Supplementary Material for StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets

Axel Sauer
University of Tübingen
Germany
a.sauer@uni-tuebingen.de

Katja Schwarz
University of Tübingen
Germany
katja.schwarz@uni-tuebingen.de

Andreas Geiger
University of Tübingen
Germany
a.geiger@uni-tuebingen.de

ACM Reference Format:
Axel Sauer, Katja Schwarz, and Andreas Geiger, . 2022. Supplementary Material for StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings (SIGGRAPH '22 Conference Proceedings), August 7–11, 2022, Vancouver, BC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3528233.3530738

In this supplemental document, we elaborate on the current limitations, describe the procedure to increase the resolution of ImageNet to one megapixel, and specify the implementation details of our approach. The supplemental video shows additional samples and interpolations. We use the same mathematical notation as in the paper.

A LIMITATIONS

Architectural Limitations. First, StyleGAN-XL is three times larger than StyleGAN3, constituting a higher computational overhead when used as a starting point for finetuning. Therefore, it will be worth exploring GAN distillation methods [Chang and Lu 2020] that trade-off performance for model size. Second, StyleGAN-XL uses translation-equivariant layers of StyleGAN3-T. As described above, StyleGAN3-R tends to produce overly symmetrical images and adds significant computational overhead. Finding a more efficient rotational-equivariant architecture is an important future direction. Second, we find StyleGAN3, and consequently, StyleGAN-XL, harder to edit, e.g., high-quality edits via $\mathcal{W}$ are noticeably easier to achieve with StyleGAN2. As already observed in [Karras et al. 2021], StyleGAN3's semantic controllability is reduced for the sake of equivariance. However, techniques using the *StyleSpace* [Wu et al. 2021], e.g., StyleMC [Kocasari et al. 2022], tend to yield better results in our experiments, confirming the findings of concurrent work by [Alaluf et al. 2022]. Furthermore, we remark that our framework can also easily be used with StyleGAN2 layers.

Comparison to Diffusion Models. Our model is larger than earlier StyleGANs, yet it is still several orders of magnitudes faster than ADM; we compare inference speeds in the supplementary. Low data coverage is a known problem of GANs, and StyleGAN-XL makes notable headway on this issue. However, StyleGAN-XL is still outperformed by diffusion models regarding data coverage. Furthermore, classes of unaligned humans and human faces are particularly challenging for all compared approaches, likely due to ImageNet's emphasis on non-human objects [Dhariwal and Nichol 2021]. For such classes, we observe that ADM [Dhariwal and Nichol 2021] generates more convincing human faces than BigGAN [Brock et al. 2019] or StyleGAN-XL. Both GANs can synthesize realistic faces; however, the main challenge in this setting is that the dataset is unstructured, and the humans are not aligned. [Brock et al. 2019] remarked the particular challenge of classes containing details to which human observers are more sensitive. We show examples in Fig. 1. Whether the points above are a general limitation of GANs remains an interesting open question for future research.

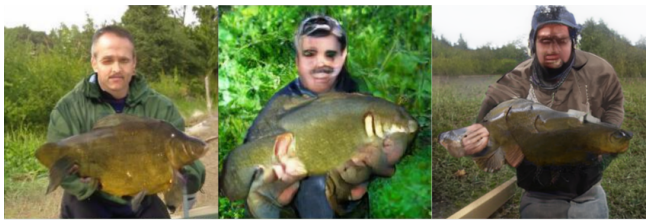

Figure 1: *Imagenet Classes Containing Humans.* *Samples for BigGAN and ADM are taken from [Dhariwal and Nichol 2021].*

B PREPROCESSING IMAGENET

An initial challenge is the lack of high-resolution data; the mean resolution of ImageNet is 469×387. Similar to the procedure used for generating CelebA-HQ[Karras et al. 2018], we preprocess the whole dataset with SwinIR-Large [Liang et al. 2021], a recent model for real-world image super-resolution. Of course, a trivial way of achieving good performance on this dataset would be to draw samples from a 256^2 generative model and passing it through SwinIR. However, SwinIR adds significant computational overhead as it is 60 times slower than our upsampling stack. Furthermore, this way, StyleGAN-XL's weights can be used for initialization when finetuning on other high-resolution datasets. Lastly, combining StyleGAN-XL and SwinIR would impair translation equivariance.

C INFERENCE SPEED

GANs generate samples in a single forward pass, unlike diffusion models that must be applied several hundred or thousand times to generate a sample. Table 1 compares StyleGAN-XL to ADM. We find that StyleGAN-XL is several orders of magnitude faster. In defense of diffusion models, speeding up their sampling is an active area of research, and novel techniques [Watson et al. 2021] may be able to reduce this gap in the future.

SIGGRAPH '22 Conference Proceedings, August 7–11, 2022, Vancouver, BC, Canada

ACM ISBN 978-1-4503-9337-9/22/08…$15.00
https://doi.org/10.1145/3528233.3530738

Table 1: *Inference speed comparison.*. *We measure the time required for a forward pass with batch size* 1 *in V100-seconds. ADM uses classifier guidance.*

Model	Inference Time ↓		
	Res. 128^2	**Res.** 256^2	**Res.** 512^2
ADM	27.07	40.26	91.54
StyleGAN-XL	0.05	0.07	0.10

Table 2: *Results on Unimodal Datasets.*.

Model	FID	Model	FID
FFHQ 1024^2		**Pokémon** 1024^2	
StyleGAN2	2.70	FastGAN	56.46
StyleGAN3	2.79	Projected GAN	33.96
StyleGAN-XL	**2.02**	StyleGAN-XL	**25.47**

D RESULTS ON UNIMODAL DATASETS

StyleGAN-XL is designed to enable training on large and diverse datasets. However, applying it to big and small unimodal datasets is straightforward. In contrast to the configuration for ImageNet, we begin with ten layers at the lowest stage and add two layers per resolution stage. Furthermore, we do not employ classifier guidance. Table 2 reports the results for both datasets at resolution 1024^2, StyleGAN-XL achieves state-of-the-art performance on both.

E ADDITIONAL QUALITATIVE RESULTS

In the following, we present additional qualitative results. Fig. 2 shows additional interpolations between samples from different classes. Fig. 4 and Fig. 5 show samples on FFHQ 1024^2 and Pokemon 1024^2 respectively. Lastly, we compare BigGAN, ADM, and StyleGAN-XL on different ImageNet classes. For a fair comparison, we do not use truncation or classifier guidance. Instead, we show images with the largest logits given by a VGG16 which corresponds to individual image quality.

F IMPLEMENTATION DETAILS

Inversion. Following [Karras et al. 2020], we use basic latent optimization in $\mathcal{W}$ for inversion. Given a target image, we first compute its average style code $\bar{\mathbf{w}}$ by running 10000 random latent codes $\mathbf{z}$ and target specific class samples $\mathbf{c}$ through the mapping network. As the class label of the target image is unknown, we pass it to a pretrained classifier. We then use the classifier logits as a multinomial distribution to sample $\mathbf{c}$. In our experiments, we use Deit-base [Touvron et al. 2021] as a classifier, but other choices are possible. At the beginning of optimization , we initialize $\mathbf{w} = \bar{\mathbf{w}}$. The components of $\mathbf{w}$ are the only trainable parameters. The optimization runs for 1000 iterations using the Adam optimizer [Kingma and Ba 2015] with default parameters. We optimize the LPIPS [Zhang et al. 2018] distance between the target image and the generated image. For StyleGAN-XL, the maximum learning rate is $\lambda_{max} = 0.05$. It is ramped up from zero linearly during the first 50 iterations and ramped down to zero using a cosine schedule during the last 250 iterations. For BigGAN, we empirically found $\lambda_{max} = 0.001$ and a ramp-down over the last 750 iterations to yield the best results. All inversion experiments are performed at resolution 512^2 and computed on $5k$ images (10% of the validation set). We report the results in Table 3 and show qualitative results in Fig. 3.

Table 3: *Inversion Results.* *The metrics are computed between the inversions obtained by the model and the reconstruction targets.*

Model	MSE ↓	PSNR ↑	SSIM ↑	FID ↓
BigGAN	0.10	10.85	0.26	47.48
StyleGAN-XL	**0.06**	**13.45**	**0.33**	**21.73**

Table 4: *Results on ImageNet at Lower Resolutions.*.

Model	FID ↓		
	Res. 16^2	**Res.** 32^2	**Res.** 64^2
StyleGAN-XL	0.73	1.10	1.51

Training StyleGAN3 on ImageNet. For training StyleGAN3, we use the official PyTorch implementation[1]. The results in Fig. 1 are computed with the StyleGAN3-R configuration on resolution 256^2 until the discriminator has seen 10 million images. We find that StyleGAN3-R and StyleGAN3-T converge to similar FID without any changes to their training paradigm. The run with the best FID score was selected from three runs with different random seeds. We use a channel base of 16384 and train on 8 GPUs with total batch size 256, $\gamma = 0.256$. The remaining settings are chosen according to the default configuration of the code release. For the ablation study in Table 1 , we use the StyleGAN3-T configuration as baseline since StyleGAN-XL builds upon the translational-equivariant layers of StyleGAN3. We train on 4 GPUs with total batch size 256 and batch size 32 per GPU, $\gamma = 0.25$, and disable augmentation.

Training & Evaluation. For all our training runs, we do not use data amplification via *x-flips* following [Karras et al. 2020]. Furthermore, we evaluate all metrics using the official StyleGAN3 codebase. For the baseline values in Table 2, we report the numbers of [Dhariwal and Nichol 2021]. The official codebase of ADM[2] provides files containing 50k samples for ADM and BigGAN. We utilize the provided samples to compute rFID. Following [Dhariwal and Nichol 2021], we compute precision and recall between 10k real samples and 50k generated samples. Table 4 reports the results on ImageNet at lower resolutions.

Layer configurations. We start progressive growing at resolution 16^2 using 11 layers. The layer specifications are computed according to [Karras et al. 2021] and remain fixed for the remaining training. For the next stage, at resolution 32^2, we discard the last 2 layers and add 7 new ones. The specifications for the new layers are computed according to [Karras et al. 2021] for a model with resolution 32^2 and

[1] https://github.com/NVlabs/stylegan3.git
[2] https://github.com/openai/guided-diffusion

16 layers. Continuing this strategy up to resolution 1024^2 yields the flexible layer specification of StyleGAN-XL in Fig. 9.

REFERENCES

Yuval Alaluf, Or Patashnik, Zongze Wu, Asif Zamir, Eli Shechtman, Dani Lischinski, and Daniel Cohen-Or. 2022. Third Time's the Charm? Image and Video Editing with StyleGAN3. *arXiv.org* abs/2201.13433 (2022). https://arxiv.org/abs/2201.13433

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *Proc. of the International Conf. on Learning Representations (ICLR)*. OpenReview.net. https://openreview.net/forum?id=B1xsqj09Fm

Ting-Yun Chang and Chi-Jen Lu. 2020. TinyGAN: Distilling BigGAN for Conditional Image Generation. In *Proc. of the Asian Conf. on Computer Vision (ACCV) (Lecture Notes in Computer Science, Vol. 12625)*, Hiroshi Ishikawa, Cheng-Lin Liu, Tomás Pajdla, and Jianbo Shi (Eds.). 509–525. https://doi.org/10.1007/978-3-030-69538-5_31

Prafulla Dhariwal and Alex Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. (2021).

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *Proc. of the International Conf. on Learning Representations (ICLR)*. OpenReview.net. https://openreview.net/forum?id=Hk99zCeAb

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-Free Generative Adversarial Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE, 8107–8116. https://doi.org/10.1109/CVPR42600.2020.00813

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proc. of the International Conf. on Learning Representations (ICLR)*.

Umut Kocasari, Alara Dirik, Mert Tiftikci, and Pinar Yanardag. 2022. StyleMC: Multi-Channel Based Fast Text-Guided Image Generation and Manipulation. *Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV)* (2022). https://arxiv.org/abs/2112.08493

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. 2021. SwinIR: Image Restoration Using Swin Transformer. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV) Workshops*. IEEE, 1833–1844. https://doi.org/10.1109/ICCVW54120.2021.00210

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *Proc. of the International Conf. on Machine learning (ICML) (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 10347–10357. http://proceedings.mlr.press/v139/touvron21a.html

Daniel Watson, Jonathan Ho, Mohammad Norouzi, and William Chan. 2021. Learning to Efficiently Sample from Diffusion Probabilistic Models. *CoRR* abs/2106.03802 (2021). https://arxiv.org/abs/2106.03802

Zongze Wu, Dani Lischinski, and Eli Shechtman. 2021. StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE, 12863–12872. https://openaccess.thecvf.com/content/CVPR2021/html/Wu_StyleSpace_Analysis_Disentangled_Controls_for_StyleGAN_Image_Generation_CVPR_2021_paper.html

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE Computer Society, 586–595. https://doi.org/10.1109/CVPR.2018.00068

Figure 2: *Interpolations.* *StyleGAN-XL generates smooth interpolations between samples of different classes.*

Figure 3: *Inversion of a Given Source Image.* *For BigGAN, we invert to its latent space* $\mathbf{z}$*, for StyleGAN-XL we invert to style codes* $\mathbf{w}$*.*

Figure 4: ***Samples on FFHQ*** 1024^2.

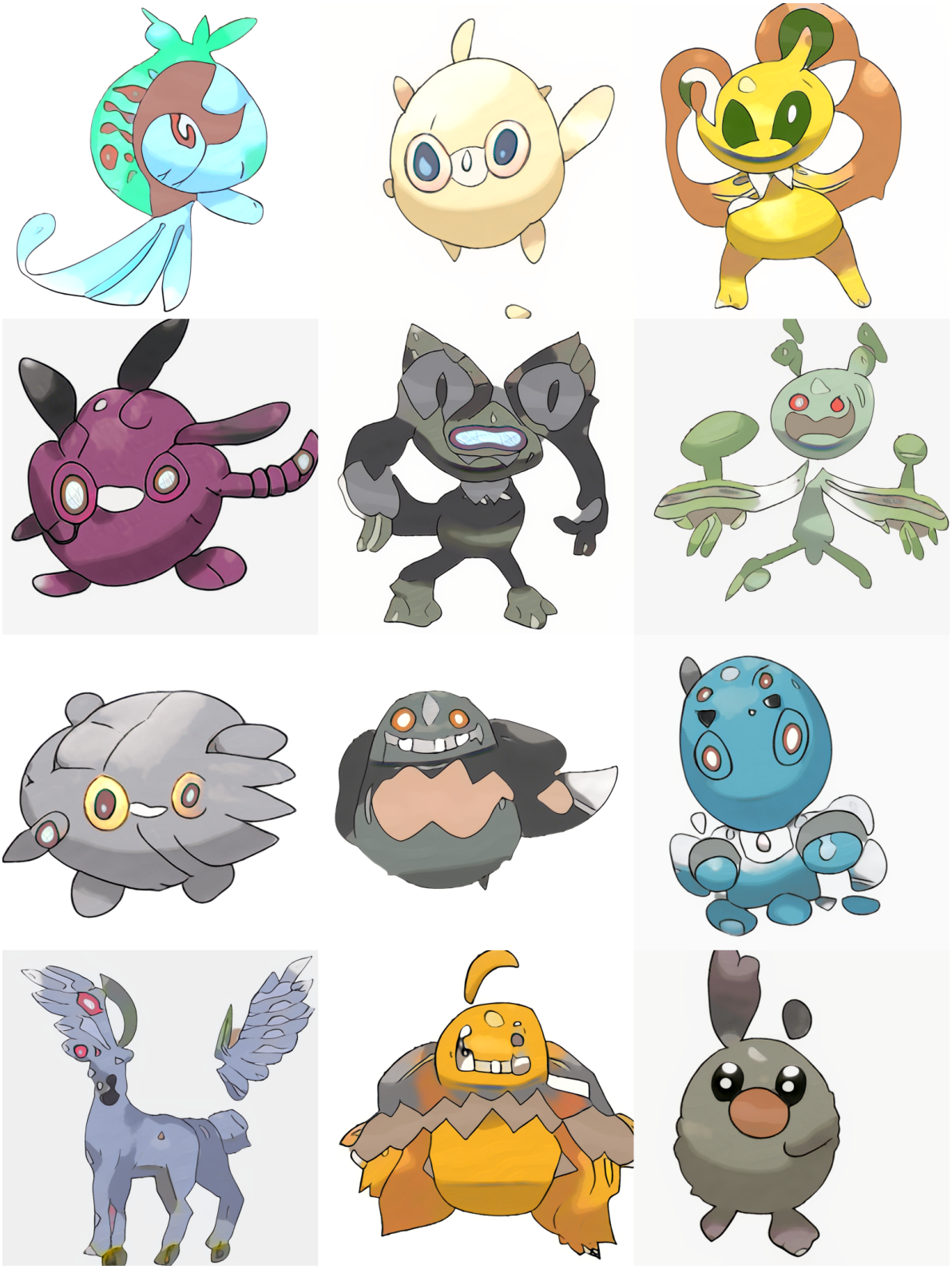

Figure 5: ***Samples on Pokemon*** 1024^2.

Figure 6: *Qualitiative Comparison on ImageNet* 256^2**.** *We compare BigGAN (left column), ADM (middle column), and StyleGAN-XL (right column). Classes from top to bottom: pizza, valley, daisy, dough, comic book.*

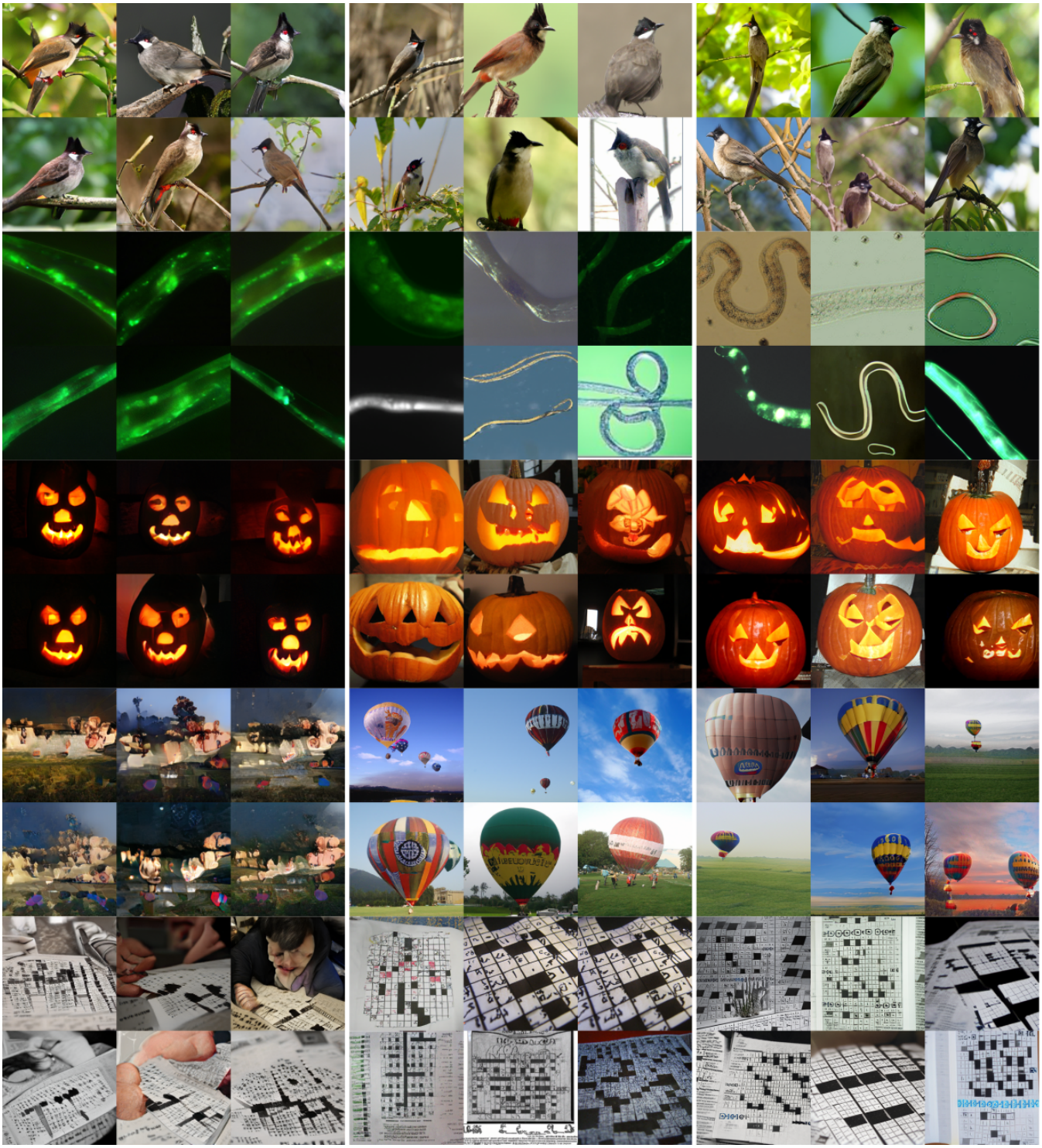

Figure 7: *Qualitiative Comparison on ImageNet* 256^2**.**. *We compare BigGAN (left column), ADM (middle column), and StyleGAN-XL (right column). Classes from top to bottom: bulbul, nematode, jack-o'-lantern, balloon, crossword puzzle.*

Figure 8: *Qualitiative Comparison on ImageNet* 256^2**.** *We compare BigGAN (left column), ADM (middle column), and StyleGAN-XL (right column). Classes from top to bottom: agaric, orange, Tibetian mastiff, espresso, paddlewheel.*

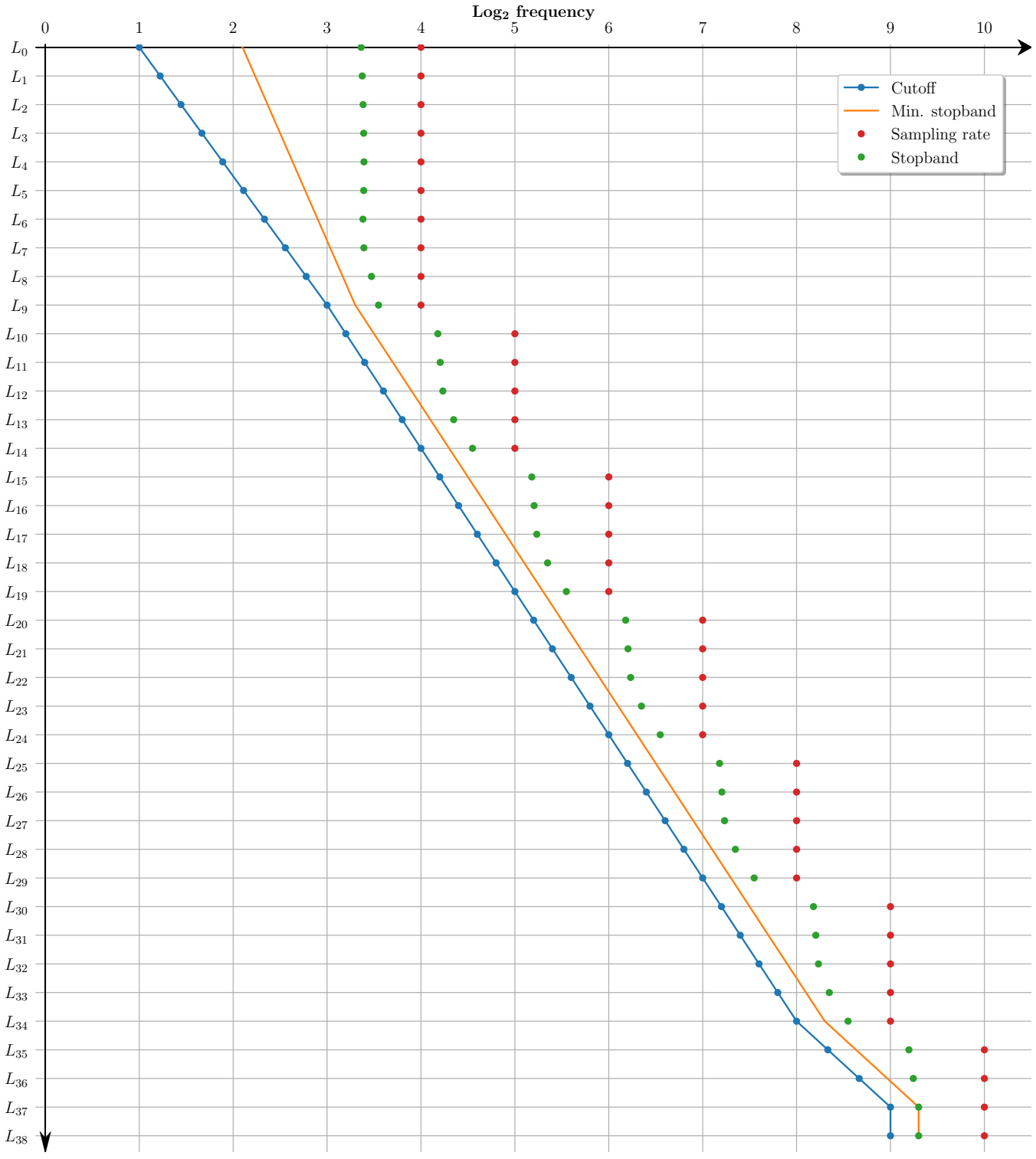

Figure 9: *Flexible Layer Specification of Stylegan-XL.* *StyleGAN-XL consists of* 39 *layers at resolution* 1024^2. *Cutoff (blue) and minimum acceptable stopband frequency (orange) obey geometric progression over the layers; sampling rate (red) and actual stopband (green) are computed according to our design constraints.*