Dual Contrastive Loss and Attention for GANs (Supplementary Material)

Ning Yu[1,2] Guilin Liu[3] Aysegul Dundar[3,4]
Andrew Tao[3] Bryan Catanzaro[3] Larry Davis[1] Mario Fritz[5]
[1]University of Maryland [2]Max Planck Institute for Informatics [3]NVIDIA
[4]Bilkent University [5]CISPA Helmholtz Center for Information Security
{ningyu,lsdavis}@umd.edu
{guilinl,adundar,atao,bcatanzaro}@nvidia.com
fritz@cispa.saarland

1. Different GAN backbones for dual contrastive loss

In Table 1 we show the consistent and significant advantages of our dual contrastive loss on two other GAN backbones: SNGAN [10] and StyleGAN [6].

Method	FFHQ	Bedroom	Church	Horse	CLEVR
SNGAN	11.28	11.14	7.37	13.87	29.19
+ Contr	**8.98**	**10.79**	**6.51**	**13.59**	**18.23**
StyleGAN	6.83	5.30	5.12	7.27	12.43
+ Contr	**6.42**	**4.76**	**4.48**	**6.26**	**8.96**

Table 1. Comparisons in FID on different GAN backbones.

2. Self-attention at different generator resolutions

It is empirically acknowledged that the optimal resolution to replace convolution with self-attention in the generator is specific to dataset and image resolution [15]. For the state-of-the-art attention module SAN [16] in Table 3 in the main paper, we find that it achieves the optimal performance at 32×32 generator resolution consistently over all the limited-scale 128×128 datasets, and therefore we report these FIDs.

For the large-scale datasets with varying resolutions in Table 6 in the main paper, we conduct an analysis study on their optimal resolutions as shown in Table 2.

We find there is a specific optimal resolution for each dataset, and the FID turns monotonically deteriorated when introducing self-attention one resolution up or down. We reason that each dataset has its own spatial scale and complexity. If longer-range dependency or consistency counts more than local details in one dataset, e.g., CLEVR, it is more favorable to use self-attention in an earlier layer, thus at a lower resolution. We stick to the optimal resolution and report the corresponding FID for each dataset in Table 6 in the main paper.

Resolution	FFHQ	Bedroom	Church	Horse	CLEVR
8^2	6.08	4.43	5.10	4.24	10.44
16^2	5.81	4.21	5.24	**3.58**	**8.96**
32^2	5.69	**3.48**	**4.38**	3.75	9.04
64^2	**5.13**	3.69	4.57	3.94	12.48
128^2	5.75	6.69	4.82	6.82	18.40

Table 2. FID w.r.t. the resolution at which we replace convolution with SAN [16] in the generator.

3. Different reference-attention configurations

Eq. 8 in the main paper provides the flexibility of how to cooperate between reference and primary images. We empirically explore the other configurations of sources to the key, query, and value components in the reference-attention. The following two equations, Eq. 1 and Eq. 2, correspond to the two configuration variants we compare to.

Configuration	CelebA	Animal Face	Bedroom	Church
Eq. 1	10.39	65.16	20.22	17.85
Eq. 2	10.95	32.33	11.05	8.33
Eq. 8 in main	**7.48**	**31.08**	**8.32**	**7.86**

Table 3. FID w.r.t. different reference-attention configurations in the discriminator. For computationally efficient comparisons, we use the 30k subset of each dataset at 128×128 resolution.

$$\mathbf{O}^{ref} \doteq \mathsf{attn}\left(\mathbf{K}(\mathbf{T}_{pri}), \mathbf{Q}(\mathbf{T}_{ref}), \mathbf{V}(\mathbf{T}_{ref})\right) + \mathbf{T}_{pri} \quad (1)$$

$$\mathbf{O}^{ref} \doteq \mathsf{attn}\left(\mathbf{K}(\mathbf{T}_{pri}), \mathbf{Q}(\mathbf{T}_{ref}), \mathbf{V}(\mathbf{T}_{pri})\right) + \mathbf{T}_{pri} \quad (2)$$

Data size	CelebA StyleGAN2	CelebA + ref attn	Animal Face StyleGAN2	Animal Face + ref attn	Bedroom StyleGAN2	Bedroom + ref attn	Church StyleGAN2	Church + ref attn
1K	55.71	**43.19**	181.26	**123.08**	230.40	**79.81**	107.31	**43.05**
5K	23.48	**18.48**	89.88	**61.17**	57.68	**19.64**	29.30	**17.85**
10K	14.73	**12.72**	61.36	**45.49**	40.70	**12.29**	17.94	**12.13**
30K	9.84	**7.48**	36.55	**31.08**	19.33	**8.32**	11.02	**7.86**
50K	**6.59**	7.09	28.92	**28.43**	14.01	**7.15**	8.88	**7.09**
100K	**5.61**	6.86	**22.85**	28.37	9.42	**6.89**	7.32	**7.08**

Table 4. Comparisons in FID between StyleGAN2 config E baseline and that with our reference-attention in the discriminator. Our method consistently improves the baseline when dataset size varies between 1k and 30k images. For computationally efficient comparisons, we use each dataset at 128×128 resolution. See Fig. 6 in the main paper for the corresponding plots.

Method	Loss	FFHQ FID	FFHQ PPL	FFHQ P	FFHQ R	FFHQ Sep	Bedroom FID	Bedroom PPL	Bedroom P	Bedroom R	Church FID	Church PPL	Church P	Church R	Horse FID	Horse PPL	Horse P	Horse R	CLEVR FID	CLEVR PPL	CLEVR P	CLEVR R
BigGAN [1]	Adv	11.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
U-Net GAN [12]	Adv	7.48	**32**	0.68	0.19	**2.00**	17.6	**504**	0.48	0.03	11.7	**318**	**0.62**	0.07	20.2	**296**	0.57	0.13	33.3	202	0.04	0.08
StyleGAN2 [7]	Adv	4.86	47	0.69	0.42	5.08	4.01	976	0.59	0.32	4.54	511	0.57	0.42	3.91	637	0.63	0.40	9.62	582	0.46	0.56
StyleGAN2 w/ attn	Adv	5.13	54	0.69	0.41	4.18	3.48	1384	0.59	0.36	4.38	611	0.59	0.41	3.59	636	**0.64**	0.39	8.96	**67**	0.47	0.63
StyleGAN2	Contr	**3.98**	50	**0.71**	**0.44**	3.76	3.86	1054	**0.60**	0.31	3.73	619	0.60	0.40	3.70	740	**0.64**	0.39	6.06	816	0.57	0.65
StyleGAN2 w/ attn	Contr	4.63	65	0.70	0.41	3.60	**3.31**	1830	0.59	**0.37**	**3.39**	1239	0.60	**0.45**	**2.97**	1367	**0.64**	**0.43**	**5.05**	106	**0.58**	**0.70**

Table 5. Comparisons to the state-of-the-art GANs in various metrics on the large-scale datasets. We highlight the best in **bold** and second best with underline. "w/ attn" indicates using the self-attention in the generator. "Contr" indicates using our dual contrastive loss instead of conventional GAN loss.

Resolution	CelebA	Animal Face	Bedroom	Church
8^2	**7.48**	**31.08**	**8.32**	**7.86**
16^2	31.36	118.82	11.05	11.42
32^2	55.07	195.82	146.85	61.83

Table 6. FID w.r.t. the resolution at which we replace convolution with reference-attention in the discriminator. For computationally efficient comparisons, we use the 30k subset of each dataset at 128×128 resolution.

From Table 3, we validate that Eq. 8 in the main paper is the best setting. We reason that the value embedding is relatively independent of the key and query embeddings. Hence we should encode value from one source, and key and query from the other source. Also, because the value and residual shortcut contribute more directly to the discriminator output, we should feed them with the primary image, and feed the key and query with the reference image to formulate the spatially adaptive kernel.

4. Reference-attention at different discriminator resolutions

In Table 6, we analyze the relationship between generation quality and the resolution to replace convolution with reference-attention in the discriminator. We stop investigation to higher resolutions because the training turns easily diverging. We conclude introducing reference-attention at the lowest possible resolution is most beneficial. We reason that the deepest features are the most representative for cooperating between reference and primary images. Also because the primary and reference images are not pre-aligned, the lowest resolution covers the largest receptive field and therefore leads to the largest overlap between the two images that should be corresponded. We stick to the 8×8 resolution for all the experiments involving reference-attention.

5. FID w.r.t. data size for reference-attention

We report in Table 4 the detailed values for Fig. 6 in the main paper. Our method consistently improves the baseline when dataset size varies between 1k and 30k images.

6. Comparisons to the state of the art in various metrics

We extend Table 6 in the main paper with additional evaluation metrics for GANs, which are proposed and used in StyleGAN [6] and/or StyleGAN2 [7]: Perceptual Path Length (PPL), Precision (P), Recall (R), and Separability (Sep). See Table 5.

Consistent with FID rankings, our attention modules and dual contrastive loss also improve from StyleGAN2 baseline for Precision, Recall, and Separability in most cases. It is worth noting that the rankings of PPL are negatively correlated to all the other metrics, which disqualifies it as an effective evaluation metric in our experiments. E.g., U-Net GAN has the best PPL in most cases but in fact it contradicts against its worst FID and worst visual quality in Fig. 1, 2, 3,

Method	Loss	CelebA	Animal Face	Bedroom	Church
StyleGAN2 [7]	Adv	9.84	36.55	19.33	11.02
StyleGAN2 w/ self-attn-G	Adv	8.60	32.72	16.36	9.62
StyleGAN2 w/ self-attn-G	Contr	7.55	25.83	10.99	8.12
StyleGAN2 w/ self-attn-G ref-attn-D	Adv	7.48	31.08	**8.32**	**7.86**
StyleGAN2 w/ self-attn-G ref-attn-D	Contr	**6.00**	**25.03**	12.84	8.75

Table 7. Comparisons in FID to StyleGAN2 config E baseline on the limited-scale datasets. Our configurations consistently improve the baseline, the relative improvements of which are even more significant than those on the large-scale datasets. We use the 30k subset of each dataset at 128×128 resolution.

4, and 5.

7. Comparisons on the limited-scale datasets

Besides comparisons on the large-scale datasets, we also compare to StyleGAN2 [7] baseline on the limited-scale datasets in Table 7. We use the 30k subset of each dataset at 128×128 resolution. We find:

(1) Comparing across the first, second, and third rows, self-attention generator, dual contrastive loss, and their synergy significantly and consistently improve on all the limited-scale datasets, more than what they improve on the large-scale datasets: from 18.1% to 23.3% on CelebA [9] and Animal Face [8], from 17.5% to 43.2% on LSUN Bedroom [14], and from 25.2% to 26.4% on LSUN Church [14]. It indicates the limited-scale setting is more challenging and leaves more space for our improvements.

(2) Comparing between the first and fourth rows, the reference-attention discriminator improves significantly and consistently on all the datasets up to 57.0% on LSUN Bedroom. We reason that the arbitrary pair-up between reference and primary images results in a beneficial effect similar in spirit to data augmentation, and consequently generalizes the discriminator representation and mitigates its overfitting.

(3) However, according to the fifth row, reference-attention discriminator is sometimes not compatible with contrastive learning because they may together overly augment the classification task: contrastive learning for one pair of primary and reference input against a batch of other pairs makes adversarial training unstable. This observation differs from that of pairwise contrastive learning in the unsupervised learning scenario [4, 13, 2, 3] or GAN applications with reconstructive regularization [11].

Even though in this paper our main scope is GANs on large-scale datasets, we believe these findings to be very interesting for researchers to design their networks for limited-scale datasets.

8. Uncurated generated samples

For comparisons to the state of the art, we show more uncurated generated samples in Figure 1, 2, 3, 4 and 5. Our generation significantly outperforms the baselines U-Net GAN [12] and StyleGAN2 [7] in terms of quality, long-range dependencies, and spatial consistency.

9. Self-attention maps

For self-attention maps in the generator, we show more results in Figure 6, 7, 8, and 9. The attention maps strongly align to the semantic layout and structures of the generated images, which enable long-range dependencies across objects.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2018. 2

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[3] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020. 3

[4] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 3

[5] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 8, 10

[6] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2, 4

[7] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2, 3

[8] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *ICCV*, 2019. 3

[9] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 3

[10] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018. 1

[11] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *ECCV*, 2020. 3

U-Net GAN
FID = 7.48

StyleGAN2
FID = 4.86

Ours
FID = 3.98

Figure 1. Uncurated generated samples at 256×256 for FFHQ dataset [6]. To align the comparisons, we use the same real query images for pre-trained generators to reconstruct. Our generation significantly outperforms the baselines in terms of quality, long-range dependencies, and spatial consistency.

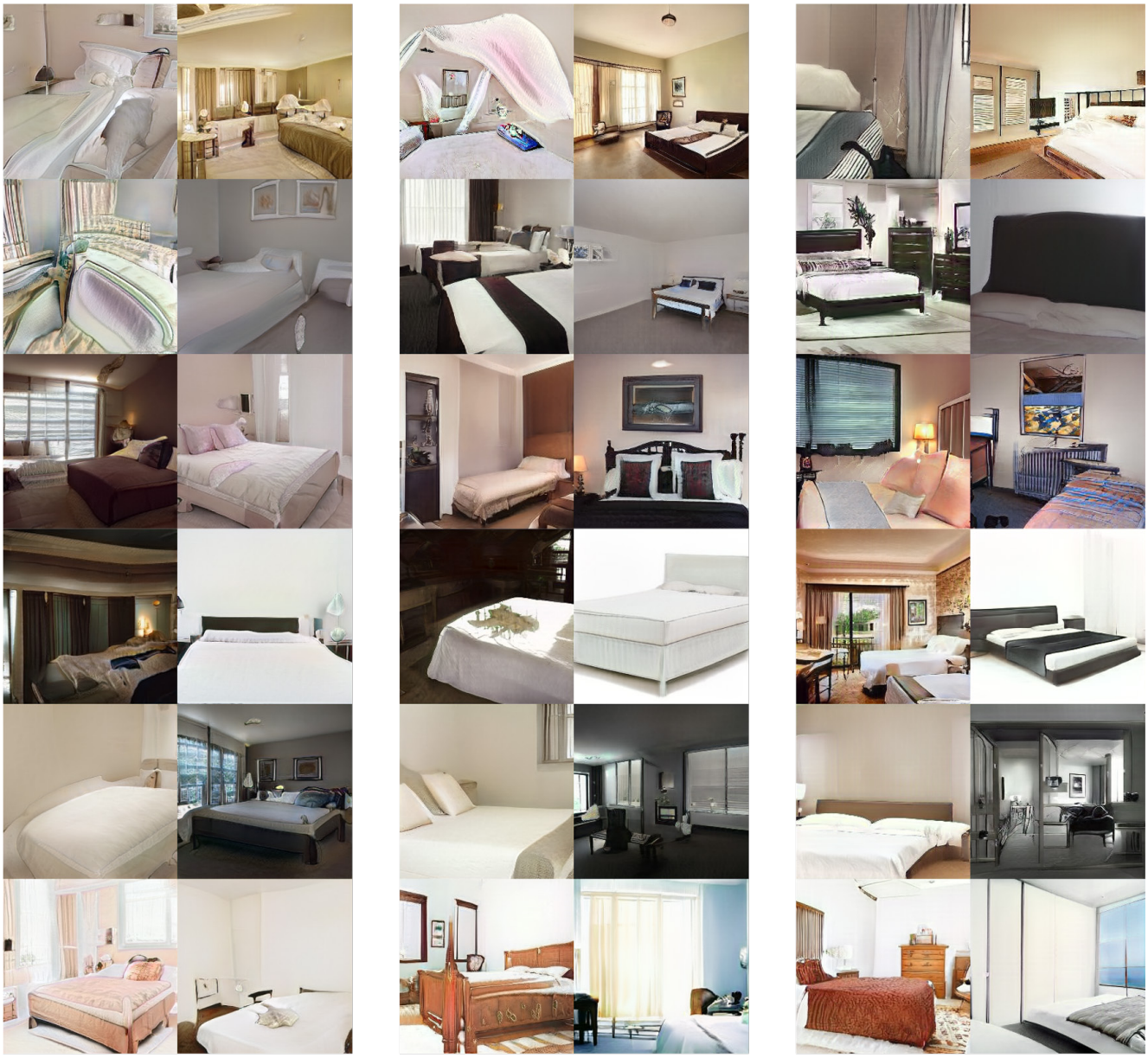
[12] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. A u-net based discriminator for generative adversarial networks. In *CVPR*, 2020. 2, 3

[13] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv*, 2019. 3

[14] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv*, 2015. 3, 5, 6, 7, 9, 10

[15] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 1

[16] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *CVPR*, 2020. 1



U-Net GAN
FID = 17.64

StyleGAN2
FID = 4.01

Ours
FID = 3.31

Figure 2. Uncurated generated samples at 256×256 for LSUN Bedroom dataset [14]. To align the comparisons, we use the same real query images for pre-trained generators to reconstruct. Our generation significantly outperforms the baselines in terms of quality, long-range dependencies, and spatial consistency.

U-Net GAN
FID = 11.74

StyleGAN2
FID = 4.54

Ours
FID = 3.39

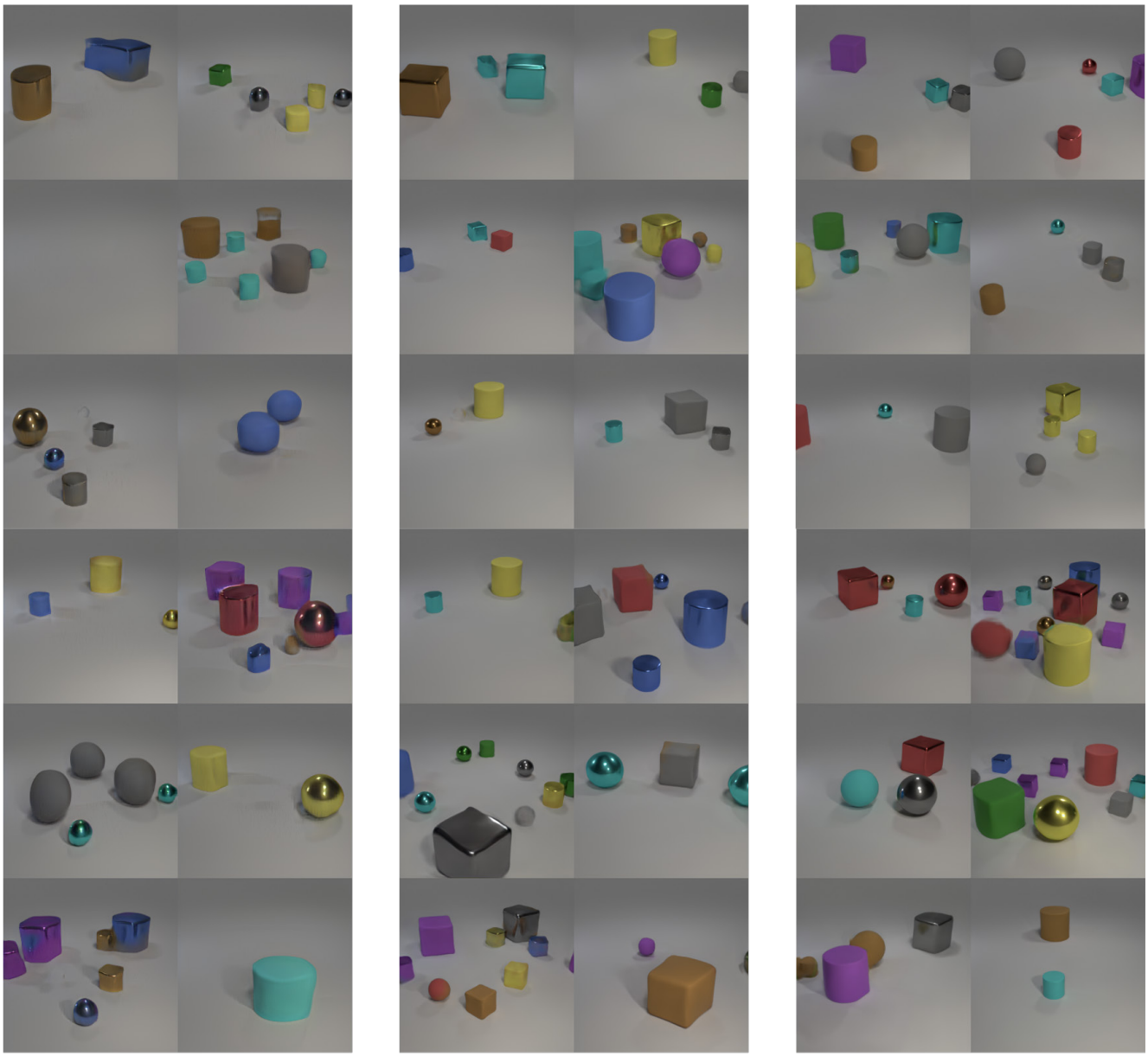
Figure 3. Uncurated generated samples at 256×256 for LSUN Church dataset [14]. To align the comparisons, we use the same real query images for pre-trained generators to reconstruct. Our generation significantly outperforms the baselines in terms of quality, long-range dependencies, and spatial consistency.

U-Net GAN
FID = 20.19

StyleGAN2
FID = 3.91

Ours
FID = 2.97

Figure 4. Uncurated generated samples at 256×256 for LSUN Horse dataset [14]. To align the comparisons, we use the same real query images for pre-trained generators to reconstruct. Our generation significantly outperforms the baselines in terms of quality, long-range dependencies, and spatial consistency.



U-Net GAN
FID = 33.32

StyleGAN2
FID = 9.62

Ours
FID = 5.05

Figure 5. Uncurated generated samples at 256×256 for CLEVR dataset [5]. To align the comparisons, we use the same real query images for pre-trained generators to reconstruct. Our generation significantly outperforms the baselines in terms of quality, long-range dependencies, and spatial consistency.

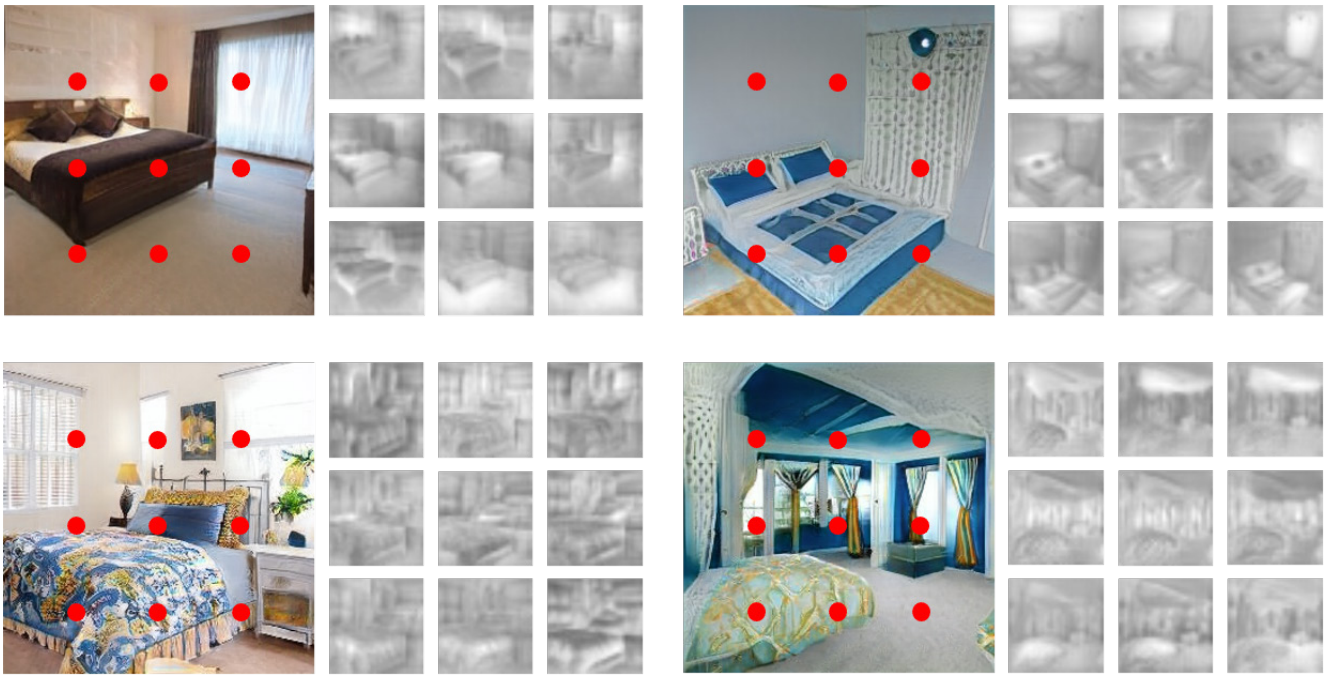


Figure 6. StyleGAN2 + SAN generated LSUN Bedroom [14] samples at 256×256 and their self-attention maps at 32×32 in the generator for the corresponding dot positions. Considering there is an attention weight kernel $\mathbf{w} \in \mathbb{R}^{s \times s \times c}$ for each position, we visualize the norm for each spatial position of $\mathbf{w}$. The attention maps strongly align to the semantic layout and structures of the generated images, which enable long-range dependencies across objects.

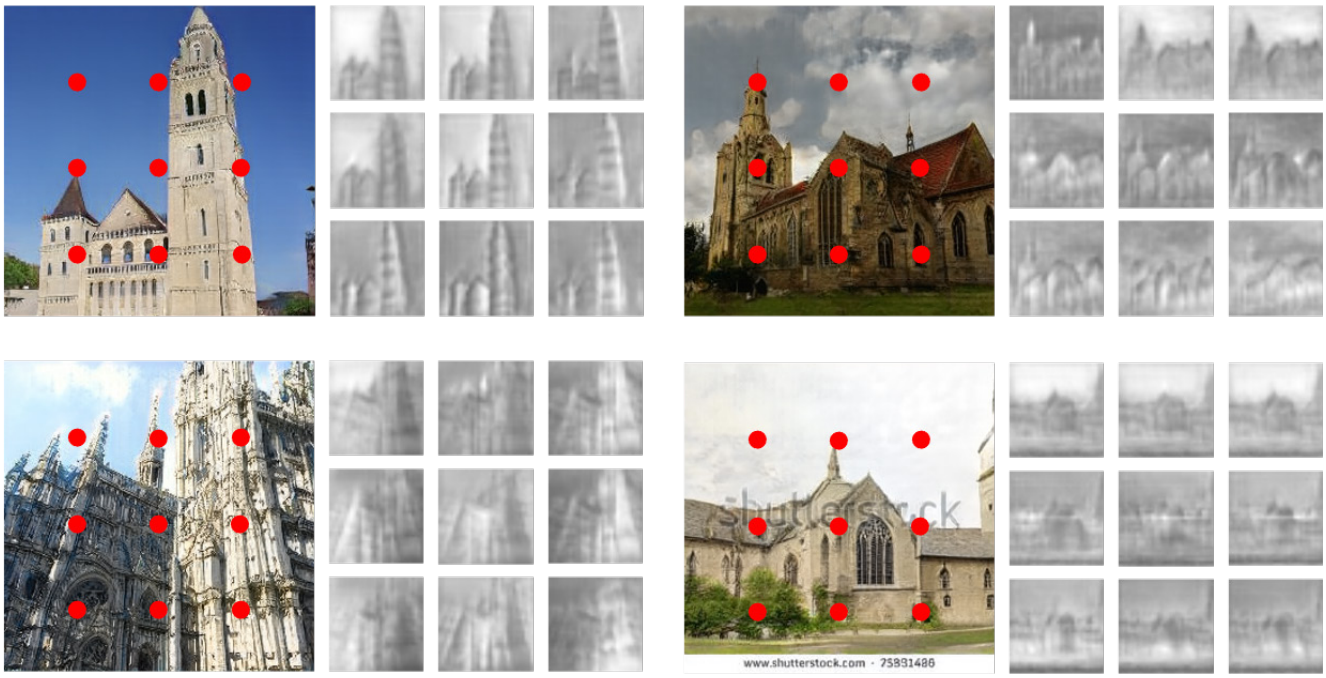


Figure 7. StyleGAN2 + SAN generated LSUN Church [14] samples at 256×256 and their self-attention maps at 32×32 in the generator for the corresponding dot positions. Considering there is an attention weight kernel $\mathbf{w} \in \mathbb{R}^{s \times s \times c}$ for each position, we visualize the norm for each spatial position of $\mathbf{w}$. The attention maps strongly align to the semantic layout and structures of the generated images, which enable long-range dependencies across objects.

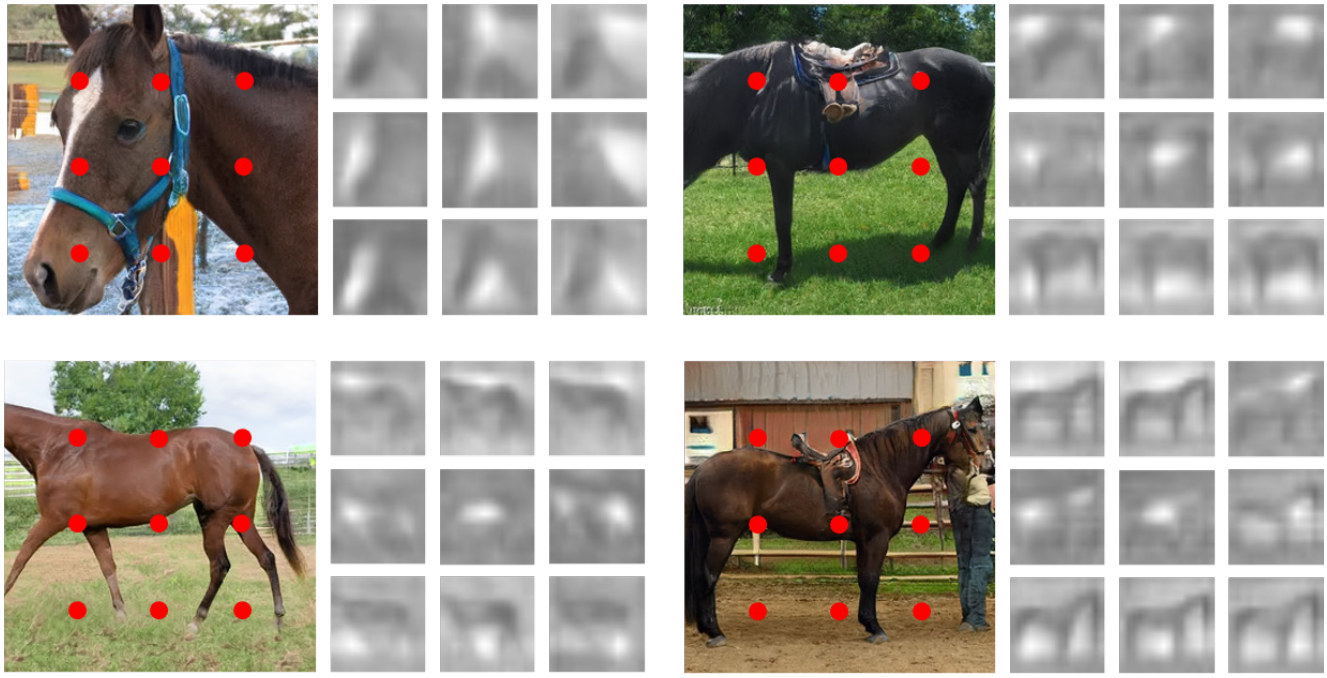


Figure 8. StyleGAN2 + SAN generated LSUN Horse [14] samples at 256×256 and their self-attention maps at 16×16 in the generator for the corresponding dot positions. Considering there is an attention weight kernel $\mathbf{w} \in \mathbb{R}^{s \times s \times c}$ for each position, we visualize the norm for each spatial position of $\mathbf{w}$. The attention maps strongly align to the semantic layout and structures of the generated images, which enable long-range dependencies across objects.

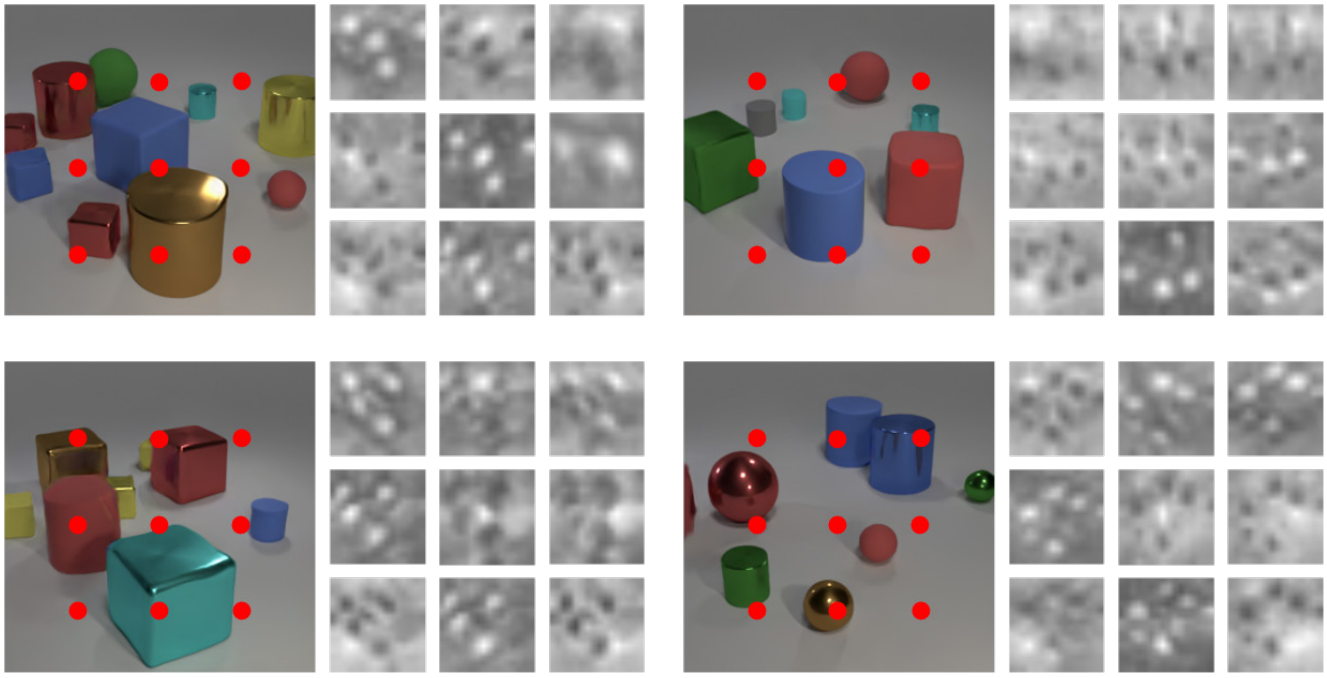


Figure 9. StyleGAN2 + SAN generated CLEVR [5] samples at 256×256 and their self-attention maps at 16×16 in the generator for the corresponding dot positions. Considering there is an attention weight kernel $\mathbf{w} \in \mathbb{R}^{s \times s \times c}$ for each position, we visualize the norm for each spatial position of $\mathbf{w}$. The attention maps strongly align to the semantic layout and structures of the generated images, which enable long-range dependencies across objects.