Supplementary Material for Modeling the Visual Ambiguity of Human Sketches

Yang Zhou[1] Ping Ni[1] Jin Wang[1*] Senyun Jia[1] Jingdan Yan[1] Kaixiang Huang[1]
Guodong Lu[1] Jingru Yang[2] Shengfeng He[3]
[1]Zhejiang University, Hangzhou, China
[2]Carnegie Mellon University, Pennsylvania, USA
[3]Singapore Management University, Singapore
{22260043, 22425047, dwjcom, 22360562, 22460668, kaixianghuang, lugd}@zju.edu.cn
jingruy@andrew.cmu.edu shengfenghe@smu.edu.sg

A. Proof of Visual Ambiguity

We randomly choose 8 categories and count the semantic cues contained in each image within the existing SBIR dataset to verify the existence of semantic ambiguity. As shown in Fig. 1, the mean number of semantics present in an image for some categories in the popular ZS-SBIR datasets Sketchy [7] and TU-Berlin [3] is larger than 3, and for some categories, it is even larger than 9. This result indicates that each image contains at least three categories of semantic objects.

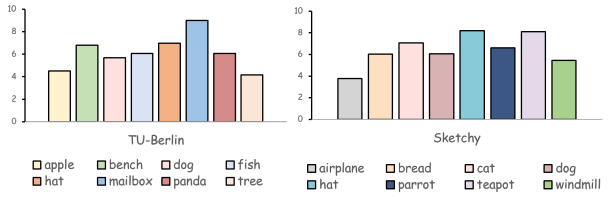

Figure 1. The semantic statistics on Sketchy and TU-Berlin are derived via a standard DeepLabv3 [2] semantic segmentation model pre-trained on the COCO-Stuff dataset [1] (background instance is not included in the count).

B. More on AmbiScore

We test several different semantic prompts for quantifying semantic ambiguity in single images, including

"`a photo of [category]`",
"`a picture of [category]`",
"`this is a photo of [category]`",
"`a close-up photo of [category]`".

The results are shown in Fig. 2. It can be observed that different handcrafted prompts indeed yield different AmbiScores. We introduce a temperature parameter τ to mitigate the sensitivity of the softmax to varying data distributions, ensuring that the fluctuations of AmbiScore remain within a reasonable range. Using the mean or the maximum value as the final AmbiScore is an alternative option, but it incurs additional computational cost.

Photo	Prompt	AmbiScore
	A photo of dog	0.6741
	A picture of dog	0.6777
	This is a photo of dog	0.6902
	A close-up photo of dog	0.7517
	A photo of pear	0.7964
	A picture of pear	0.8136
	This is a photo of pear	0.8382
	A close-up photo of pear	0.7891
	A photo of hat	0.9791
	A picture of hat	0.9538
	This is a photo of hat	0.9869
	A close-up photo of hat	0.9604

Figure 2. AmbiScore produced by different handcrafted prompts.

C. Additional `Person` Prompt

We observe that human-related content frequently appears in existing datasets, often introducing unintended semantic noise, particularly when the target category is non-human. We statistically calculate the proportion of human semantics appearing in each category in the Sketchy dataset as shown in Fig. 3. Specifically, we calculate the AmbiScore for every image using the prompt "`a photo of person`" and count how many images per category exceed a threshold of 0.25. It can be observed that a considerable

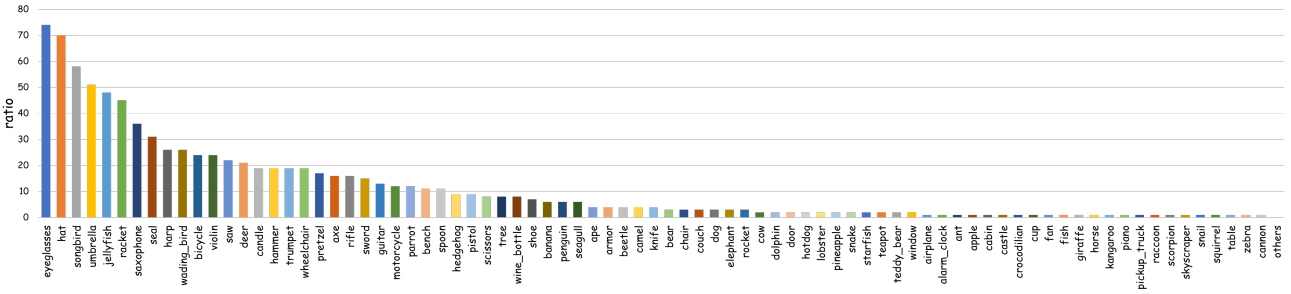

Figure 3. Statistics on the proportion of noise in the *person* category within the Sketchy dataset.

number of categories exhibit an extremely high proportion of "person" noise. For instance, the eyeglasses category exceeds 70%. This reveals non-trivial cross-category human interference, especially in categories such as "scissors" or "guitar", where human hands or figures frequently appear. Consequently, to ensure robust ambiguity estimation, we explicitly incorporate the person prompt during AmbiScore computation, even for datasets or categories that do not include person as a semantic label. This helps account for spurious human cues and improves the fidelity of ambiguity assessment.

D. Details of *Sketchy-Q* Dataset

To quantify sketch quality, Yang *et al.* [9] proposed the geometric aware classification layer, which replaces the dense/softmax combination to predict a quality score between 0 and 1 through classification. However, as this method relies on stroke sequences and lacks subjective human evaluation, we adopt a more concise approach.

The Sketchy dataset [7] provides five quality labels: (1) correct, (2) contains environment details or shading, (3) incorrect pose or perspective, (4) ambiguous, and (5) erroneous. Ignoring "incorrect pose or perspective" and "contains environment details or shading," we redefine "correct" as good (1,137 sketches), "ambiguous" as medium (1,137 sketches), and "erroneous" as bad (917 sketches). Using these labels, we construct a dataset for quality prediction and train MobileNetV3 [4], following the implementation details of Liu *et al.* [5], achieving a final recognition accuracy of 90.85% (with 10% of sketches for validation).

The trained model predicts sketch quality across three SBIR datasets: Sketchy Ext, TU-Berlin, and QuickDraw. As shown in Fig. 4, QuickDraw sketch creators do not have any professional drawing skills, and the drawing time is limited to 20 seconds, resulting in most sketches being rough and abstract. TU-Berlin sketch creators are given 30 minutes to draw, with a potential list of images provided as prompts, leading to higher sketch quality. The goal of Sketchy is to create fine-grained and instance-level associations with natural images. Sketch creators must base their sketches on specific photos to ensure sufficient similarity, with no time limit on drawing, resulting in the highest sketch quality.

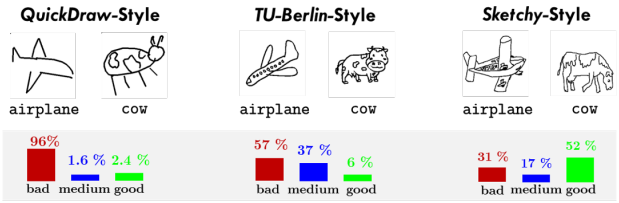

Figure 4. Sketch datasets with different sketch qualities.

E. Visualization of Purified Latent Space

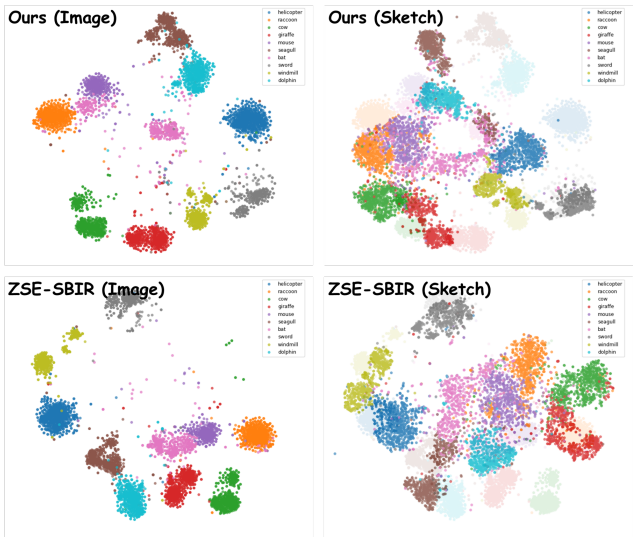

Figure 5. Visualized t-SNE results. Image representations in the sketch latent space are preserved and lightened for comparison.

Our objective is to improve the model's ability to distinguish sketch and image categories in the latent space. We use t-SNE [8] to visualize the sketch–image embeddings learned by DisAmb (based on ViT-B/16) on the Sketchy Ext validation set, and we compare them with those of the state-of-the-art ZSE-SBIR [6] (also based on ViT-B/16). We randomly select 10 categories for illustration, as shown in Figure 5. DisAmb enhances the separability of semantically

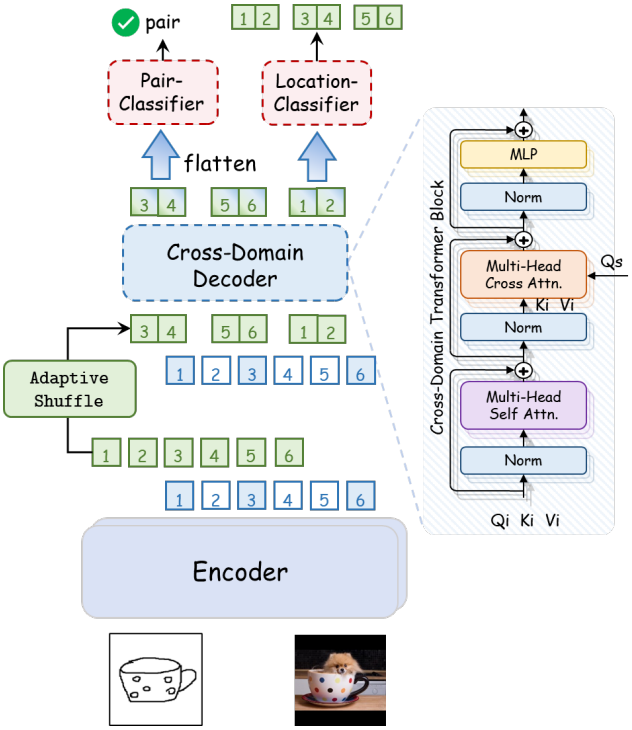

Figure 6. Detailed model architecture of shape jigsaw.

Figure 7. Ambiguous images and the corresponding AmbiScore.

ambiguous challenging samples, increases the margin between classes, and accordingly reduces misclassifications. For example, "bat" and "rat" are highly similar in semantics, and ZSE-SBIR tends to mix them in the embedding space. In contrast, our method achieves better discrimination between these categories under the zero-shot setting. Furthermore, DisAmb significantly reduces outlier dispersion and promotes tighter intra-class clustering, as observed in the "windmill" category. As for sketch representations, our method also demonstrates better inter-class separability and intra-class compactness.

F. Details of Shape Jigsaw

We add more details of the *shape jigsaw* (Fig. 6) for reproducibility. Given a sketch token sequence $\mathbf{v}_s \in \mathbb{R}^{l\times d}$ (excluding the `[CLS]` token), we first reshape the tokens into a 2D spatial grid of size $H \times W$, where $H = W = \sqrt{l}$. To obtain a spatially coherent partition, we apply adaptive average pooling to divide the grid into $K = k \times k$ segments: $\mathbf{v}^{grid} = \text{AdaptivePool}(x_s) \in \mathbb{R}^{K\times D}$. Each pooled token corresponds to one local region of the sketch.

We then generate a random permutation π and construct the shuffled sketch representation:

$$\mathbf{v}^r_{s(i)} = \mathbf{v}^{grid}_{\pi(i)}, \quad i = 1, \ldots, K, \tag{1}$$

and record the ground-truth permutation indices $l_{\text{gt}} = \pi$ for supervision. Given the masked image features $\mathbf{v}_{\text{mask}} \in \mathbb{R}^{l\times d}$, the shuffled sketch feature $\mathbf{v}^r_s$ serve as queries in our multi-layer cross-attention module. We classify whether the shuffled sketch and image share the same category using a binary classifier. For positive pairs, the model is further required to recover the original spatial order of the shuffled segments.

G. More Ambiguous Cases

We show more ambiguous images and their AmbiScore in Fig. 7.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 1

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[3] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Transactions on graphics (TOG)*, 31(4):1–10, 2012. 1

[4] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu,

Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 2

[5] Hanhui Li, Xudong Jiang, Boliang Guan, Ruomei Wang, and Nadia Magnenat Thalmann. Multistage spatio-temporal networks for robust sketch recognition. *IEEE Transactions on Image Processing*, 31:2683–2694, 2022. 2

[6] Fengyin Lin, Mingkang Li, Da Li, Timothy Hospedales, Yi-Zhe Song, and Yonggang Qi. Zero-shot everything sketch-based image retrieval, and in explainable style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23349–23358, 2023. 2

[7] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016. 1, 2

[8] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 2

[9] Lan Yang, Kaiyue Pang, Honggang Zhang, and Yi-Zhe Song. Annotation-free human sketch quality assessment. *International Journal of Computer Vision*, pages 1–22, 2024. 2