Toward Spatially Unbiased Generative Models

Jooyoung Choi[1] Jungbeom Lee[1] Yonghyun Jeong[3] Sungroh Yoon[1,2,*]
[1] Data Science and AI Laboratory, Seoul National University, Korea
[2] ASRI, INMC, and Interdisciplinary Program in AI, Seoul National University, Korea
[3] AI Research Team, Samsung SDS
{jy_choi, jbeom.lee93, sryoon}@snu.ac.kr yhyun.jeong@samsung.com

Abstract

Recent image generation models show remarkable generation performance. However, they mirror strong location preference in datasets, which we call ***spatial bias****. Therefore, generators render poor samples at unseen locations and scales. We argue that the generators rely on their implicit positional encoding to render spatial content. From our observations, the generator's implicit positional encoding is translation-variant, making the generator spatially biased. To address this issue, we propose injecting explicit positional encoding at each scale of the generator. By learning the spatially unbiased generator, we facilitate the robust use of generators in multiple tasks, such as GAN inversion, multi-scale generation, generation of arbitrary sizes and aspect ratios. Furthermore, we show that our method can also be applied to denoising diffusion probabilistic models. Our code is available at:* https://github.com/jychoi118/toward_spatial_unbiased.

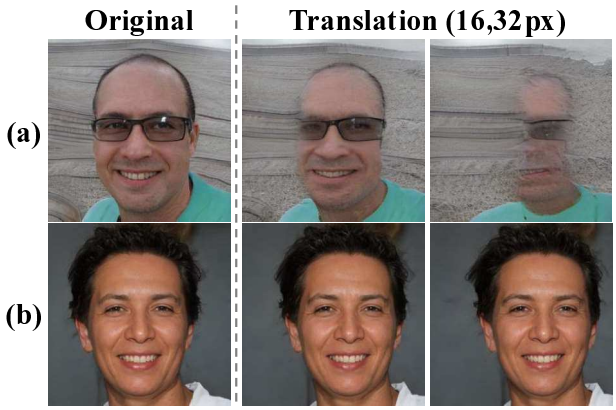

Figure 1: **Generation at translated locations.** (a) Baseline StyleGAN2; (b) Spatially unbiased StyleGAN2. 256^2 images generated at 16, 32 pixels rolled locations. (a) shows distorted and harmed samples. With our method, a model can generate well-structured images in unseen locations.

1. Introduction

Recent CNN-based generative models [5, 14, 17, 19, 20] generate images of remarkable quality by learning the distribution of well-designed datasets. The careful design of a dataset [19, 38], along with architecture design [5, 17, 19], is an important factor [11] in the development of generative model performance. First, humans carefully collect high-quality images [19]. Then, to benefit learning, the collected data are pre-processed according to human inductive bias. Pre-processing may contain center-cropping [38], resizing, mirror padding, and alignment based on landmark annotation [17, 22]. Eventually, the images may have a strong location preference, which we call ***spatial bias***, due to the photographer's bias in the initial data collection and pre-processing of the collected data. Both are used so the model learns the underlying concepts of the dataset.

The collection and processing of the dataset are intended to improve learning, but the resulting spatial bias can be an unintended shortcut for the model. If the spatial bias is a cue that helps the model learn the data distribution easily, the model can rely on this shortcut [4] despite the sufficient model capacity. We demonstrate that generator components, such as constant tensors and zero-padding, jointly work as implicit positional encoding. Since they are located at fixed absolute positions, such implicit positional encoding is translation-variant. Thus, it encourages the model to rely on spatial bias. This phenomenon may be undesirable from a generalization perspective. As seen in Fig. 1(a), a spatially biased model generates distorted and harmed samples at shifted positions.

Therefore, our goal is to utilize a well-designed dataset to learn rich representations without relying on spatial bias. We explore implicit positional encoding of the model, analogous to studies in discriminative tasks [16, 21]. At each scale of the generative models, positional information introduced from the previous scale and zero-padding in the

*Correspondence to: Sungroh Yoon (sryoon@snu.ac.kr)

current scale jointly work as implicit encoding. To render spatial content, the generator relies on implicit positional encoding. We demonstrate that this implicit cue is translation-variant and therefore suffers from spatial bias. To avoid reliance on implicit cues, we propose injecting an explicit scale-specific positional embedding into the generator. Our method learns spatially unbiased and translation-equivariant generators, which generate content according to a given positional encoding.

Since we provide explicit scale-specific positional information, the generator no longer needs to rely on its implicit cue. Therefore, the generator creates content according to the given positional information, which enables translation-equivariant generation. Fig. 1(b) shows consistent facial features generated at unseen positions. We illustrate that multi-scale positional encoding enables the generation of an image at an unseen position by manipulating the positional embedding at the inference time.

Concurrent to our work, Karras *et al.* [18] observe that previous GANs are not translation-equivariant and the details stick to the absolute location. Based on signal processing, [18] introduce multiple architectural changes to enable the coarsest feature to control the position of finer features. In constrast, we control positions with explicit scale-specific embeddings, without further architectural changes.

Through extensive experiments, we demonstrate that multi-scale positional encoding is effective in learning spatially unbiased generators. With only minimal modification in model architectures, our method maintains the generation quality of previous well-performing models, as shown by comparison of FID [13] as well as Precision and Recall [30]. Nevertheless, our method leads to the spatially unbiased generator. Removing spatial bias further encourages extension to various applications. We demonstrate robustness to affine transformations in GAN inversion [1, 6, 12, 37, 41] and image edition [1, 12, 33]. We show the effectiveness in multi-scale generation with a model trained on a single resolution. Our method enables the generation of arbitrary sizes and aspect ratios by expanding and extrapolating positional embedding in the test time. Furthermore, we also apply our method on Denoising Diffusion Probabilistic Models (DDPM) [14, 34], which is a recent family of generative models. The main contributions of this work are threefold:

- We discuss that existing generators rely on translation-variant implicit positional encoding, making the generators spatially biased.
- We propose a method for learning a spatially unbiased generator with a minimal modification in well-performing architectures.
- We validate our method on multiple tasks: GAN inversion, multi-scale generation, generation of arbitrary sizes and aspect ratios, and adaptation to DDPM.

2. Related Work

2.1. Translation Variance of CNN

Translation invariance is an essential feature of convolution operations with advantages in computer vision tasks such as robustness of classification [39] and object detection [26]. However, despite the translation-invariant convolutional operation, modern CNN architectures are not translation-invariant. Various reasons have been studied: photographer's bias [26] in a dataset, zero-padding implicitly providing absolute positional information [16, 21, 36], and lack of consideration of the sampling theory in downsampling operations [3, 39]. The resulting translation-variant model can cause inconsistent inferences depending on the input position [2, 3, 21, 26, 39].

2.2. Positional Encoding

Positional encoding plays a simple yet powerful role in various domains such as natural language processing (NLP) and computer vision. In the transformer architecture [35], the order of the sequence is provided by adding sinusoidal embedding to the input feature. Similarly, the transformer-based object detection model DETR [7] informs the spatial order of features at each location with 2D sinusoidal embedding. Several studies [7, 10] have shown that the presence of positional embedding significantly improve the performance, while different embedding variants (e.g., sinusoidal embedding, Cartesian grid, learned constant) yield similar performances. Positional encoding also plays an important role in 3D vision studies. NeRF [28] maps input coordinates to a high-dimensional space with sinusoidal embedding to learn high-frequency 3D representations.

2.3. Bias in Generative Models

Recent studies [11, 27, 40] on the bias of generative models examine that the model can reflect the bias of the learned dataset. Esser *et al.* [11] demonstrated that a high-quality dataset leads to high-quality generation by disentangling data-biased and unbiased content. In PULSE [27], a super-resolution study with a StyleGAN [19] model, it was discovered that there is a facial color bias in the generated images due to severe imbalance in the dataset and GAN's mode-collapse behavior. In contrast, we consider spatial bias. Huh *et al.* [15] found an object-center bias in the BigGAN [5]. Therefore, it was difficult to project a real image into a latent space. The paper addressed this using a separate transformation optimization method. Analogous to our work, Xu *et al.* [36] found that zero-padding works as implicit positional encoding in SinGAN [31], then provided explicit encoding at the coarsest scale of the model. In contrast, we investigate implicit encoding at each scale and observe that explicit encoding at the coarsest scale is insufficient for learning spatially unbiased models.

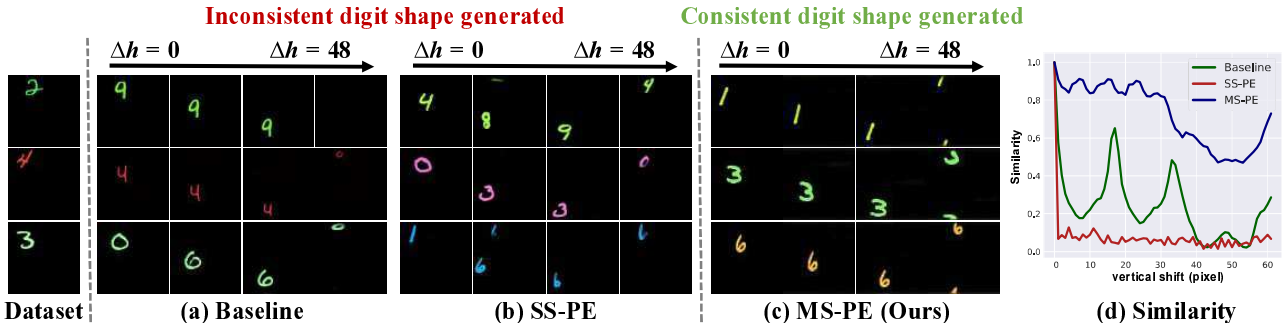

Figure 2: **Generation at shifted locations.** We generated digits at locations shifted vertically by $\triangle h$, sharing the same inputs. We trained generators on spatially biased images shown on the left. (a) and (b) exhibit inconsistent digits at shifted locations, while our method (c) generates consistent digits. (d) exhibits the similarity measured during successive translations.

3. Method

Before we analyze the spatial bias of generative models, we introduce the background of StyleGAN [19, 20], which is our baseline, in Sec. 3.1. In Sec. 3.2, we introduce multi-scale positional encoding (MS-PE), the method of learning spatially unbiased generative model. Sec. 3.3 shows that existing generators rely on implicit positional encoding, which is translation-variant, and that our method addresses this issue. We describe how to generate at an unseen position and resolution in Sec. 3.4.

3.1. Preliminary

To analyze the spatial bias of the CNN-based generation models, we focus on StyleGAN [19, 20], where three inputs are provided: 1D latent vector w, 2D noise map ϵ, and a learnable constant tensor c. Each is known to serve as globally coherent style [19, 32], spatial details [19], and implicit positional encoding [36] respectively.

First, we investigate the spatial bias of StyleGAN using our customized dataset. We create a spatially biased color-MNIST, where digits are located only in the upper-left corner. In our experiments regarding shifts, we used a circular shift where elements shifted beyond the boundary are "rolled" to the other side.

The original StyleGAN, which we denote as "baseline", generates inconsistent shapes from rolled constant tensor c, as shown in Fig. 2(a). Thus, we can infer that a constant tensor c does not serve as an accurate positional encoding.

3.2. Multi-Scale Positional Encoding

In this section, we introduce a method of providing scale-specific positional information, which we call *multi-scale positional encoding* (MS-PE). Let the L-scale be a generator with scale $l \in \{1, ..., L\}$, spatial dimensions H^l, W^l, channel number C^l and feature h^l. We use a simple yet powerful sinusoidal positional embedding widely used in transformer literatures [7, 10, 35]. Sinusoidal positional embedding is a predefined code consisting of different sine or cosine function values for each location. We add the code to the convolutional feature. The dimension is equal to the number of feature channels. The code representing the location (i, j) is as follows:

$$PE_{(i,j)} = [pe_{(i,0)}, ..., pe_{(i,2d-1)}, pe_{(j,0)}, ..., pe_{(j,2d-1)}], \tag{1}$$

$$pe_{i,2k} = \sin(i/100^{k/d}),\ pe_{i,2k+1} = \cos(i/100^{k/d}), \tag{2}$$

where $k = 0, 1, ..., d-1$ and d is a quarter of the channel dimension. PE^l is the positional embedding of the l-th scale, a set of $PE_{(0,0)}, ..., PE_{(H^l,W^l)}$. Note that PE^l and h^l have the same dimensionality.

Additionally, we multiply the learnable scalar γ^l before adding the positional encoding of each scale to the feature. Learnable scalars allow the generator to adjust the scale of positional information as desired. The features at each scale are calculated as follows:

$$h^{l+1} = f^l(PE^l, w^l, h^l) = G^l(w^l, h^l) + \gamma^l PE^l, \tag{3}$$

where G^l and w^l are the convolutional layer and latent vector of the l-th scale, respectively. Note that the feature at each scale is a function of the latent vectors and positional embeddings. As an exception, positional encoding is given as a feature input in the coarsest scale:

$$h^1 = G^0(PE^0, w^0). \tag{4}$$

Xu *et al.* [36] also provided explicit positional encoding as an input tensor in the coarsest scale to offer spatial inductive bias to a generator. However, in the next section, we demonstrate that this is insufficient, and that the multi-scale positional encoding (MS-PE) of Eq. 3 is required.

Combining Eq. 3 and Eq. 4, the generative process can be denoted as:

$$I = f^L(PE^L, w^L, f^{L-1}(..., f^0(PE^0, w^0))). \tag{5}$$

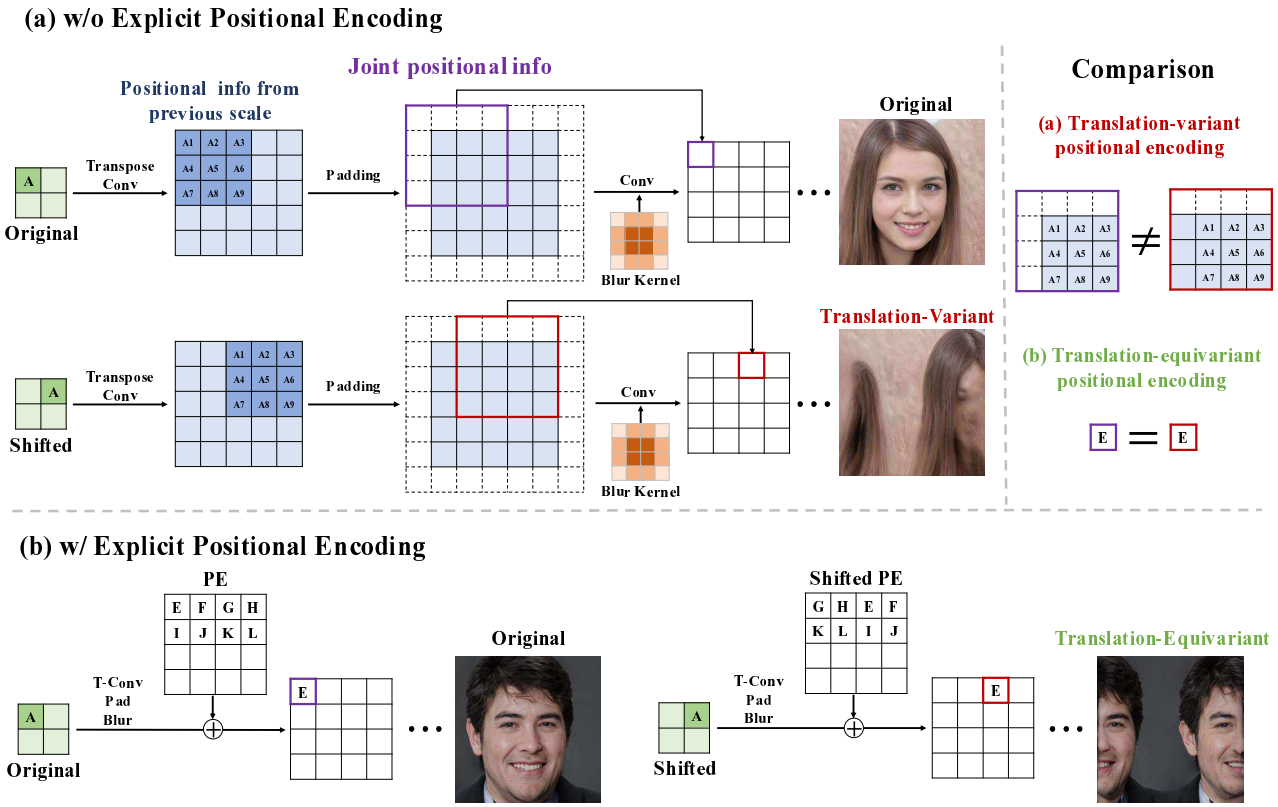

Figure 3: **Illustrative example of shifted generation.** We illustrate how implicit and explicit positional encoding affect generation at shifted location. Alphabets represent location indices. **(a)** Positional information is extracted from previous scale and zero-padding. Shift in previous scale causes inconsistent shift of positional encoding in current scale. Therefore destructive image is generated. **(b)** Positional information is given explicitly. We simply shift explicit encoding according to the shift of the previous scale, thus provide consistent positional encoding.

3.3. Why Does MS-PE Work Better?

Before we demonstrate the effectiveness of MS-PE, we first study the implicit positional encoding of the generator. Fig. 3 exhibits an example sequence of transpose convolution and anti-aliasing (blur) [39] of StyleGAN [19, 20] upsampling a 2×2 feature to a 4×4 feature. From location A in 2×2 feature, upsampling sequence introduces detailed positional information A1–9. Then, the shape of zero-paddings [16] is encoded by anti-aliasing convolution. Positional information A1–9 from the previous scale and zero-padding jointly serve as implicit positional encoding, providing unique positional information to each location. These implicit cues are located in fixed absolute locations. As a consequence, implicit positional encoding at shifted location differs from that at the original location, as illustrated in the top-right corner of Fig. 3. The different shape of zero-padding is encoded in the shifted location. Therefore, the implicit positional encoding is translation-variant.

Baseline StyleGAN's learnable constant tensor c and zero-padding are known to jointly work as implicit positional encoding [36]. The generator renders spatial content relying on these implicit cues. However, since the generator's implicit cue is translation-variant , it leads to poor generation quality at shifted locations, as shown in Fig. 3(a). In other words, the generator is spatially biased, failing to generalize to unseen positions. The generator renders reasonable images only at the learned positions. Fig. 1(a) and Fig. 2(a) show distorted samples at unseen positions.

Our motivation is to provide positional information explicitly to prevent the generator from relying on implicit cues. As a starting point, we replace the constant tensor c with explicit positional encoding following [36], which we call single-scale positional encoding (SS-PE). Surprisingly, the SS-PE in Fig. 4 generates the different shapes of digits according to the 2D noise map ϵ, despite the same latent vector w. As mentioned previously, the constant tensor c of the baseline implicitly encodes the positional information across the image. In contrast, explicit positional encoding

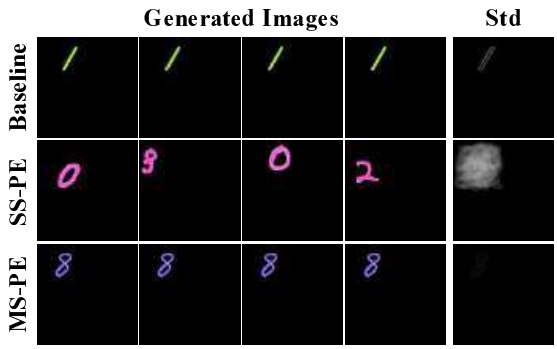

Figure 4: **Effect of 2D noise input.** Standard deviation of each pixel is calculated over 100 2D noise map instances. SS-PE generate inconsistent shapes and the largest standard deviations compared to the baseline and MS-PE.

of SS-PE captures unique positional information only for the coarsest scale (e.g., 4×4). Eventually, zero-padding is the only component that serves as an implicit encoding at other scales, which is insufficient for providing positional information to every location. Thus, the generator has no choice but to rely on spatially varying 2D noise maps ϵ to render shapes.

In contrast, the MS-PE provides unique positional information for every location at each scale. The generator no longer needs to rely on its implicit encoding. Fig. 3(b) illustrates that explicit encoding is translation-equivariant and thus spatially unbiased. Therefore the generator successfully learns the mapping between the given positional information and the related content, thus generating consistent digits, as shown in Fig. 2(c), even at the unseen positions. Fig. 2(d) shows the similarity measured during successive translations. MS-PE exhibits the most consistent generation. The similarity of the baseline are lower by a large gap, suggesting that the c is translation-variant and inaccurate positional encoding. SS-PE shows the lowest similarity since digits change rapidly according to 2D noise.

3.4. Generation from Unseen Position and Scale

MS-PE effectively learns spatially unbiased generators that generate images according to the given positional information. Therefore, a simple modification of positional encoding at inference time enables the generation of images at unseen positions and resolutions.

Let us consider a L-scale generator. A single pixel shift in the l-th scale results in a 2^{L-l} pixel shift in the image space. In the opposite direction, the k pixel shift in the image space corresponds to a shift of $k \times 2^{l-L}$ in the l-th scale. Therefore we shift the positional encoding in each scale l by $k \times 2^{l-L}$ to generate images in k pixel shifted coordinates. Following Eq. 5, $Shift_{\triangle h,\ \triangle w}(I)$ is a function of $PE^{l}_{i-\frac{\triangle h \% H}{2^{L-l}},\ j-\frac{\triangle w \% W}{2^{L-l}}}$.

Sinusoidal embedding allows a decimal number of coordinates, which enables generation from continuously translated coordinates. In Fig. 2(d), MS-PE shows consistent similarity at shifted locations by a single pixel. Meanwhile, as the constant tensor c does not support shift by a decimal number, the similarity of the baseline plunge at shifts except 16 and 32. Similarly, 32 pixels shifted sample of Fig. 1(a) shows destructive facial features.

Moreover, explicit positional encoding can express different resolutions through a simple interpolation [36]. This motivates multi-scale generation in Sec. 4.2, where images of multiple resolutions are generated with a model designed for a single resolution.

4. Experiment

We validate our method with StyleGAN2 [20] trained on customized color-MNIST, FFHQ [19] and LSUN-Church [38] and show translation-equivariant generations in Fig. 1 and Fig. 2. We further demonstrate how a spatially unbiased generative model is beneficial for various generation tasks. Sec. 4.1 demonstrates robustness in GAN inversion. Sec. 4.2 presents multi-scale generation with a single model. Sec. 4.3 presents generation of arbitrary sizes and aspect ratios. Sec. 4.4 reports the quantitative results. In Sec. 4.5, we show an extension of our method to the denoising diffusion probabilistic model. See the supplementary material for details on implementation and evaluation.

4.1. GAN Inversion

Recent GAN inversion literatures [1, 12, 15, 27] leverage well-performed GANs on various image processing tasks. They first project [1, 20] the real image into the latent space of the GAN by optimization. Then, exploring the latent space enables various image manipulations and restorations such as style transfer [1], class editing [15], face editing [32], and super-resolution [12, 27]. However, as studied in the Im2StyleGAN [1], GAN inversion is sensitive to affine transformations such as translation and resizing.

To demonstrate our robustness to transformations in GAN inversion, we optimized the StyleGAN [19, 20] latent vector for a given image following StyleGAN2 [20] literature. In addition, we edited images by moving the optimized latent vector along the directions found with SeFa [33]. Since our method frees the generative model from spatial bias, it enables robust GAN inversion and edition, as shown in Fig. 5. Baseline StyleGAN fails to reconstruct unaligned face images, whereas ours reconstructs faces shifted in the vertical and horizontal directions and zoomed faces in various positions. Edited images are even more destructive than the reconstructed images. We expect these results to broaden the applicability of well-performed generative models in image-processing tasks.

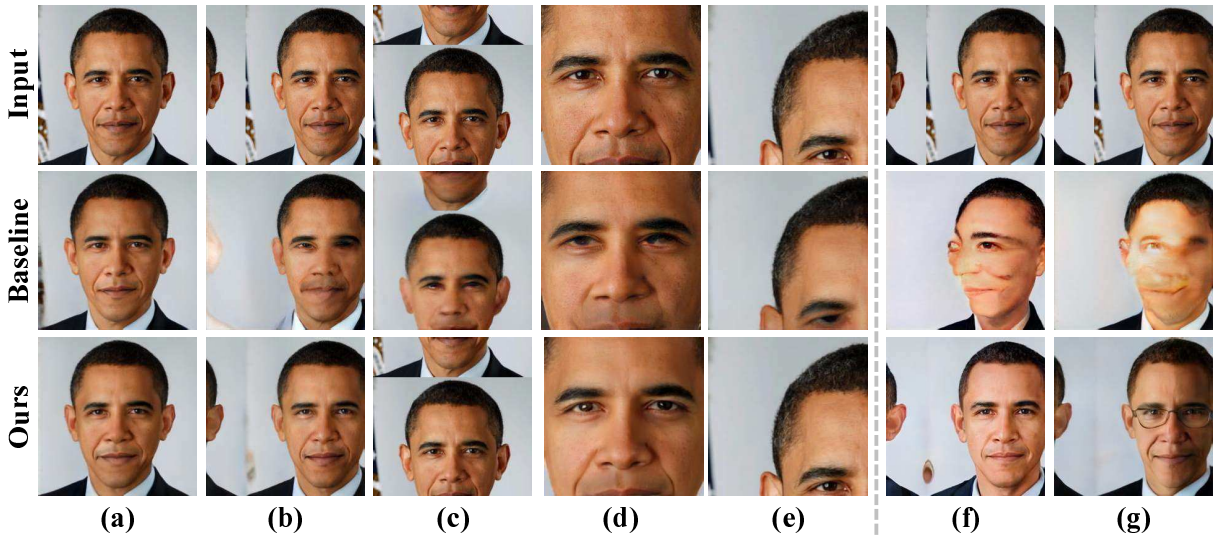

Figure 5: **GAN Inversion.** (a) Standard position. (b) Translation of 64 pixels to the right. (c) Translation of 64 pixels to the bottom. (d) Zoom in by 2x in top-left corner. (f,g) Edition following [33]. Generator learned by our method show robust GAN inversion and edition at various locations and scales.

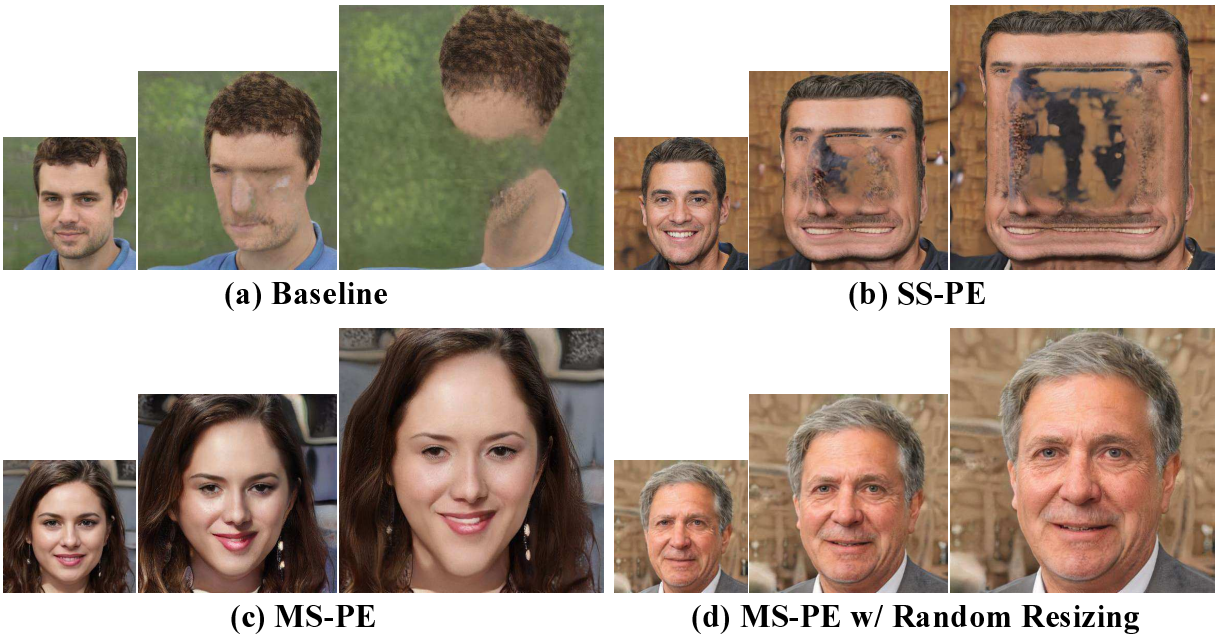

Figure 6: **Ablation on Multi-Scale Generation**. 256^2, 384^2, 512^2 resolution images generated with a single model. Generated by interpolating (a) constant tensor, (b) positional encoding in the coarsest scale, (c, d) positional encoding in multi-scale. Models learned by our method (c, d) generate reasonable faces in various resolutions. (b) Face structure in the boundary implies strong influence of zero-padding.

4.2. Multi-Scale Generation

Fig. 6 shows the qualitative comparison for multi-scale generation. As the constant tensor of the baseline leads to the generator biased to 256^2 resolution, Fig. 6(a) shows destructive samples for the unseen scales due to the reliance on constant tensor c and zero-padding. Since SS-PE lacks positional information from the previous scale as argued in Sec. 3.3, only zero-padding operates as positional cue.

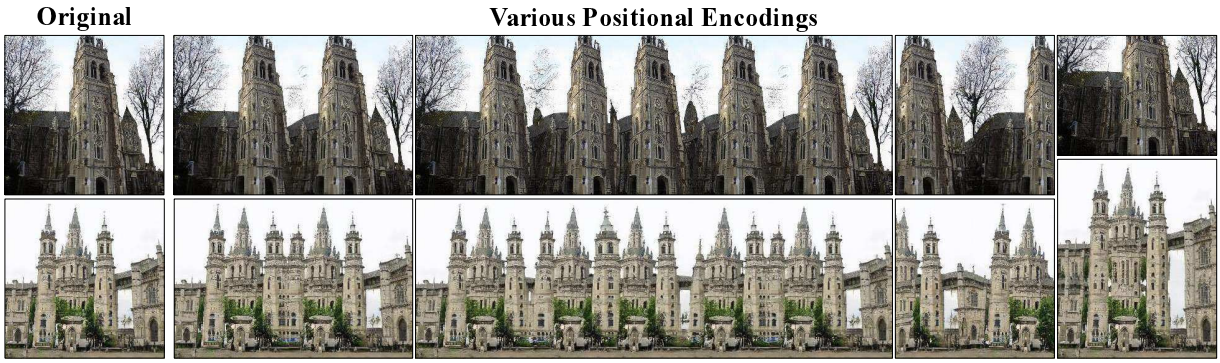

Figure 7: **Expansion to arbitrary sizes and ratios.** Generated images expanded to arbitrary sizes and ratios while preserving their patch distribution. Towers, trees, and bridges are repeated in various locations.

Figure 8: **Generation beyond the boundaries.** 256x384, 256x512 resolution images generated with a model trained on 256x256 by extrapolating positional encodings beyond the boundaries. Red box denotes original area the model learned.

Therefore, the generator relies strongly on zero-padding and shows reasonable structure only at the boundary in Fig. 6(b). As shown in Fig. 6(c), our method generates reasonable global structures at various scales. In table 1, MS-PE exhibits better FID scores in all resolutions than the others by a large gap. However, generated images are awkward in detail since convolution has a scale-variant nature. Therefore, we train the model to generate multiple resolutions by randomly resizing (MS-PE w/ Random Resizing) the explicit positional encoding at each iteration. Still, the discriminator learns only on a single resolution. To match the discriminator's input resolution, we resize generated image of random sizes to 256^2. With this learning strategy, Fig. 6(d) shows consistent high-quality samples in various scales generated with a single model.

4.3. Generation of arbitrary sizes and aspect ratios

From Sec. 3.3, the generator learns a mapping between explicit positional encoding and the corresponding content. This property motivates generation from expanded coordinate grids of arbitrary sizes and aspect ratios. By repeating the desired part of the positional encoding, images of arbitrary sizes and aspect ratios can be generated while preserving their patch distribution. Fig. 7 shows samples of various resolutions generated with a single model trained at a resolution of 256^2. As can be observed, towers, trees, and bridges are rendered in various locations and orders. Expanding generated images is studied in SinGAN [31], which train an unconditional model on a single image. However, we used a model trained on a large dataset.

Our method not only expands images but also enables generation beyond the image boundary. Since our method learns a spatially unbiased generator, the generator can also generate contents beyond the learned position. We trained model on the coordinate grid $[0, H^l] \times [0, W^l]$ for each scale l, and we can generate from the coordinate grid, for example, $[0, H^l] \times [-0.5W^l, 1.5W^l]$ at each scale. Fig. 8 shows realistic samples even beyond the boundary. Extra objects such as trees, humans, buildings are generated. A similar experiment on extrapolation was presented in [24, 25], how-

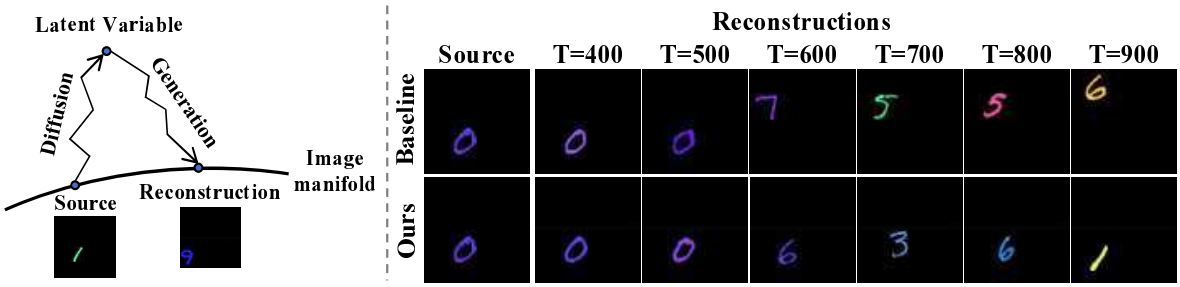

Figure 9: **DDPM Reconstructions.** From T=600, baseline DDPM generates digits in the upper-left corner as it learned spatial bias. Ours can consistently generate in given position.

ever, our work differs in that we generated a full-resolution image with an unconditional model.

4.4. Quantitative Results

We evaluated our method with FID [13] as well as Precision and Recall [30]. We compare our MS-PE with baseline and SS-PE, those described in Sec. 3.1. Our method shows better performance in FFHQ [19], suggesting that the model learned rich representation from spatial bias in the dataset while not relying on it. Recall increased rapidly, suggesting that unbiased learning led to improved diversity. Our method maintains performance in LSUN-Church [38], while enabling diverse generations as shown in Sec.4.3.

4.5. Additional Experiment on DDPM

Denoising Diffusion Probabilistic Models (DDPM) [14, 34] is a recently introduced family of generative models which have shown powerful generation performance in various domains [8, 9, 14, 23, 29]. It learns the reverse of the *diffusion process* where 2D noise maps are gradually added to the image. Since each step of the generation process (1000 steps in total) is stochastic, highly diverse samples are generated from the same latent variable. See supplementary materials for details on DDPM.

Ho *et al*. [14] presented stochastic reconstruction, where images are encoded into an intermediate time step and then decoded by a learned generation process. Both encoding and decoding processes are stochastic. Therefore, as the image is encoded into a larger time step, more diverse samples are generated. We train the DDPM on color-MNIST biased to the upper-left corner, then provide inputs biased to the bottom-left corner. In Fig. 9, when reconstructed from time step over 500, baseline generates digits in the upper-left corner. This indicates that the baseline DDPM is spatially biased. However, the model trained by our method reconstructs digits at the bottom-left, which is an unseen position. This example indicates that MS-PE can also be applied to train spatially unbiased DDPM.

Config	@256	@384	@512
Baseline	7.21	43.11	154.40
SS-PE	7.37	94.44	187.78
MS-PE (Ours)	**6.75**	**16.03**	**30.41**

Table 1: **Ablation on multi-scale generation.** FID scores with 256^2, 384^2, 512^2 images generated by 256^2 models. Our method shows superior performances.

	FFHQ			LSUN-Church		
Config	FID↓	P↑	R↑	FID↓	P↑	R↑
Baseline	7.21	64.07	35.20	**4.70**	59.74	**39.74**
SS-PE	7.37	**64.19**	36.97	5.01	**60.53**	36.43
MS-PE	**6.75**	62.97	**42.27**	6.42	55.91	34.20

Table 2: **Quantitative comparison.** FID, Precision & Recall (%) calculated with 256^2 images.

5. Conclusion

We introduced a simple method for learning spatially unbiased generative models. We demonstrated that generators rely on translation-variant implicit positional encoding to generate spatial content. This caused spatially biased generator, which show degraded performance at unseen position and scale. By injecting explicit positional encoding in each scale, we learn generators that render images according to the given positional information. Concurrent work [18] presented unaligned version of existing datasets. We are looking forward to training on such unaligned datasets.

Acknowledgements: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT)[2018R1A2B3001628], AIRS Company in Hyundai Motor and Kia through HMC/KIA-SNU AI Consortium Fund, and the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2021.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 2, 5

[2] Bilal Alsallakh, Narine Kokhlikyan, Vivek Miglani, Jun Yuan, and Orion Reblitz-Richardson. Mind the pad–cnns can develop blind spots. *arXiv preprint arXiv:2010.02178*, 2020. 2

[3] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018. 2

[4] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pages 528–539. PMLR, 2020. 1

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1, 2

[6] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Neural photo editing with introspective adversarial networks. *arXiv preprint arXiv:1609.07093*, 2016. 2

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2, 3

[8] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020. 8

[9] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 8

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[11] Patrick Esser, Robin Rombach, and Björn Ommer. A note on data biases in generative models. *arXiv preprint arXiv:2012.02516*, 2020. 1, 2

[12] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3012–3021, 2020. 2, 5

[13] 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. *arXiv preprint arXiv:1706.08500*, 2017. 2, 8

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020. 1, 2, 8

[15] Minyoung Huh, Richard Zhang, Jun-Yan Zhu, Sylvain Paris, and Aaron Hertzmann. Transforming and projecting images into class-conditional generative networks. In *European Conference on Computer Vision*, pages 17–34. Springer, 2020. 2, 5

[16] Md Amirul Islam, Sen Jia, and Neil DB Bruce. How much position information do convolutional neural networks encode? In *International Conference on Learning Representations*, 2020. 1, 2, 4

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 1

[18] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *arXiv preprint arXiv:2106.12423*, 2021. 2, 8

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2, 3, 4, 5, 8

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 1, 3, 4, 5

[21] Osman Semih Kayhan and Jan C van Gemert. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14274–14285, 2020. 1, 2

[22] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014. 1

[23] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020. 8

[24] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. Cocogan: Generation by parts via conditional coordinating. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4512–4521, 2019. 7

[25] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *arXiv preprint arXiv:1807.03247*, 2018. 7

[26] Marco Manfredi and Yu Wang. Shift equivariance in object detection. In *European Conference on Computer Vision*, pages 32–45. Springer, 2020. 2

[27] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2445, 2020. 2, 5

[28] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020. 2

[29] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *arXiv preprint arXiv:2102.09672*, 2021. 8

[30] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *arXiv preprint arXiv:1806.00035*, 2018. 2, 8

[31] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4570–4580, 2019. 2, 7

[32] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020. 3, 5

[33] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021. 2, 5, 6

[34] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2, 8

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 2, 3

[36] Rui Xu, Xintao Wang, Kai Chen, Bolei Zhou, and Chen Change Loy. Positional encoding as spatial inductive bias in gans. *arXiv preprint arXiv:2012.05217*, 2020. 2, 3, 4, 5

[37] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5485–5493, 2017. 2

[38] 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, 5, 8

[39] Richard Zhang. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning*, pages 7324–7334. PMLR, 2019. 2, 4

[40] Shengjia Zhao, Hongyu Ren, Arianna Yuan, Jiaming Song, Noah Goodman, and Stefano Ermon. Bias and generalization in deep generative models: An empirical study. *arXiv preprint arXiv:1811.03259*, 2018. 2

[41] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, pages 597–613. Springer, 2016. 2