SceneTrilogy: On Human Scene-Sketch and its Complementarity with Photo and Text. (Supplementary Material)

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

A. Additional Details for Subjective Captioning

We provide additional details of our pilot study in Sec. 3.2 that compare the performance of subjective captioning when using part-of-speech (POS) [4], mouse trace [8] or sketch as a guiding signal into the image captioning pipeline. Instead of choosing a common baseline to compare subjective captioning when using POS, mouse trace, and sketches, we measure the relative performance over the standard baselines used in recent literature to study the contribution of every guiding signal. (i) For POS [4], we measure the relative performance using Wang *et al*. [17] as baseline. Without using POS, i.e., (w/o)-POS gives a B-4/C score of $31.1/100$ as compared to with POS, i.e., (w)-POS that gives $31.6/104/5$. (ii) For mouse trace [4], we use [10] to get (w/o)-Trace B-4/C score of $8.1/29.3$ as compared to (w)-Trace score of $24.6/106.5$. This leads to a large relative improvement of $16.5/77.2$ to show the significant contribution of using mouse trace as guiding signal. (iii) For sketch, we follow [3] to use [6] as baseline to get (w/o)-Sketch B-4/C score of $31.8/42.7$. We use cross-attention mechanism in [8] to inject sketch as a guiding signal into our baseline [7] to give a (w)-Sketch score of $42.7/121.6$. This gives a relative improvement of $10.9/16.1$, which shows that sketch as a guiding signal is better than POS and competitive as mouse trace. Hence, we advocate for sketch as a guiding signal to depict saliency since unlike POS [4] or mouse trace [8], sketches are more expressive that can capture artistic interpretation like caricature [5].

B. Modelling more than three modalities ($M > 3$)

Sec. 4.4 optionally models the modality-agnostic components of sketch or text using the function $\mathcal{G}(\cdot)$ that consists of a multihead cross-attention module $\texttt{MH}(\cdot)$ followed by an attention-based pooling $\texttt{PMA}(\cdot)$. For $M = 3$, $\mathcal{L}_{cls}^{tot}$ is defined as,

$$\begin{aligned}\mathcal{L}_{cl}^{tot} =& \mathcal{L}_{cl}(\mathcal{G}(f_{\mathbf{s}}^{ag}, f_{\mathbf{t}}^{ag}), f_{\mathbf{p}}^{ag}) \\ &+ \mathcal{L}_{cl}(\mathcal{G}(f_{\mathbf{s}}^{ag}, f_{\mathbf{p}}^{ag}), f_{\mathbf{t}}^{ag}) + \mathcal{L}_{cl}(\mathcal{G}(f_{\mathbf{p}}^{ag}, f_{\mathbf{t}}^{ag}), f_{\mathbf{s}}^{ag})\end{aligned} \tag{6}$$

In this section, we show how $\mathcal{G}(\cdot)$ can be extended to more than three modalities $M > 3$. Given a set of modality-agnostic components as $\Psi = \{f_{\mathbf{1}}^{ag}, f_{\mathbf{2}}^{ag}, \ldots, f_{\mathbf{M}}^{ag}\}$, we can solve for $\mathcal{L}_{cl}^{tot}$ as,

$$\mathcal{L}_{cl}^{tot} = \sum_{j=1}^{\mathbf{M}} \mathcal{L}_{cl}(\mathcal{G}(\Psi - \{f_j^{ag}\}), f_j^{ag}) \tag{7}$$

We further elaborate Eq. (7) using Algorithm 1.

Algorithm 1 Compute generalised $\mathcal{L}_{cl}^{tot}$ for $M > 3$

Require: $\mathcal{P} \in \mathbb{R}^{1\times 480}$ ▷ Learned weights.
$\Psi = \{f_{\mathbf{1}}^{ag}, f_{\mathbf{2}}^{ag} \ldots, f_{\mathbf{M}}^{ag}\}, \in \mathbb{R}^{M\times 480}$
$\mathcal{L}_{cl}^{tot} \leftarrow 0$
for $j \leftarrow 1$ to M **do**
  $\mathcal{S}_M \leftarrow \Psi - \{f_{\mathbf{i}}^{ag}\}$ ▷ $(M-1)\times 480$
  $H_M \leftarrow \texttt{MH}(\mathcal{S}_M)$ ▷ $(M-1)\times 480$
  $f_M = \texttt{PMA}(H_M) = \sigma(\mathcal{P}H_M^T)H_M$ ▷ (1×480)
  $\mathcal{L}_{cl}^{tot} \leftarrow \mathcal{L}_{cl}^{tot} + \mathcal{L}_{cl}(f_j^{ag}, f_M)$
end for
return $\mathcal{L}_{cl}^{tot}$

C. Derivation of Disentanglement Loss in Eq. 3

For optionality across tasks, we disentangle the information from sketch, text, and photo, given by $\mathbf{k} \in \{\mathbf{s}, \mathbf{t}, \mathbf{p}\}$ into a discriminative part $f_{\mathbf{k}}^{ag}$ shared across modalities, and a generative part specific to one modality $f_{\mathbf{k}}^{sp}$. This information split of $f_{\mathbf{k}} = [f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp}]$ is achieved in Sec. 4.3 by minimising the mutual information between the modality-agnostic and modality-specific components defined as,

$$\begin{aligned}\mathcal{I}(f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp}) &= \int_{f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp}} \mathbb{p}(f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp}) \log \frac{\mathbb{p}(f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp})}{\mathbb{p}(f_{\mathbf{k}}^{ag})\mathbb{p}(f_{\mathbf{k}}^{sp})} \\ &= \int_{f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp}} \mathbb{p}(f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp}) \log \frac{\mathbb{p}(f_{\mathbf{k}}^{sp}|f_{\mathbf{k}}^{ag})}{\mathbb{p}(f_{\mathbf{k}}^{sp})}\end{aligned} \tag{8}$$

Given a variational distribution $\mathfrak{q}(f_{\mathbf{k}}^{sp})$, due to positivity of KL divergence we have,

$$\int \mathfrak{p}(f_{\mathbf{k}}^{sp}) \log \mathfrak{p}(f_{\mathbf{k}}^{sp}) \geq \int \mathfrak{p}(f_{\mathbf{k}}^{sp}) \log \mathfrak{q}(f_{\mathbf{k}}^{sp}) \quad (9)$$

Hence, approximating the modality-specific prior $\mathfrak{p}(f_{\mathbf{k}}^{sp})$ with variational distribution $\mathfrak{q}(f_{\mathbf{k}}^{sp})$ in Eq. (8) we get,

$$\mathcal{I}(f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp}) \leq \int_{f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp}} \mathfrak{p}(f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp}) \log \frac{\mathbb{p}(f_{\mathbf{k}}^{sp} | f_{\mathbf{k}}^{ag})}{\mathfrak{q}(f_{\mathbf{k}}^{sp})} \quad (10)$$

Assuming a uniform prior distribution $\mathfrak{p}(\eta)$, and its definition in Eq. 2 via conditional invertible neural network $\tau_{\mathbf{k}}$, we have,

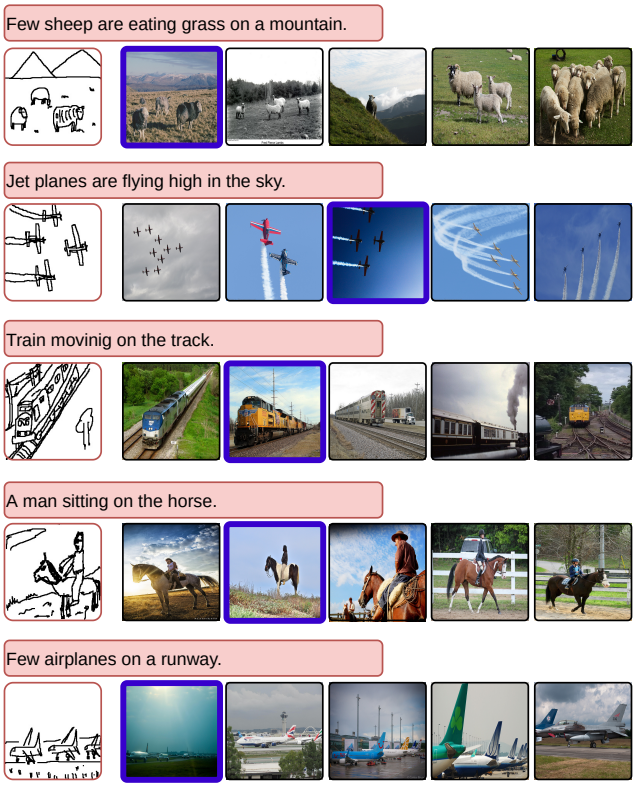
$$\begin{aligned} \mathcal{L}_{\tau_{\mathbf{k}}} = &- \mathbb{E}_{f_{\mathbf{k}}^{sp}, f_{\mathbf{k}}^{ag}} \{ \log \mathfrak{q}(\tau_{\mathbf{k}}^{-1}(f_{\mathbf{k}}^{sp} \mid f_{\mathbf{k}}^{ag})) \\ &+ \log |\det J_{\tau_{\mathbf{k}}^{-1}}(f_{\mathbf{k}}^{sp} \mid f_{\mathbf{k}}^{ag})| \} - H(f_{\mathbf{k}}^{sp} | f_{\mathbf{k}}^{ag}) \end{aligned} \quad (11)$$

where, $H(f_{\mathbf{k}}^{sp} | f_{\mathbf{k}}^{ag})$ is the constant data entropy which is ignored in the final optimisation in Eq. 3.

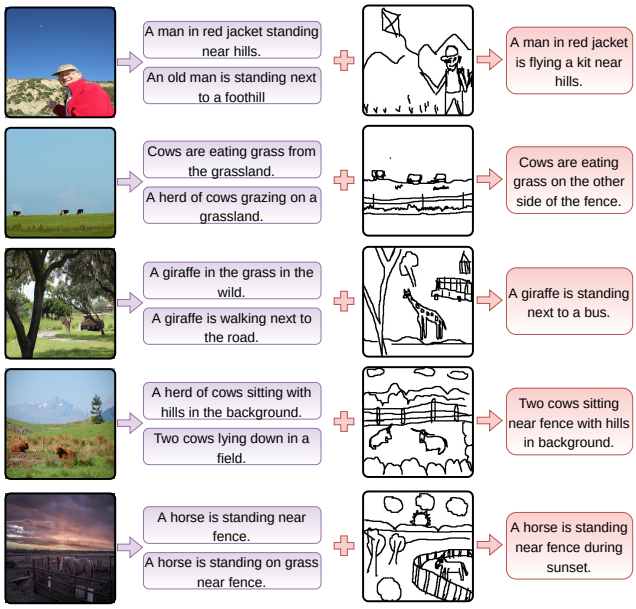
D. Comparison with a parallel work [12]

A parallel work surfaced while writing this paper by Sangkloy *et al*. [12] can optionally perform text-based image retrieval (TBIR), sketch-based image retrieval (SBIR), or both sketch+text based image retrieval (STBIR). However, the motivation of [12] is crucially different from ours. While we focus on improving the latent space via disentanglement into a modality-specific and modality-agnostic component to support optionality across tasks (retrieval and captioning) and modalities (using only sketch, only text, or both as query), Sangkloy *et al*. [12] focused on improving the encoders for sketch, text, and photo by adapting the recently popular pre-trained CLIP [11]. To model only sketch, only text, or both sketch+text for image retrieval, [12] used a rather simple late-fusion technique performing element-wise addition of sketch and text features. While the training code of the proposed model in [12] is not been released yet, our re-implementation of [12] using simple element-wise addition of sketch and text features with CLIP encoders lead to STBIR performance of $23.9/53.5$ in Acc.@1/Acc.@10 which is significantly lower than our proposed method by $15.6/35.2$ on FS-COCO [3]. Although CLIP [11] is highly generalisable to open-set setups, it is difficult to adapt to small downstream datasets like FS-COCO [3] and simultaneously outperform task-specific encoders like VGG-16 [13] used in the proposed method. A similar trend was also observed in Chowdhury *et al*. [3].

E. Sketch and Text as Query for Image Retrieval



F. Image Captioning v/s Subjective Captioning



G. Complex Faliure Cases

We show qualitative results below where sketch + text performs poorly. We observe this happens when both the input sketch or text is ambigious (i.e., badly drawn sketch or unprecise short textual phrases).

H. Clarification on Contributions

Our goal is not to design a model that is state-of-the-art for ALL retrieval (e.g., FG-STBIR, FG-SBIR, FG-TBIR) and generative (e.g., image, sketch, and subjective captioning) tasks. Instead, we (i) design a generalisable model that is competitive with a myriad of baselines (large models like CLIP-LN or small ones like VGG) across multiple tasks; (ii) we show how the benefits of sketch modality (acknowledged by several prior works [3, 15]) can be optionally combined with multiple modalities like text and photo.

I. Comparison with Matrix Factorization

While our baseline MulCap performs feature multiplication similar to matrix factorization [9, 16], we additionally adopt [16] to get subjective captioning (BELU-1, CIDEr) score of (79.2 ± 0.6, 113.5 ± 1.1).

J. Different training seeds and evaluation of Shoes

Training on 5 different seeds, we report accuracy on FG-STBIR task. For FS-COCO [3] we get Acc.@1 and Acc.@10 of 25.6 ± 0.5 and 55.3 ± 0.3 respectively. Further experimenting on shoe dataset [18], we get FG-STBIR Acc.@1 and Acc.@10 scores of 53.2 ± 0.5 and 88.1 ± 0.2.

K. Additional Details on Pilot Study

Our pilot study aims to: (i) compare sketch vs. text as a query for fine-grained image retrieval. For this, we use standard baselines Triplet-SN (for SBIR) and CLIP-LN (for TBIR) on 3000 sketch/photo, and text/photo pairs in FS-COCO [3]. We observe that for some instances sketch is a better query for image retrieval as it can depict complex shapes, multiple objects, and spatial alignment. However, not all objects are easy to draw (e.g., differentiate a 'donkey' vs. a 'horse') but could be easily described via text. (ii) For subjective captioning, we compare the relative improvements in standard captioning metrics (like M, R, C, S) when using users' sketch (to generate subjective captions) vs. without using sketches (to generate subjective captions).

L. Comparison with Aytar *et al.* [2]

Ayatar *et al.* [2] learns a joint embedding space across image, sound, and text. This is similar to our method, which also aims to learn a joint embedding space across image, sketch, and text. However, there are some key differences: (i) [2] lacks the ability to combine multiple modalities like sound+text for image retrieval. The ability to optionally combine multiple modalities for image retrieval is crucial to our motivation, e.g., fine-grained sketch-based image retrieval (FG-SBIR), fine-grained text-based image retrieval (FG-TBIR), and fine-grained sketch+text based image retrieval (FG-STBIR). (ii) The embedding space of [2] only supports discriminative tasks. This fails to support the generative objectives of our method, like image captioning, sketch captioning, and subjective captioning. Nevertheless, we compare Acc.@1 with [2] on FS-COCO [3] for FG-SBIR and FG-TBIR to get 23.5% and 7.1% respectively.

M. Differences from prior works

Prior works like (i) Aytar *et al.* [1] study only cross-modal transfer between a pair of modalities (sketch/photo, or text/photo), not a combination of multiple modalities (sketch+text, or sketch+photo) nor feature disentanglement (modality-agnostic and modality-specific) which is crucial for tasks like FG-STBIR and subjective captioning. (ii) Song *et al.* [14] combines sketch+text for image retrieval via a weighted sum of sketch-photo and text-photo distances computed independently. This simple setup is (a) limited to retrieval (i.e., no captioning), and (b) lacks feature disentanglement to filter our irrelevant modality-specific information (drawing style) when combining multiple modalities (sketch+text). We bring new insights into scene understanding by showing the need for feature disentanglement to (i) optionally combine multiple modalities, and (ii) support both discriminative and generative tasks.

References

[1] Yusuf Aytar, Lluis Castrejon, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Cross-modal scene networks. *IEEE TPAMI*, 2018. 3

[2] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, Hear, and Read: Deep Aligned Representations. *arXiv preprint arXiv:1706.00932*, 2017. 3

[3] Pinaki Nath Chowdhury, Aneeshan Sain, Yulia Gryaditskaya, Ayan Kumar Bhunia, Tao Xiang, and Yi-Zhe Song. Fs-coco: Towards understanding of freehand sketches of common objects in context. In *ECCV*, 2022. 1, 2, 3

[4] Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander Schwing, and D. A. Forsyth. Fast, diverse and accurate image captioning guided by part-of-speech. In *CVPR*, 2019. 1

[5] Xiaoguang Han, Kangcheng Hou, Dong Du, Yuda Qiu, Yizhou Yu, Kun Zhou, and Shugang Cui. Caricatureshop:

Personalized and photorealistic caricature sketching. *IEEE TVCG*, 2018. 1

[6] Shweta Mahajan, Iryna Gurevych, and Stefan Roth. Latent normalizing flows for many-to-many cross-domain mappings. In *ICLR*, 2020. 1

[7] Shweta Mahajan and Stefan Roth. Diverse image captioning with context-object split latent spaces. In *NeurIPS*, 2020. 1

[8] Zihang Meng, Licheng Yu, Ning Zhang, Tamara Berg, Babak Damavandi, Vikas Singh, and Amy Bearman. Connecting what to say with where to look by modeling human attention traces. In *CVPR*, 2021. 1

[9] Nils Murrugarra-Llerena and Adriana Kovashka. Cross-Modality Personalization for Retrieval. In *CVPR*, 2019. 3

[10] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 1

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2

[12] Patsorn Sangkloy, Wittawat Jitkrittum, Diyi Yang, and James Hays. A sketch is worth a thousand words: Image retrieval with text and sketch. In *ECCV*, 2022. 2

[13] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 2

[14] Jifei Song, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. Fine-grained image retrieval: the text/sketch input dilemma. In *BMVC*, 2017. 3

[15] Aditay Tripathi, Rajath R. Dani, Anand Mishra, and Anirban Chakraborty. Sketch-guided object localization in natural images. In *ECCV*, 2020. 3

[16] Andreas Veit, Maximilian Nickel, Serge Belongie, and Laurens van der Maaten. Separating Self-Expression and Visual Content in Hashtag Supervision. In *CVPR*, 2018. 3

[17] Liwei Wang, Alexander G. Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *NeurIPS*, 2017. 1

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