Sketch Your Own GAN Supplemental Material

Sheng-Yu Wang[1] David Bau[2] Jun-Yan Zhu[1]
[1]Carnegie Mellon University [2]MIT CSAIL

A. Implementation Details.

Training details We use the same training hyperparameters as [5]. In particular, we are using softplus for GAN loss, and R1 regularization [7] on both the sketch and image discriminator, D_Y and D_X. We do not use path length regularization, as it has no effect on the latent mapping network. Also, we set the batch size to 4 for all of our experiments, except when the sketch inputs are less than four, where we set the batch size to 1.

Hyperparameters. We use the same hyperparameters for our **full** method in all of our experiments. In particular, we use $\lambda_{\text{image}} = 0.7$

In the main text (Sec 4.1), we compared several variants of our method in our ablation studies. To make the comparison fair, for each variant, we tuned the loss weights for optimal performance. In Table 2, we list the hyperparameters used for each variant. The only exception is that we use $\lambda_{\text{weight}} = 50$ for the $\mathcal{L}_{\textbf{sketch}}$ **+** $\mathcal{L}_{\textbf{weight}}$ and $\mathcal{L}_{\textbf{sketch}}$ **+** $\mathcal{L}_{\textbf{weight}}$ **+ aug.** variant model trained on the **standing cat** task. Also, if the variants are not listed in the table, the same loss weights as the full method are used. The search space of the λ_{image} is $[0.3, 0.5, 0.7, 1.0]$, and the search space of λ_{weight} is $[0.1, 1, 10, 50, 100, 1000]$.

Data collection. In the main text (Sec. 4.1), we selected sets of 30 sketches with similar shapes and poses to designate as the user input: examples of sketches from these sets are shown in Figure 1. To evaluate generation quality, we collected images that match the input sketches from LSUN [10]. To retrieve matching images, we experimented with two sketch-image cross-domain matching methods. We applied both the SBIR method of Bui *et al*. [2] and chamfer distance [1]. Both of these retrieval results are shown in Figure 2. We observe that with chamfer distance, the retrieved images match poses of the sketches more faithfully. As a result, we adopt this method to generate our evaluation sets. However, we notice that there still exists outliers after the retrieval; hence, we hand-selected 2,500 images out of top 10,000 matches to curate the evaluation sets. A comparison between curated dataset and top chamfer matches are shown in Figure 3.

Evaluation procedure. To evaluate each model, we sample 2,500 images without truncation and save them into png files. Likewise, the evaluation set described in the main text (Sec. 4.1) consists of 2,500 256×256 images stored in png. We evaluate the Fréchet Inception Distance values using the CleanFID code [8].

B. Additional results

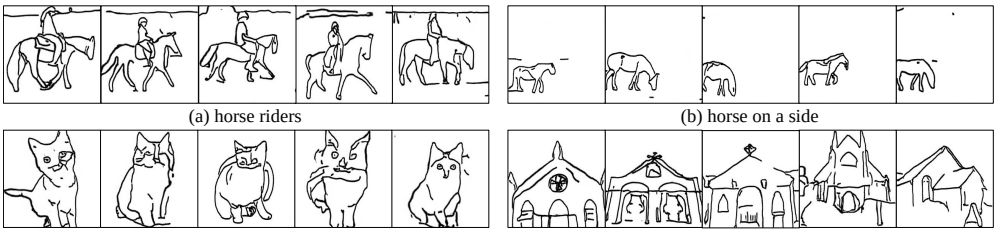
Other evaluation metrics. We report Perceptual Path Length (PPL) [4] in Table 1. We find that our method improves the original models' PPL, and beats the baselines. We note that our model focuses on fewer modes than the original one, so interpolations are smoother on average, leading to smaller PPL.

In addition, Precision, and Recall metrics [6] are reported in Table 1. The precision measures the proportion of generated samples that are close to the real dataset in VGG feature space [9], and the recall measures the proportion of real dataset that are close to generated samples in VGG feature space. We note that models with better results often have higher precision and lower recall. We expect our method to increase precision as it refines the generated distribution to better match the target distribution. But since our task aims at finding a subset of the source distribution, our method theoretically *cannot* increase the recall: increasing recall would require synthesizing new modes of real data without access to any new real examples. In our setting, the ideal maximizes precision while maintaining recall unchanged from the pretrained model. A drop in recall reveals some loss in diversity, and measures headroom for improving upon our method.

Additional qualitative results. In Figure 4, we show additional results on latent space editing with our customized models. In Figure 5, we show interpolation between customized model by interpolating the W-latents or the model weights. Also, we show uncurated samples for our models in Figure 6 (horse rider), Figure 7 (horse on a side), Figure 8 (standing cat) and Figure 9 (gabled church).

Family	Name	Training settings		Test cases											
				Horse rider			Horse on a side			Standing cat			Gabled church		
		No. Samples	Aug.	PPL ↓	Prec. ↑	Rec. ↑	PPL ↓	Prec. ↑	Rec. ↑	PPL ↓	Prec. ↑	Rec. ↑	PPL ↓	Prec. ↑	Rec. ↑
Pre-trained	Original	N/A		338.87	0.22	0.63	338.87	0.33	0.57	438.11	0.21	0.54	342.73	0.46	0.49
Baseline	Bui *et al*. [2]	30		356.56	0.24	0.53	343.48	0.26	**0.60**	433.05	0.22	**0.58**	346.48	0.49	0.48
	Chamfer	30		353.07	0.30	**0.56**	371.11	0.35	0.57	418.91	0.26	0.55	340.12	**0.50**	**0.52**
Ours	Full (w/o aug.)	30		353.71	0.42	0.52	266.69	0.42	0.49	**150.89**	**0.65**	0.20	344.24	0.48	0.48
	Full (w/ aug.)	30	✓	**306.81**	**0.50**	0.50	**232.95**	**0.44**	0.39	263.99	0.50	0.41	**336.67**	0.46	0.51

Table 1. **Other metrics.** We report the Perceptual Path Length (PPL), Precision (Prec.), and Recall (Rec.) of the original models, baselines and our methods on four different test cases. The details of the baselines are in the main text (Sec. 4.1). ✓ indicates translation augmentation is applied. ↑, ↓ indicate if higher or lower is better. Evaluations on the original models are in gray, and the best value is highlighted in **black**.



(a) horse riders (b) horse on a side (c) standing cat (d) gabled church

Figure 1. **Example of sketches used for training.** For each task (a, b, c, d), 30 sketches with similar shapes and layouts are hand-selected as training samples, where above shows subsets of 5 sketches.

Figure 2. **Comparison between retrieval methods.** We compare retrieval methods between chamfer distance [1] and SBIR method of Bui *et al*. [2]. We find that the retrievals using chamfer distance matches the input sketches better than those using Bui *et al*. Left shows the example query out of the 30 sketches used for the retrieval.

Figure 3. **Curated evaluation set.** We show random samples from the top 2,500 matches using chamfer distance [1] **(left)** and 2,500 hand-selected images **(right)**. The quality of the evaluation set is improved after curation.

	λ_{image}	λ_{weight}
$\mathcal{L}_{\text{sketch}}$	0	0
$\mathcal{L}_{\text{sketch}}$+aug.	0	0
$\mathcal{L}_{\text{sketch}}$+$\mathcal{L}_{\text{weight}}$	0	100
$\mathcal{L}_{\text{sketch}}$+$\mathcal{L}_{\text{weight}}$+aug.	0	100

Table 2. **Loss weights for each variant.** For a fair comparison, we use different loss weights for several variants. We find that using the above weights gives optimal performance. The variants not listed in this table is using the same hyperparameters as the **Full** method.

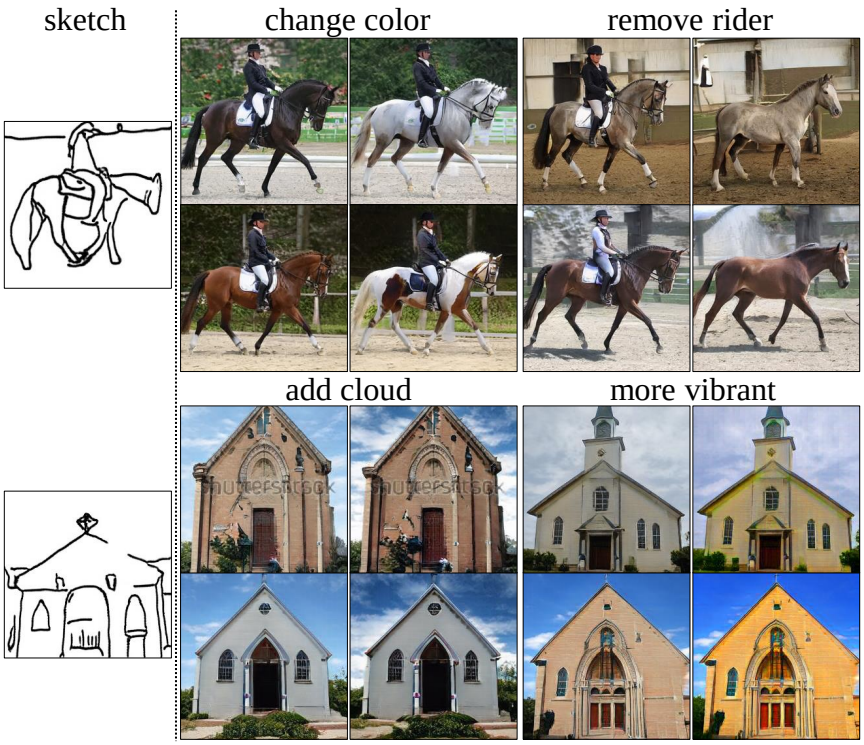


Figure 4. **Additional latent edit results.** Similar to Figure 9 in the main text, we show additional results of applying GANSpace [3] edits to our customized models, **horse rider** (top) and **gabled church** (bottom).

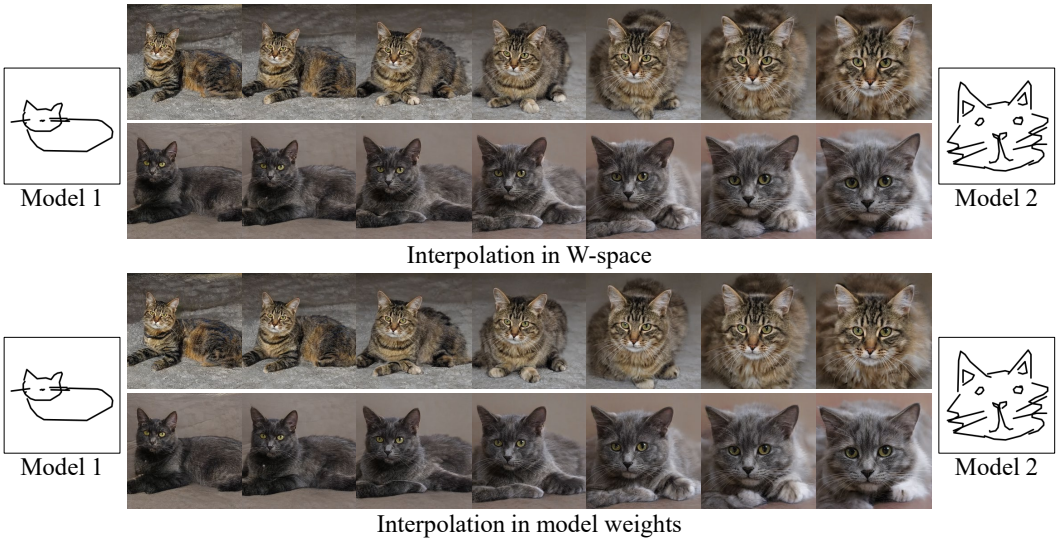


Figure 5. **Interpolating between customized models.** We can interpolate between the customized model by interpolating (top) the W-latents or (bottom) the model weights. Model 1 and 2 are from Figure 6 and Figure 5 in the main text, respectively.

Figure 6. Uncurated samples of the **horse rider** model. Truncation $\psi = 0.5$ is applied to generate the images.

Figure 7. Uncurated samples of the **horse on a side** model. Truncation $\psi = 0.5$ is applied to generate the images.

Figure 8. Uncurated samples of the **standing cat** model. Truncation $\psi = 0.5$ is applied to generate the images.

Figure 9. Uncurated samples of the **gabled church** model. Truncation $\psi = 0.5$ is applied to generate the images.

References

[1] Harry G Barrow, Jay M Tenenbaum, Robert C Bolles, and Helen C Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. Technical report, SRI INTERNATIONAL MENLO PARK CA ARTIFICIAL INTELLIGENCE CENTER, 1977. 1, 2

[2] Tu Bui, Leonardo Ribeiro, Moacir Ponti, and John Collomosse. Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network. *Computer Vision and Image Understanding*, 2017. 1, 2

[3] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Advances in Neural Information Processing Systems*, 2020. 3

[4] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[5] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[6] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1

[7] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International Conference on Machine Learning (ICML)*, 2018. 1

[8] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On buggy resizing libraries and surprising subtleties in fid calculation. *arXiv preprint arXiv:2104.11222*, 2021. 1

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 1

[10] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 1