Supplementary material for You'll Never Walk Alone: A Sketch and Text Duet for Fine-Grained 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. Composed Fashion Image Retrieval

Although our method is primarily focused on *fine-grained* composed image retrieval, here we showcase an interesting application of our method namely, *sketch+text composed fashion image retrieval*. It aims to retrieve fashion garment images from abstract sketches and additional textual queries. We train and test our method on FashionIQ [5] dataset, having ∼15K paired triplets of query-image, textual description and target-image. Due to the lack of paired sketches, we generate synthetic sketches from query images using [3]. Compared to other baselines (Fig. 1), images retrieved by the proposed method hold more semantic coherence with the sketch+text composed query.

Figure 1. Qualitative comparison with baselines for sketch+text composed fashion image retrieval on FashionIQ [5]. GT photos are green-bordered. Notably, even though the images retrieved by B-Sketch+Text are mostly of the same shape as the query sketch, they lack the desired appearance given by textual description.

B. Details on Neutral Text Set

We aim to restrict the adverse effects of sketch-photo difference signal Δ to a minimum. For that, we regularise the training via a *"neutral-text"* set containing a list of 3-5 word *generic* descriptions of a freehand abstract sketch (Sec. 4.2), generated via a lightweight GPT [1]. The full neutral-text set is given below:

```
"in abstract lines",
"as sparse contours",
"with a line drawing",
"as an artwork",
"as a ballpoint sketch",
"in chirography",
"as a blueprint",
"in cartography",
"as a cartoon",
"in charcoal sketch",
"as conceptual drawings",
"in design sketches",
"as concept sketches",
"in professional sketches",
"as freehand sketches",
"as contour maps",
"as edge maps",
"of pencil curves",
"as a sparse diagram",
"as doodles",
"in drawing",
"in ink etching",
"as a deformed figure",
"as a line diagram",
"in geometrical drawing",
"containing sparse strokes",
"as graphical representations",
"containing hatching",
```

```
"as an illustration",
"as a sparse imitation",
"with sketch impression",
"as an ink sketch",
"as an ink doodle",
"as an ink drawing",
"in graphical layout",
"as line drawing",
"as line art",
"as line sketch",
"in chirographic representation",
"as monochrome drawing",
"with black and white outlines",
"in sketchy pattern",
"as a pencil drawing",
"as a pencil sketch",
"as a pencil doodle",
"containing pencil strokes",
"as perspective drawing",
"with hatching representation",
"as a contour plot",
"in sketch portrayal",
"as preliminary drawing",
"as preliminary sketch",
"in stroke rendering",
"as doodle representation",
"as rough draft",
"as rough drawing",
"as a rough outline",
"in contour schematic",
"as a scribble",
"as a rough shape",
"in silhouette",
"as a skeleton drawing",
"as a stipple",
"containing line traces",
"as a line tracing",
"as rough doodle",
"with artistic rendering",
"as a graphite sketch",
"as a freeform drawing",
"as a quick sketch",
"with pen and ink drawing",
"as an unconstrained sketch",
"as a rapid sketch",
"as an impromptu drawing",
"having loose pen strokes",
"as a rough sketch",
"as a gesture drawing",
"as a cartoon",
"as an abstract sketch",
"as a thumbnail sketch",
"as a hand drawn illustration",
"as fine art",
"as imaginary sketch",
"as casual drawing",
"as spontaneous sketch",
"as organic drawing",
"as experimental drawing",
"as an unstructured sketch",
"as a naturalistic drawing",
"as a contour drawing",
"as expressive sketch",
"containing whimsical sketch strokes",
"as an artistic drawing",
"with uninhibited sketch strokes",
"as minimalist drawing",
"as a free-flowing drawing",
"as a symbolic sketch",
"having unrestricted strokes",
"as a playful doodle",
"with monochrome brush strokes",
```

C. Details on Handcrafted Prompts

Our *text-to-text* generalisation loss (Sec. 4.3) enforces the learned prompts to be similar to handcrafted English language prompts, in the text embedding space for better generalisation. In particular, at every instance, we randomly pick one handcrafted fixed language prompt from the below set, which we curate from existing prompt learning literature [2, 4, 6].

```
"a photo of",
"an image of",
"itap of a",
"a photo of the hard to see",
"a low resolution photo of the",
"a rendering of a",
"a bad photo of the",
"a photo of a person doing",
"a high-resolution photo of",
"an origami of",
"a cropped photo of the",
"a pixelated photo of the",
"a bright photo of the",
"a close-up photo of the",
"a low resolution photo of a",
"a rendition of the",
"a clear photo of the",
"a blurry photo of a",
"a pixelated photo of a",
"itap of the",
"a photo of the small",
"a photo of the large",
"a black and white photo of",
"an art of the",
"a photo of a",
```

```
"a photo of many",
"a cropped photo of a",
"a photo of the",
"a good photo of the",
"a rendering of the",
"a photo of a large",
"a jpeg corrupted photo of the",
"a good photo of a"
"a photograph of a"
"a 4K photo of"
"a photorealistic image of a"
"a snapshot of"
"a good picture of"
"a bad picture of"
"a sharp photograph of"
"a blurry photograph of"
```

D. Full List of Connecting Words

In this paper, we represent an input sketch as a *pseudo-word token* that emulates its visual concept in equivalent word-embedding space. Subsequently, we combine it with the textual description via connecting phrases to obtain "⟨pseudo-word token⟩ ⟨connecting phrase⟩ ⟨text description⟩" that forms our composed query. The full list of such connecting phrases is given below:

```
is, in, having, are, and, with, upon,
on, containing, comprising, of, beside,
including, over, under, alongside,
plus, without, above, below, as.
```

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-Shot Learners. *NeurIPS*, 2020. 1

[2] Adrian Bulat and Georgios Tzimiropoulos. LASP: Text-to-Text Optimization for Language-Aware Soft Prompting of Vision & Language Models. In *CVPR*, 2023. 2

[3] Caroline Chan, Frédo Durand, and Phillip Isola. Learning to generate line drawings that convey geometry and semantics. In *CVPR*, 2022. 1

[4] 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. 2

[5] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback. In *CVPR*, 2021. 1

[6] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional Prompt Learning for Vision-Language Models. In *CVPR*, 2022. 2