Improving GAN Equilibrium by Raising Spatial Awareness Supplementary Material

Jianyuan Wang[1,2] Ceyuan Yang[1] Yinghao Xu[1] Yujun Shen[1] Hongdong Li[2] Bolei Zhou[3]
[1]The Chinese University of Hong Kong [2]The Australian National University
[3]University of California, Los Angeles

A. Overview

As discussed in the paper, our EqGAN-SA also enables the interactive spatial editing of the output image. We build an interactive interface to better visualize this property, as illustrated in the demo video. In the appendix, we provide the following content. Sec. B includes the implementation details. Sec. C provides a description of the demo video, while Sec. D provides more ablation studies and analysis. Sec. E consists of the discussion on the validity of the disequilibrium indicator DI, our comparison to the methods of GAN manipulation, and some other discussions on manipulation.

B. Implementation

Training. We run the experiments on a computing cluster using a environment of PyTorch 1.8.1 and CUDA 9.0. For the convenience of reproducibility, we use the official PyTorch implementation of StyleGAN2 as our codebase. All the experiments follow the configuration of 'paper256' in the codebase. Specifically, we use the Adam optimizer with a learning rate of 2.5×10^{-3}. The minibatch size is 64 and the group size for the minibatch standard deviation layer is 8. The depth of the mapping network is 8. For all the datasets, we set the R_1 regularization weight γ as 1. We also adopt mixed-precision training for a speedup.

Architecture of SEL$_{concat}$. Same as its counterpart, SEL$_{concat}$ first uses a convolutional layer to extract features from the input heatmap, with a dimension of 64. It then concatenates the extracted features with the input feature map. Two convolutional layers are used after concatenation, with an intermediate dimension of 256. Similarly, SEL$_{concat}$ adopts a residual connection, and employs another convolutional layer for post-processing. All the convolutional layers use a kernel size of 3×3.

C. Interactive Editing

Although our method is proposed to improve GAN equilibrium and enhance image synthesis quality, it additionally supports a hierarchical manipulation on the generated image. The Fig. 5 and Fig. 6 of the main paper have shown this property. For better illustration, we provide an interactive interface for EqGAN-SA and show it in the demo video. Specifically, given a well-trained EqGAN-SA model, the 'Reset' button will randomly sample a latent code, and generate an image using the default spatial heatmaps. Users can move heatmap centers through dragging. The movement of centers updates the heatmaps in real time, shown in the second column from right. From top to down, the heatmaps correspond to 4×4, 8×8, and 16×16 feature resolution, *i.e.*, level $0, 1, 2$. Once setting the heatmaps, the users can click on the button 'Generate' to produce an image with the unchanged latent code and moved heatmaps. They can also click the 'auto' button under 'Generate', which enables automatic generation after each heatmap movement. Please note the level 1 and level 2 centers would automatically move with level 0 center, due to the design of hierarchical sampling.

In the demo video, we can see how the church tower reacts with heatmap moving, and how EqGAN-SA tries to produce high-quality results even under some extreme cases. We also observe that controlling a level 2 heatmap center can lead to a result like 'shaking' the ear of a cat. Besides the interface, we provide more dynamic samples in the end of the demo to show such manipulation is valid for diverse cases.

D. Ablation Study and Analysis

Hyper-parameters for Heatmap Sampling. As mentioned in the paper, we heuristically use two sub-heatmaps in the 8×8 feature resolution and four sub-heatmaps in 16×16. Here we provide an ablation study in Tab. 1 to show this setting is effective on the LSUN Cat dataset. In addition, though other settings may not be best, they all achieve reasonable results, and a clear improvement over the baseline. It verifies that the proposed method is robust to heatmap sampling hyper-parameters.

Hyper-parameters for Alignment. In the paper, Fig.

Table 1. **Ablation study on the hyper-parameters of spatial heatmap sampling,** on the LSUN Cat [12] dataset. Heuristically, using 2 heatmap centers in the 8×8 (level 1) feature resolution and 4 centers in the 16×16 (level 2) resolution lead to a good result. The baseline does not use spatial heatmaps, denoted as 'N/A'.

	Baseline	Level 1			Level 2		
Num	N/A	1	2	4	2	4	8
FID ↓	8.36	6.93	6.81	6.97	6.90	6.81	7.02
DI ↓	3.64	2.47	2.39	2.48	2.43	2.39	2.55

Table 2. **Ablation study on the hyper-parameters of $\mathcal{L}_{align}$,** on the LSUN Cat dataset. We explore the effect of the loss weight and the truncation threshold τ. Overall, the alignment regularization $\mathcal{L}_{align}$ is robust to various loss weights and τ is beneficial.

Loss Weight	0.25	0.50	1.00	1.50	2.00
FID ↓	6.99	6.88	6.81	6.79	6.83
DI ↓	2.46	2.41	2.39	2.43	2.41
Threshold τ	0.00	0.10	0.25	0.35	0.50
FID ↓	7.10	6.93	6.81	6.82	6.87
DI ↓	2.58	2.45	2.39	2.37	2.42

7 provides a qualitative analysis to support $\mathcal{L}_{align}$. Here we quantitatively explore the effect of its loss weight and truncation threshold τ, illustrated in Tab. 2. On the LSUN Cat dataset, different loss weights can generally lead to a satisfactory performance, where a number of 1.0 or 1.5 is close to best. Therefore, we use a loss weight of 1.0 for the experiments on all the datasets. We also prove that the truncation operation is beneficial, since our spatial heatmaps cannot perfectly match the real GradCAM maps with complex structures. Truncating the samples those have been 'good enough' reduces the difficulty of optimization. With the help of τ, we improve the FID from 7.10 to 6.81. Similarly, we use $\tau = 0.25$ for all the datasets.

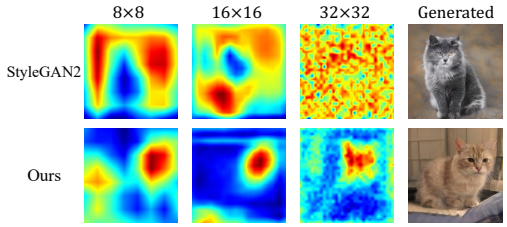

Figure 1. **Visualization of intermediate features of the generator.** A bright color indicates a high value. It can be verified that the generator of StyleGAN2 does not concentrate on the meaningful regions while ours does, especially at the resolution '32×32'.

Visualization of generator intermediate features. The spatial awareness of generator is worth investigating. However, CAM or GradCAM is not a suitable visualization tool because they both require a classification score, which is not applicable for a generator. Introducing another classifier may be a solution but it would involve the bias of the classifier. As an alternative solution, we could directly average the intermediate features of the generator along the feature dimension. Such a visualization can be viewed as *the contribution of a layer towards certain pixels*. As shown in Fig. 1, our generator shows a much clearer spatial preference than the baseline StyleGAN2 which presents random spatial focus, particularly at the 32×32 resolution. It indicates our method indeed brings spatial awareness to the generator.

E. Discussion

Claim of improving GAN equilibrium and Metrics. In this work, we use Disequilibrium Indicator (DI) to evaluate the degree of GAN equilibrium. Here, we would like to explain why DI is a reasonable metric:

> Recall that Wasserstein distance is a good indicator of GAN equilibrium [1, 2], and can be approximated as $W = \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{z \sim p(z)}[f(g(z))]$, where $f(\cdot)$ represents the discriminator. The approximation is valid as long as Lipschitz continuity stands. From this perspective, DI $= min[f(x)] - max[(f(g(z))]$ can be viewed as a specific form of W, and hence is a reasonable metric for GAN equilibrium.

Additionally, we report the results with various metrics, as shown in Tab. 3. Our method improves the baseline StyleGAN2 over all the metrics. The metric Recall is computed using the implementation of StyleGAN2.

Table 3. **Quantitative results on LSUN Cat Dataset with various metrics**. The metric W is the Wasserstein distance approximation introduced by WGAN [1].

	FID ↓	DI ↓	W ↓	Recall ↑
StyleGAN2	8.36	3.64	4.91	30.71%
+ Ours	**6.81**	**2.39**	**4.21**	**37.78%**

Comparison to GAN Manipulation Methods. Although not designed to do so, our EqGAN-SA provides an alternative way to manipulate the output synthesis of GAN. Previous methods mostly manipulate synthesis through interpolating latent code, since the latent space of GAN has been found to encode rich semantic information. For a certain attribute, they search for a certain direction in the latent space, and then alter the target attribute via moving the latent code z along the searched direction [3,4,8,10,11]. However, for each pre-trained GAN model, these methods require to annotate a collection of the generated samples. They use the annotated samples to train linear classifiers in the latent space. Instead, our EqGAN-SA achieves the manipulation ability in an unsupervised way, relying on self-emerging attention.

Recently, a method SeFa [9] proposes a closed-form factorization algorithm to find semantically meaningful directions in the latent space without supervision. Unfortunately, it does not guarantee the attributes of found directions. For example, if wanting to spatially manipulate the synthesis

like EqGAN-SA, the user has to manually check the effect of various directions, while the one corresponding to spatial manipulation may not exist. In addition, EqGAN-SA supports editing both on the overall location and the local structure, while SeFa [9] cannot. Moreover, a concurrent work [5] generates $2D$ keypoints from the latent space and associate them to appearance embeddings. It encodes the generated keypoints into styles maps, adopting a similar technical choice to ours. Compared to EqGAN-SA, it can only change the local structures of synthesis.

Dataset Limitation for Manipulation. We have observed that moving heatmaps will alter various contents on different datasets. For example, it changes the cat faces on the LSUN Cat while the church towers on the LSUN Church. This matches our design target, since the discriminative/attentive contents are different on various datasets, but yields to difficulty in manipulation. In addition, the diversity of the training dataset limits the manipulation result. For example, the face images in the FFHQ [6] dataset have been well-aligned, *i.e.*, the location of face is constrained to a vertical range (close to center). Therefore, moving heatmaps to top cannot lead to an image with the face at the top.

Artifacts during Heatmap Movement. We notice that there are artifacts when interpolating spatial heatmaps, *e.g.*, blurring at the location of heatmaps boundaries. We attribute this to the instability from the outputs of spatial encoding layers. Such an instability may be mitigated via involving a regularization in the way of path length regularization [7], *i.e.*, requiring a fixed-size step of heatmap movement to have a fixed-magnitude change in the image. We plan to solve these artifacts in the future work, which may further improve our synthesis quality.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Int. Conf. Mach. Learn.*, pages 214–223, 2017. 2

[2] David Berthelot, Thomas Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017. 2

[3] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Int. Conf. Comput. Vis.*, pages 5744–5753, 2019. 2

[4] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Adv. Neural Inform. Process. Syst.*, pages 9841–9850, 2020. 2

[5] Xingzhe He, Bastian Wandt, and Helge Rhodin. Latentkeypointgan: Controlling gans via latent keypoints. *arXiv preprint arXiv:2103.15812*, 2021. 3

[6] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4401–4410, 2019. 3

[7] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8110–8119, 2020. 3

[8] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[9] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2, 3

[10] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *Int. Conf. Mach. Learn.*, pages 9786–9796, 2020. 2

[11] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis. *Int. J. Comput. Vis.*, pages 1451–1466, 2021. 2

[12] 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. 2