chowdhury[1,2] Subhadeep Koley[1,2]
Tao Xiang[1,2] Yi-Zhe Song[1,2]
[1]SketchX, CVSSP, University of Surrey, United Kingdom.
[2]iFlyTek-Surrey Joint Research Centre on Artificial Intelligence.
{a.sain, a.bhunia, p.chowdhury, t.xiang, y.song}@surrey.ac.uk

A. Prompt Design for ZS-SBIR and ZS-FG-SBIR

Here we show in detail the way visual prompts are incorporated into the Image Encoder of CLIP [5].

For **ZS-SBIR**, we have two separate CLIP-image-encoders for photo and sketch branch with sketch and photo prompts incorporated into the respective encoders. During training the entire CLIP model is kept frozen except the LayerNorm of transformer layers and the prompts themselves, as shown in Fig. 1.

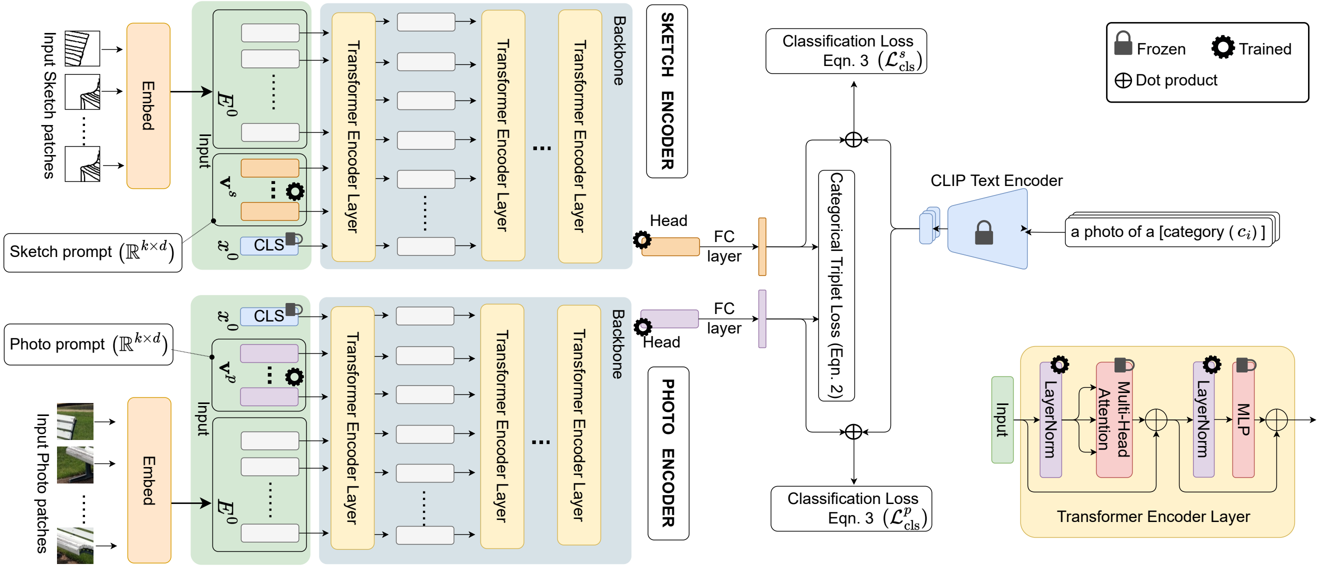
Input Sketch patches
Embed
E^0
Input
$\mathbf{v}^s$
x^0
CLS
Sketch prompt $(\mathbb{R}^{k\times d})$
Transformer Encoder Layer
Transformer Encoder Layer
...
Transformer Encoder Layer
Backbone
SKETCH ENCODER
Head
FC layer
Classification Loss Eqn. 3 $(\mathcal{L}^s_{\text{cls}})$
Categorical Triplet Loss (Eqn. 2)
CLIP Text Encoder
a photo of a [category (c_i)]
Frozen
Trained
⊕ Dot product
Photo prompt $(\mathbb{R}^{k\times d})$
x^0
CLS
$\mathbf{v}^p$
Input
E^0
Input Photo patches
Embed
Transformer Encoder Layer
Transformer Encoder Layer
...
Transformer Encoder Layer
Backbone
PHOTO ENCODER
FC layer
Head
Classification Loss Eqn. 3 $(\mathcal{L}^p_{\text{cls}})$
Input
LayerNorm
Multi-Head Attention
LayerNorm
MLP
Transformer Encoder Layer

Figure 1. Prompt Design for ZS-SBIR along with training objectives.

For **FG-ZS-SBIR**, we use one CLIP-image-encoder with *one common prompt* shared for both photo and sketch branches incorporated into the CLIP-image encoder. Similar to ZS-SBIR design, during training the entire CLIP model is kept frozen except the LayerNorm of transformer layers and the prompt itself, as shown in Fig. 2. Apart from shared image-encoders the main difference from ZS-SBIR is in considering hard-triplets within each category instead of category-level triplets as in ZS-SBIR. Furthermore, we have two more additional losses apart from the ones used for ZS-SBIR, aimed at (i) making the relative sketch-photo distances across categories uniform via f-divergence, and (ii) learning the structural correspondences between a sketch-photo pair via Patch-shuffling loss.

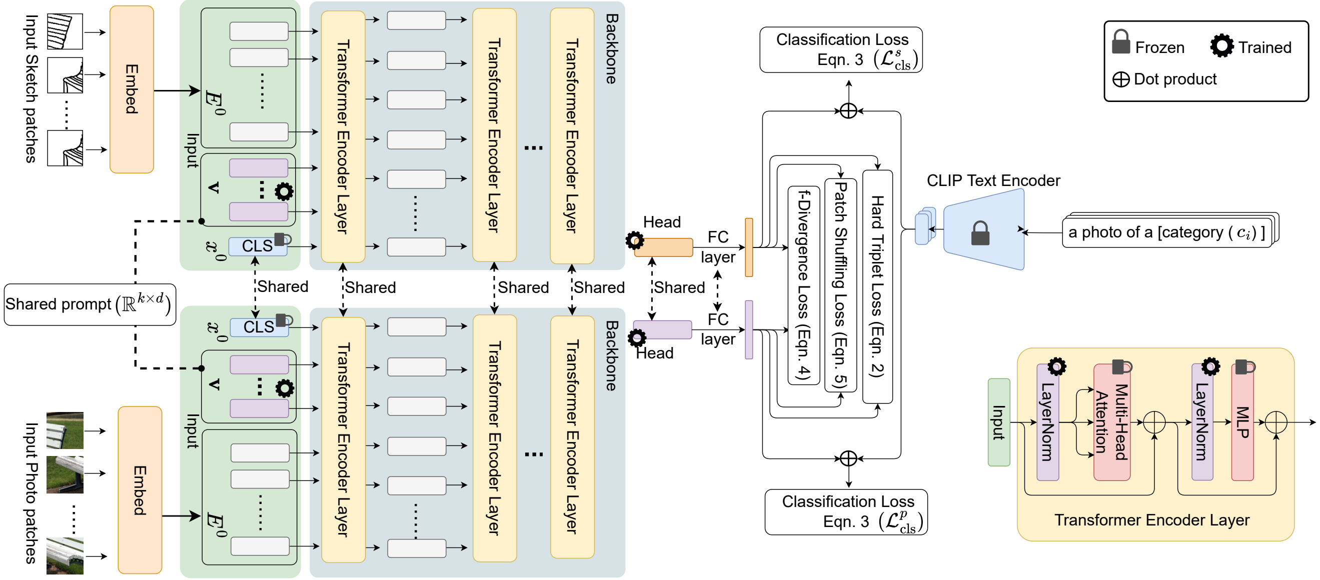
Input Sketch patches
Embed
E^0
Input
v
x^0
CLS
Transformer Encoder Layer
Transformer Encoder Layer
Transformer Encoder Layer
Backbone
Shared prompt ($\mathbb{R}^{k\times d}$)
Shared
Shared
Shared
Shared
Shared
Head
FC layer
FC layer
Head
Input Photo patches
Embed
E^0
Input
v
x^0
CLS
Transformer Encoder Layer
Transformer Encoder Layer
Transformer Encoder Layer
Backbone
Classification Loss Eqn. 3 ($\mathcal{L}^s_{\text{cls}}$)
f-Divergence Loss (Eqn. 4)
Patch Shuffling Loss (Eqn. 5)
Hard Triplet Loss (Eqn. 2)
Classification Loss Eqn. 3 ($\mathcal{L}^p_{\text{cls}}$)
CLIP Text Encoder
a photo of a [category (c_i)]
Frozen
Trained
⊕ Dot product
Input
LayerNorm
Multi-Head Attention
LayerNorm
MLP
Transformer Encoder Layer

Figure 2. Prompt Design for FG-ZS-SBIR along with training objectives.

B. Datasets

For evaluation on **ZS-SBIR** we have used three datasets:

(i) **Sketchy (extended)** [4] – Sketchy [7] contains 75,471 sketches over 125 categories having 100 images per category, with atleast 5 associated hand-drawn sketches per photo [9]. It was extended [4] further with extra 60,502 images from ImageNet [6] (Sketchy-ext), which we use here. Following [9] for zero-shot setup we split it as 104 classes for training and 21 for testing, ensuring that *test*-set images do not overlap with 1000 classes of ImageNet [6].

(ii) **TUBerlin [2]** – contains 250 categories, with 80 free-hand sketches in each, which was extended with a total of 204,489 images by [11]. We split it following [1] as 30 classes for testing and 220 for training.

(iii) **QuickDraw Extended** – The full-version contains over 50 million sketches across 345 categories, drawn by users across the internet under 20 seconds per sketch. Augmenting the sketches with images from *Flickr*, a subset of QuickDraw with 110 categories having 330,000 sketches and 204,000 photos was introduced for ZS-SBIR in [1]. We follow their split of 80 classes for training and 30 for testing to ensure no overlap of test-set photos from ImageNet [6].

For evaluation on **FG-ZS-SBIR** we require fine-grained (one-to-one matching) sketch-photo association [10] across categories for evaluation. Accordingly we resort to Sketchy [7] which has atleast atleast 5 associated hand-drawn sketches associated to every photo [9]. We use the same zero-shot categorical split of 104 training and 21 testing classes [9]. A few examples of sketch-photo association with multiple sketches per photo is illustrated in Fig. 3.

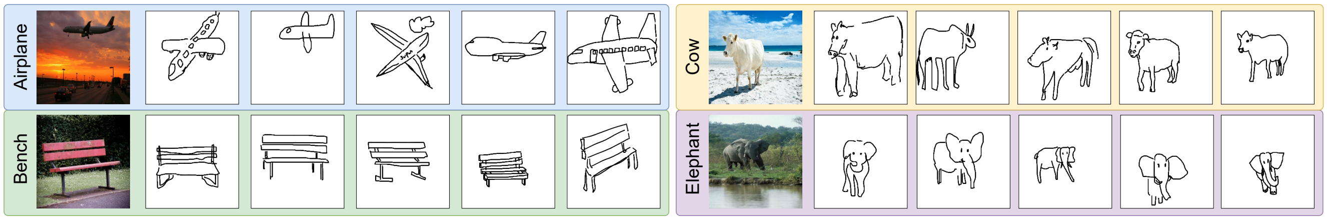

Figure 3. Some examples of Fine-grained associations across categories from Sketchy [7].

C. More on f-Divergence

To use a single (global) margin parameter μ that works for all categories, we impose a regulariser that aims to make the sketch-photo relative distance, defined as $\delta(s, p^+, p^-)$, uniform across categories. We achieve this by computing the distribution of relative distances for all triplets (s, p^+, p^-) in category c as $\mathcal{D}_c = \{\delta(s_i, p_i^+, p_i^-)\}_{i=1}^{N_s}$, where the c^{th} category has N_s sketch-photo pairs. Next, towards making the relative distance uniform across all categories, we minimise the KL-divergence [3] between a distribution of relative distances.

However, KL-divergence [3] only computes the information distance between two distributions – the length of the shortest program to describe a second distribution given the first. Comparing multiple (≥ 2) distributions however is comparatively less studied. The multi-distribution generalisation of information distance, aka., the *f-divergence* is defined by a convex function $f : [0, \infty) \to \mathbb{R}$. Despite its generalisation capability, f-divergence for multiple distribution setup is under-explored in computer vision applications. In this paper, we thus adopt a rather simplistic definition of f-divergence by Sgarro [8] – the average divergence, which is defined as,

$$\frac{1}{N_s(N_s-1)}\sum_{i=1}^{N_s}\sum_{j=1}^{N_s}\texttt{KL}(\mathcal{D}_i, \mathcal{D}_j) \tag{1}$$

D. Some Qualitative Results on Sketchy

Figures show qualitative results on Sketchy (ext) [4] for ZS-SBIR (Fig. 4) and on Sketchy [7] for FG-ZS-SBIR (Fig. 5), of baseline methods vs. ours. Baselines are constructed following [1] and [10] for ZS-SBIR and FG-ZS-SBIR respectively.

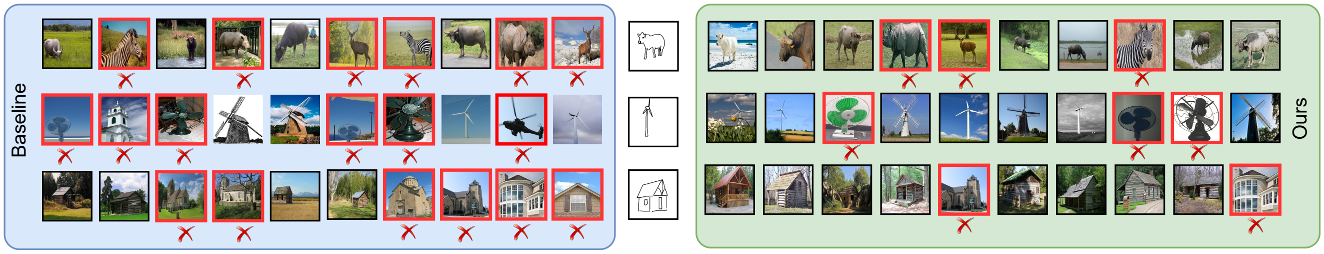

Figure 4. Qualitative results of ZS-SBIR on Sketchy [4] by a baseline (blue) method vs Ours (green).

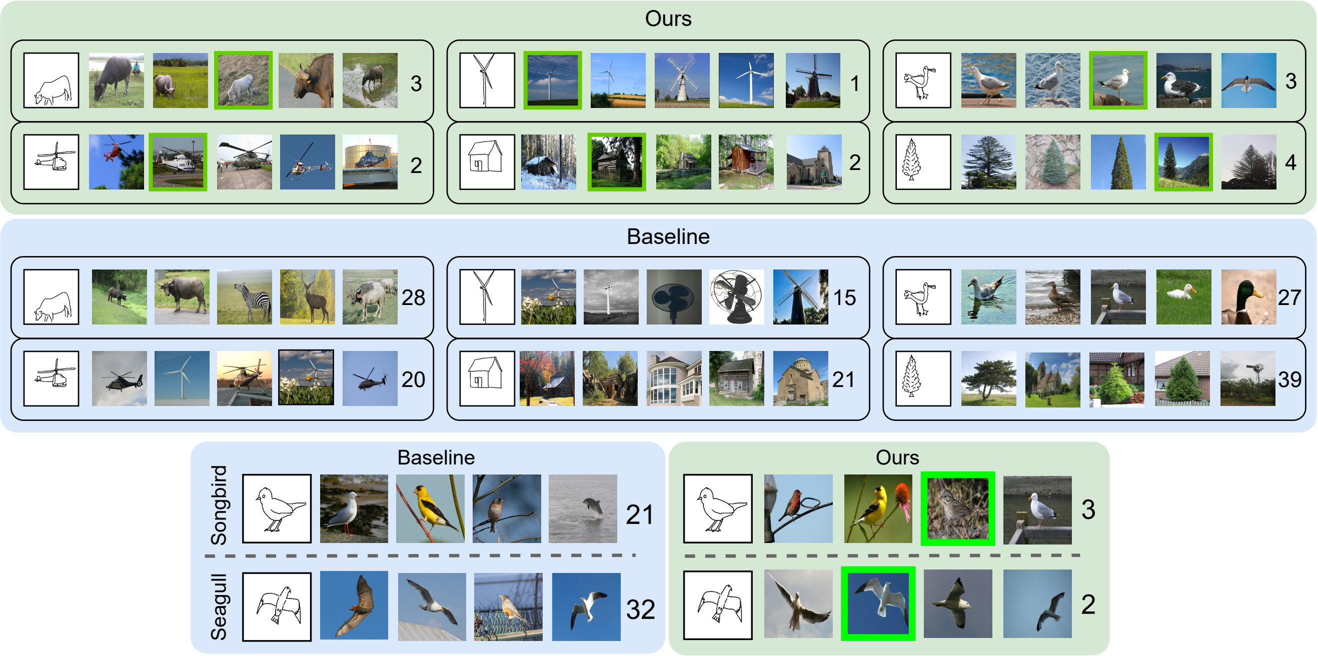

Figure 5. Qualitative results of FG-ZS-SBIR on Sketchy [7] by a baseline (blue) method vs Ours (green). The images are arranged in increasing order of the ranks beside their corresponding sketch-query, i.e the left-most image was retrieved at rank-1 for every category. The true-match for every query, if appearing in top-5 is marked in a green frame. Numbers denote the rank at which that true-match is retrieved for every corresponding sketch-query.

E. Limitations

We observed two plausible limitations of our method which we keep for addressal in a future work. *(i)* The assumption that CLIP covers almost all classes during training, might fail in certain niche cases. *(ii)* Being trained on internet-scale data (400M image-text pairs), thorough zero-shot evaluation on an unseen class is challenging. However both limitations are universal to all CLIP-based applications.

References

[1] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *CVPR*, 2019. 2, 3

[2] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM TOG*, 2012. 2

[3] Solomon Kullback and Richard Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 1951. 3

[4] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, 2017. 2, 3

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[6] 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. *IJCV*, 2015. 2

[7] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM TOG*, 2016. 2, 3

[8] Andrea Sgarro. Information divergence and the dissimilarity of probability distributions. *Calcodo*, 1981. 3

[9] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, 2018. 2

[10] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *CVPR*, 2016. 2, 3

[11] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. Sketchnet: Sketch classification with web images. In *CVPR*, 2016. 2