Supplementary material for
How to Handle *Sketch-Abstraction* in Sketch-Based Image Retrieval?

Subhadeep Koley[1,2] Ayan Kumar Bhunia[1] Aneeshan Sain[1] Pinaki Nath Chowdhury[1]
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.
{s.koley, a.bhunia, a.sain, p.chowdhury, t.xiang, y.song}@surrey.ac.uk

A. Additional Qualitative Results

Figs. 1-6 depict additional qualitative retrieval results for various retrieval scenarios and datasets [1, 17] using our framework. To delineate the abstraction-agnostic behaviour of our method, we abstracted the input sketches using the GDSA [9] method at different abstraction budgets ($\{10, 30, 100\}\%$). Fig. 1 and Fig. 2 show how our method reasonably retrieves the ground truth paired photo even in the case of *extreme* abstraction of 10%.

Figure 1. Top-10 retrieved images for inputs abstracted (by [9]) at different budgets ($10\%, 30\%, 100\%$). Paired photo is red bordered.

Figure 2. Top-10 retrieved images for inputs abstracted (by [9]) at different budgets ($10\%, 30\%, 100\%$). Paired photo is red bordered.

Fig. 3 and Fig. 4 qualitatively depict our method's efficacy over Triplet-SN [23] for the case of different sketching styles (*i.e.*, good, reasonable, and abstract) of the same object. It is evident from Fig. 3 and Fig. 4 that the proposed method equipped with dynamic abstraction identification surpasses SoTA Triplet-SN [23] in every case.

Figure 3. Proposed (blue) method's efficacy over Triplet-SN [23] (green) against different sketching styles of the same shoe (red bordered).

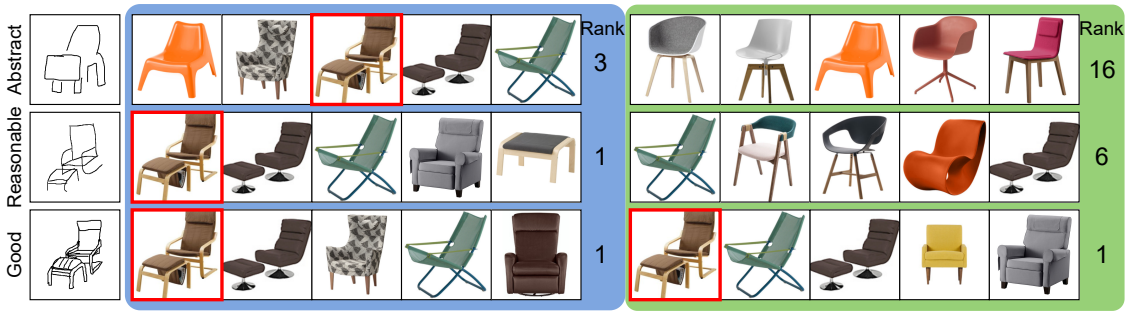

Figure 4. Proposed (blue) method's efficacy over Triplet-SN [23] (green) against different sketching styles of the same chair (red bordered).

Finally, Fig. 5 and Fig. 6 show qualitative retrieval results of the proposed method on sketches from ShoeV2 [1, 17] and ChairV2 [1, 17] datasets. It is worth noticing in Fig. 5 and Fig. 6 how the retrieved images *transition smoothly* from rank-1 to rank-10. This importantly ensures that most of the images retrieved by the proposed method are *semantically relevant* and *correspond to the input sketch*. We posit that this behaviour is driven by the regularisation provided by the disentangled and smooth latent space of StyleGAN [8].

Figure 5. Top-10 qualitative retrieval results of the proposed method on sketches from ShoeV2 dataset [1, 17]. Paired photo is red bordered.

Figure 6. Top-10 qualitative retrieval results of the proposed method on sketches from ChairV2 dataset [1, 17]. Paired photo is red bordered.

B. Quantitative Analysis of Dynamic Structural Latent Code Selection

To quantitatively demonstrate the relevance of the proposed dynamic structural latent code selection through the abstraction identification head ($\mathcal{A}$), we perform a few experiments. Here, instead of the automatic prediction of embedding matrix dimension via the $\mathcal{A}$ module, we force the system to always use either 3, 6, or 9 structural latent codes *regardless* of the input abstraction, thus resulting in a feature embedding matrix of size $\mathbb{R}^{3\times d}$, $\mathbb{R}^{6\times d}$, or $\mathbb{R}^{9\times d}$ respectively. Additionally, in another paradigm we enforce the model to *randomly* select between the feature embedding matrix of size $\mathbb{R}^{3\times d}$, $\mathbb{R}^{6\times d}$, or $\mathbb{R}^{9\times d}$. Experimental results in Tab. 1 depict how the accuracy falls drastically in cases of fixed or random latent selection. On the other hand, the proposed method equipped with *dynamic abstraction-modelling* outperforms them all with an Acc.@1 of $45.3\%(72.1\%)$ on ShoeV2 (ChairV2) dataset.

Table 1. Quantitative analysis of dynamic latent selection.

Embedding matrix dimension	ChairV2		ShoeV2	
	Acc.@1	Acc.@5	Acc.@1	Acc.@5
$\mathbb{R}^{3\times d}$	23.6	42.8	18.9	31.7
$\mathbb{R}^{6\times d}$	44.7	53.6	29.2	49.8
$\mathbb{R}^{9\times d}$	58.5	70.1	37.1	68.1
Random ($\mathbb{R}^{3\times d}$ / $\mathbb{R}^{6\times d}$ / $\mathbb{R}^{9\times d}$)	45.4	55.1	30.5	51.3
Ours-full	**72.1**	**80.9**	**45.3**	**77.3**
Avg. Improvement	*+29.0*	*+25.5*	*+16.3*	*+27.0*

C. Utility of Abstraction-aware Feature Matrix Embedding

Recent literature [7, 8, 10, 12, 22] motivates us to exploit the abstraction hierarchy present in the StyleGAN [7] latent matrix. To justify the same, we experiment by rendering the ShoeV2 test set sketches at different stages (25-35%, 55-65%, & 90-100%) to represent three abstraction levels and forcing the model to calculate the distance with $\mathbb{R}^{3\times d}$, $\mathbb{R}^{6\times d}$, and $\mathbb{R}^{9\times d}$ dimensional feature matrices *per level*. The resultant plot (Fig. 7) shows how the proposed matrix embedding achieves *optimum* Acc.@10 for each abstraction level when the distance is calculated with the *corresponding* matrix dimension by *traversing* the rows of the matrix embedding. This underpins our hypothesis that the feature matrix embedding can efficiently *accommodate* different abstraction levels.

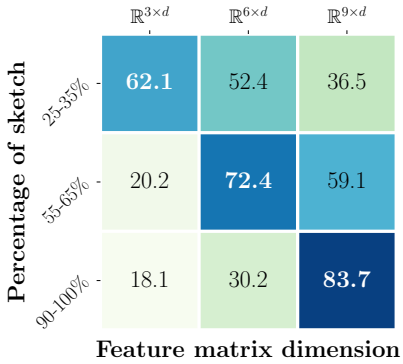

Figure 7. Acc.@10 comparison at different feature matrix dimension and sketch completion levels.

D. Choice of Backbone and Additional StyleGAN-based Baselines

In our approach, the pre-trained StyleGAN [8] is used *only* during training. During inference, we *discard* it and use the trained VGG-16 [15] feature extractor ($\mathcal{F}$), along with two sketch and photo-specific feature-matrix embedding networks $\mathcal{E}_s$ and $\mathcal{E}_p$ to calculate the matrix embeddings. However, following SoTAs [1, 2, 4, 13, 14], training an ImageNet-pretrained Inception-V3 [18] feature extractor ($\mathcal{F}$) we get a competitive Acc.@1 of $47.1(71.4)\%$ on ShoeV2 (ChairV2), thus validating our comparisons. On the other hand, to justify the proposed usage of a pre-trained StyleGAN [8] and for a fairer comparison, we amend existing SoTAs [4, 13, 14, 16, 23] with an additional StyleGAN-based regularisation. Here, given the respective SoTA backbones $\mathcal{F}(I) \in \mathbb{R}^{h\times w\times d}$, we employ 14 individual stride-two convolution blocks with LeakyReLU [21] applied over $\mathcal{F}(I)$ to convert $\mathbb{R}^{h\times w\times d} \rightarrow \mathbb{R}^{14\times d}$. This $\mathbb{R}^{14\times d}$ code upon passing through a pre-trained StyleGAN generates an image $\hat{p}$ for both photo ($\hat{p}_p$) and sketch ($\hat{p}_s$) branches, which we utilise to impose an additional reconstruction objective ($\mathcal{L}_{\text{recons}}$) apart from their respective losses.

$$\mathcal{L}_{\text{recons}} = \|p - \hat{p}_s\|_2 + \|p - \hat{p}_p\|_2 \tag{1}$$

This importantly ensures that the pre-trained StyleGAN's knowledge is distilled into the SoTA frameworks [4, 13, 14, 16, 23]. Although this trick boosts SoTA performance up to a certain extent (Tab. 2), the proposed method surpasses them with an average Acc.@1 of $5.6\%(12.2\%)$ on ShoeV2 (ChairV2). This further explains how the naive adaptation of StyleGAN fails to efficiently utilise the rich information residing in a pre-trained StyleGAN's latent space.

Table 2. Quantitative analysis of StyleGAN-based regularisation.

Methods	ChairV2		ShoeV2	
	Acc.@1	Acc.@5	Acc.@1	Acc.@5
Triplet-SN [23] + $\mathcal{L}_{\text{recons}}$	50.2	75.4	33.3	68.2
HOLEF-SN [16] + $\mathcal{L}_{\text{recons}}$	52.6	77.1	34.8	69.9
Partial-OT [4] + $\mathcal{L}_{\text{recons}}$	66.2	82.8	43.1	71.8
CrossHier [13] + $\mathcal{L}_{\text{recons}}$	64.9	80.9	42.8	72.3
StyleMeUp [14] + $\mathcal{L}_{\text{recons}}$	65.4	80.1	44.6	73.5
Ours-full	**72.1**	**80.9**	**45.3**	**77.3**
Avg. Improvement	*+12.2*	*+1.6*	*+5.6*	*+6.1*

E. Comparison with Other Surrogate Losses

Cross-modal retrieval is typically evaluated on three metrics – accuracy, precision, and recall [2, 14]. While *precision* and *recall* measure *how well* or *how many times* the model detected a certain *category* respectively, *accuracy* indicates the overall model performance irrespective of the category, thus making it the standard metric for instance-level fine-grained retrieval tasks [20]. Existing surrogate losses [3, 11] mostly optimise category-level metrics (*e.g.*, precision [3] or recall [11]), rendering them sub-optimal for our *fine-grained* setting. On the other hand, Engilberge et al. [5] proposed an LSTM-based network to learn ranking loss surrogates, but its adaptation has been limited in the consequent literature due to the alleged slow training [11]. Although SmoothAP [3] and Recall@k [11] have shown promising results in fine-grained datasets like INaturalist [19] and VehicleID [19], their off-the-shelf adaptation in our cross-modal fine-grained scenario produces sub-optimal Acc.@1 of 40.3% and 39.5% respectively in ShoeV2. On the other hand, the proposed `Acc.@q` loss being tailored for smooth approximation of the instance-level retrieval metric (*i.e.*, accuracy), outperforms existing SoTAs by a significant margin. More importantly, the parametric design of our `Acc.@q` loss allows us to use *different* variants (by changing $q = 1/5/10$) of the same loss to tackle *different* abstraction levels, which in turn provide better *retrieval granularity*.

F. Details on Human Study

Fig. 8 and Fig. 9 depict various UIs of the applet used to collect Mean Opinion Scores (MOS) [6] through a human study. After logging into the system, the participant first selects the category (*i.e.*, shoe or chair) of which class he/she wants to draw a sketch. Next, the user clicks on the "Draw" button to activate the drawing tool and starts drawing. Upon finishing, the participant clicks on the "Retrieve" button to view the images retrieved by all competing methods. The user rates every retrieved photo and clicks on "Submit & Next" to continue. We further sub-divide the MOS value levels (1(bad)→5(excellent)) into nine discreet levels (*e.g.*, $\{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5\}$) [6] for brevity and ease of rating. We *purposefully* anonymise the method names to prevent the rating from being influenced by the participant's prior knowledge of the literature.

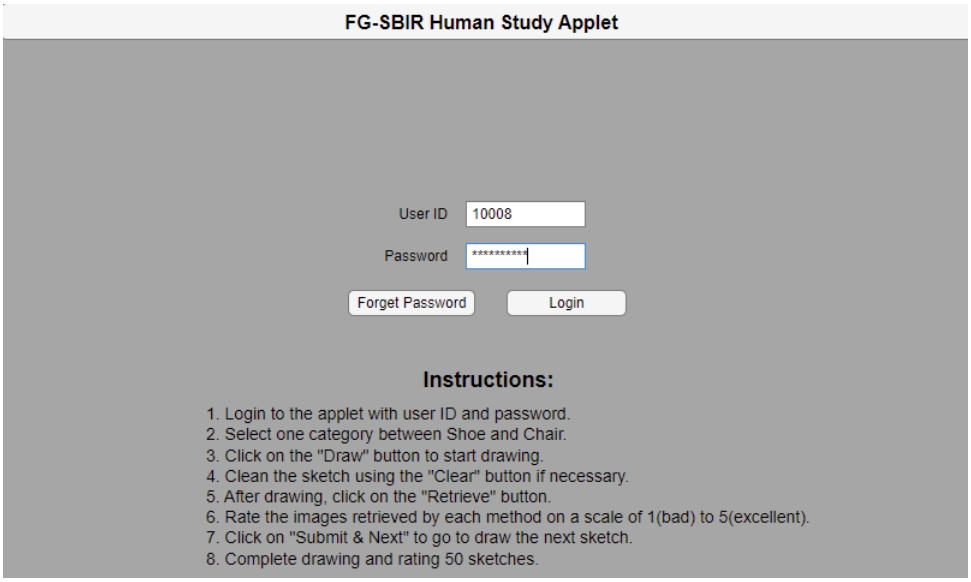

Figure 8. Login UI of the FG-SBIR human study applet

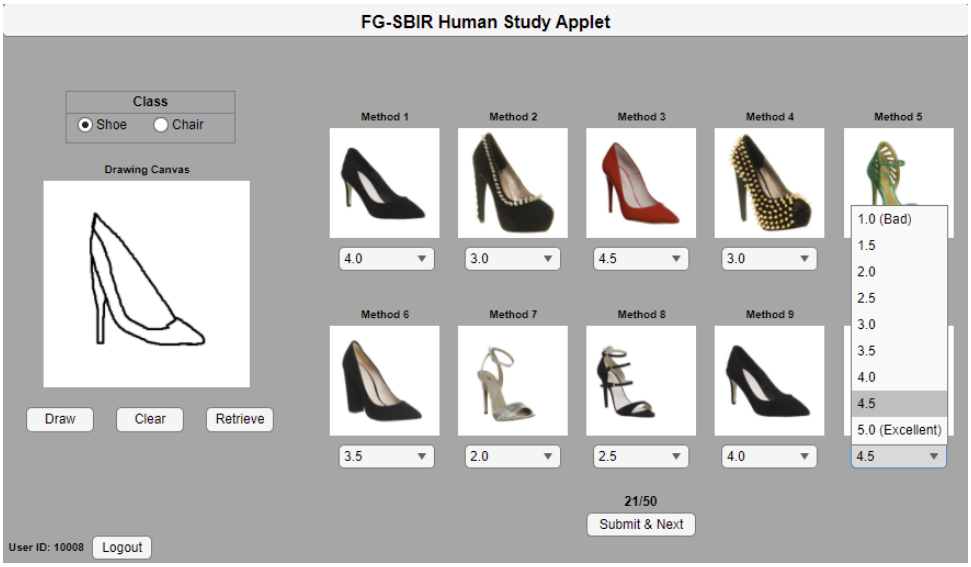

Figure 9. Scoring UI of the FG-SBIR human study applet

References

[1] 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. 1, 2, 3

[2] Ayan Kumar Bhunia, Subhadeep Koley, Abdullah Faiz Ur Rahman Khilji, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketching Without Worrying: Noise-Tolerant Sketch-Based Image Retrieval. In *CVPR*, 2022. 3, 4

[3] Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval. In *ECCV*, 2020. 4

[4] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Viswanatha Reddy Gajjala, Aneeshan Sain, Tao Xiang, and Yi-Zhe Song. Partially Does It: Towards Scene-Level FG-SBIR With Partial Input. In *CVPR*, 2022. 3, 4

[5] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. SoDeep: a Sorting Deep net to learn ranking loss surrogates. In *CVPR*, 2019. 4

[6] Quan Huynh-Thu, Marie-Neige Garcia, Filippo Speranza, Philip Corriveau, and Alexander Raake. Study of Rating Scales for Subjective Quality Assessment of High-Definition Video. *IEEE TBC*, 2010. 4

[7] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*, 2019. 3

[8] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. In *CVPR*, 2020. 2, 3

[9] Umar Riaz Muhammad, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Goal-Driven Sequential Data Abstraction. In *ICCV*, 2019. 1

[10] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *ICCV*, 2021. 3

[11] Yash Patel, Giorgos Tolias, and Jiří Matas. Recall@k Surrogate Loss With Large Batches and Similarity Mixup. In *CVPR*, 2022. 4

[12] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation. In *CVPR*, 2021. 3

[13] 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. 3, 4

[14] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. StyleMeUp: Towards Style-Agnostic Sketch-Based Image Retrieval. In *CVPR*, 2021. 3, 4

[15] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2015. 3

[16] 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. 3, 4

[17] Jifei Song, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Learning to Sketch with Shortcut Cycle Consistency. In *CVPR*, 2018. 1, 2

[18] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision . In *CVPR*, 2016. 3

[19] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist Species Classification and Detection Dataset. In *CVPR*, 2018. 4

[20] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE TPAMI*, 2021. 4

[21] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical Evaluation of Rectified Activations in Convolutional Network. *arXiv preprint arXiv:1505.00853*, 2015. 3

[22] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic Hierarchy Emerges in Deep Generative Representations for Scene Synthesis. *IJCV*, 2021. 3

[23] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch Me That Shoe. In *CVPR*, 2016. 1, 2, 3, 4