Supplementary Material
- Local Attention Pyramid for Scene Image Generation

Sang-Heon Shim, Sangeek Hyun, DaeHyun Bae, Jae-Pil Heo*
Sungkyunkwan University

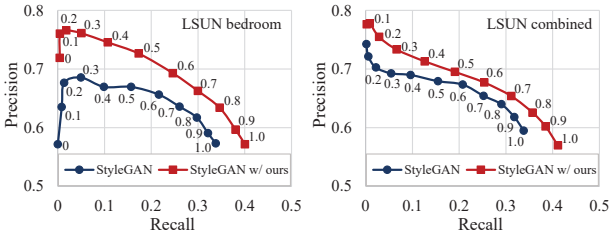

Figure 1. Precision-recall curves on LSUN bedroom (left) and combined (right). Number at each point is truncation threshold ψ.

A. Precision and recall (PR) metrics

We measure PR [5] of StyleGAN [4] and ours. We observe that higher generation quality reported in Table. 1 of the paper is coming from higher recalls (i.e. diversity). This makes sense because the diversity can be derived from dedicate generation of fine-details based on distributed activation by LAP. Since the PR generally have trade-off, we further compute PR curves with truncation parameter ψ. Specifically, a lower ψ leading conservative sampling results in a higher precision and a lower recall [2, 5]. As reported in Fig. 1, our method significantly outperforms the baseline. For instance, at the recall of 0.3, our method ($\psi = 0.7$) provides 7.3% higher precision than the baseline ($\psi = 0.8$) on LSUN bedroom. Note that, the quantities are measured with 50K real and fake samples.

B. Per class evaluation w.r.t. freq. and size

Class-wise performances are measured by relative improvement ratios, $(\frac{\text{FID of StyleGAN}}{\text{FID of Ours}})$ and $(\frac{\text{Score of Ours}}{\text{Score of StyleGAN}})$, for FID [3] and PR [5], respectively. We first plot about FID (based on Table. 3 in the paper) to Fig. 2 (top), and it shows a tendency that improvements are higher on small and less frequent objects. For class-wise PR, we evaluate precision at the equal (actually very close) recall, and vise versa, for a clear comparison. Specifically, we evaluate precision at recall of 0.3 ($\psi_{\text{Ours}} = 0.7$ and $\psi_{\text{StyleGAN}} = 0.8$), and recall at precision of 0.57 ($\psi = 1$ for both) on LSUN bedroom, as shown in Fig. 2 (bottom). The precision (e.g. visual quality) has a similar trend with FID, where higher improvements are achieved by small or less frequent objects. However, although most recalls are higher than baseline across various objects, it is hard to tell the same claim for the recall. We conjecture that high recalls for the large objects are mainly because they have many fine-detail parts where LAP helps diverse generation. In other words, the detailed parts of large objects (e.g. bed rod) are treated similarly with small object (e.g. table) by LAP, since GANs are trained in totally unsupervised way.

*Corresponding author

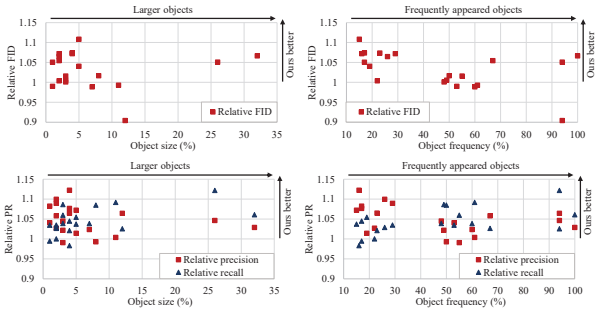

Figure 2. Relative improvements of class-wise FID (upper row) & precision and recall (lower row) over StyleGAN w.r.t. object size and frequency on LSUN bedroom.

Dataset	Resolution	Method	FID	FSD
LSUN church	256^2	StyleGAN	4.34	19.2
		StyleGAN (w/ ours)	**3.69**	**14.4**

Table 1. Experimental results on LSUN church.

C. Results onto LSUN church

We additionally validated on LSUN church, as reported in Table. 1. Our LAP achieves an FID score of 3.69, improving 0.65 points over a StyleGAN, as well as increases the FSD score [1] from 19.2 to 14.4. These results confirm again that our LAP works well on various indoor and outdoor scene. Note that, we trained all models until the discriminator sees 25M of real images.

(a) LSUN bedroom (b) LSUN combined (c) COCO-stuff

Figure 3. Samples generated by StyleGAN (top row) and by StyleGAN w/ ours (bottom row). Best viewed in zoom.

D. More qualitative results

Fig. 3 shows that StyleGAN has difficulty in rendering object-parts. For example, the seat of ottoman (1st col.), legs of chairs (4th col.) and animal (8th col.) were skipped or vanished out. On the other hands, the LAP generated the rectangular shaped ottoman (1th col.), chairs w/ straight legs (4th col.) or animal whose legs were not faded out (8th col.).

References

[1] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4502–4511, 2019.

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

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

[5] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019.