Supplementary material for
StyleMeUp: Towards Style-Agnostic Sketch-Based Image Retrieval

Aneeshan Sain[1,2] Ayan Kumar Bhunia[1] Yongxin Yang[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.

Additional explanations

Clarity on bridging domain gap:

From the viewpoint of bridging the domain gap, a gradient reversal layer is employed in Dey *et al*. [10], that is used to create a domain-agnostic embedding, which however does not differentiate if it comes from a sketch or a photo. Our motivation is different – in addition to tackling the sketch-photo domain gap, we further focus on narrowing the domain gaps that exist amongst different sketching styles (i.e., learning a style-agnostic embedding). In particular, the *feature transformation layer* helps bridge this style gap by simulating varying distributions in the intermediate layers of the encoder, and thus condition the encoder to *generalise onto unseen sketching styles*. The meta-learning paradigm further ensures that this notion of style variance is minimised over episodic training, finally resulting in a style-agnostic embedding.

Additional experimental comparison:

The results of DSH and GDH on Sketchy and TU-Berlin have been taken directly from their respective papers. For further transparency we re-run these baselines using Inception-V3 as backbone. Table 4 shows these results to be in line with our conclusions for Sketchy and TUBerlin datasets respectively –

Table 4. Quantitative analysis using Inception-V3 backbone

Method	Sketchy		TUBerlin	
	mAP	P@200	mAP	P@200
DSH	0.725	0.867	0.537	0.660
GDH	0.821	0.896	0.696	0.741
Ours	**0.905**	**0.927**	**0.778**	**0.795**

More on training details:

The hyperparameters $\lambda_{1\rightarrow3}$ have been determined empirically. The impact of $\mathcal{L}_{KL}$ is suppressed (λ_1=0.001) during initial stages of training, and increased with linear scheduling later for better training stability. We further observed that λ_2 works best if kept constant throughout. Changing λ_3 had generally produced comparatively lower results. Margin hyperparameters for triplet losses $\mu^{z_{inv}}$ and μ^{z_f} were set empirically as well. Please note that unlike few-shot adaption in MAML, there is no adaptation step here during inference. Instead, meta-learning is employed only during training to learn a style-agnostic feature encoder for better generalisation.

More on Fusing modal invariant and modal specific features:

Combining these two components helps the model in keeping important details that might have been removed during disentanglement, for image (sketch/photo) reconstruction. Furthermore, as we intend to learn *how to disentangle* modal-invariant feature from modal-specific one, combining them to obtain a proper reconstruction re-verifies that the disentanglement itself has been learned properly. However, experimental results suggested that element-wise addition performs better than concatenating the two components together. This is probably because the former establishes a clearer boundary between the disentangled components than concatenation.