Positional Encoding as Spatial Inductive Bias in GANs

CVPR 2021

2021. 7. 8

Dajin Han

Index

- Introduction
- Preliminaries
- Motivation
- Methods
- Experiments
- QA

Introduction

Key Contents

Key Contents

- Zero padding works as **implicit positional encoding**
  - Unintended design
  - Pros : Global structure information without center region
  - Cons : **Unbalanced spatial info**, Limits the diversity of a generative model

- Alternative **explicit positional encodings**
  - Cartesian Grid
  - Sinusoidal Positional Encoding

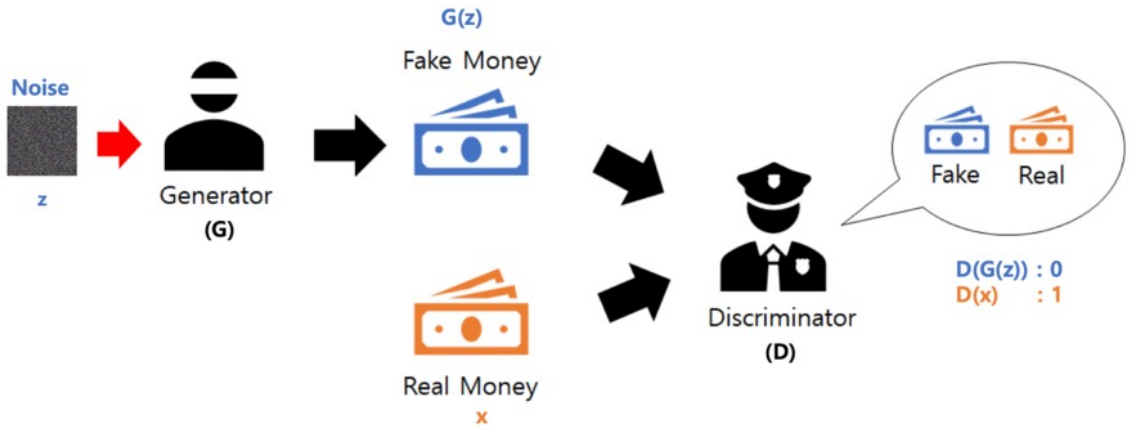
- New multi-scale training strategy
  - MS-PIE : **Scaling encoding**, not image

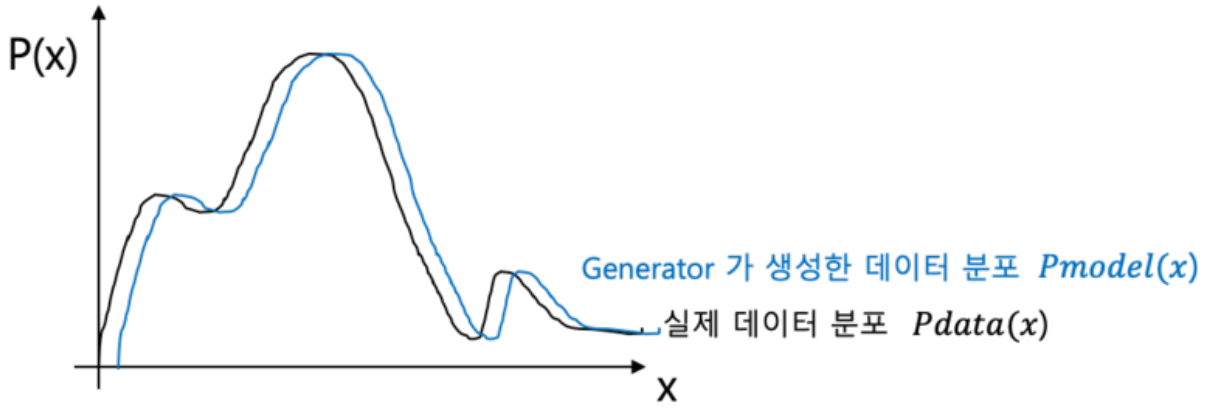
Preliminaries

GAN, Positional Encoding, Padding effect

GAN

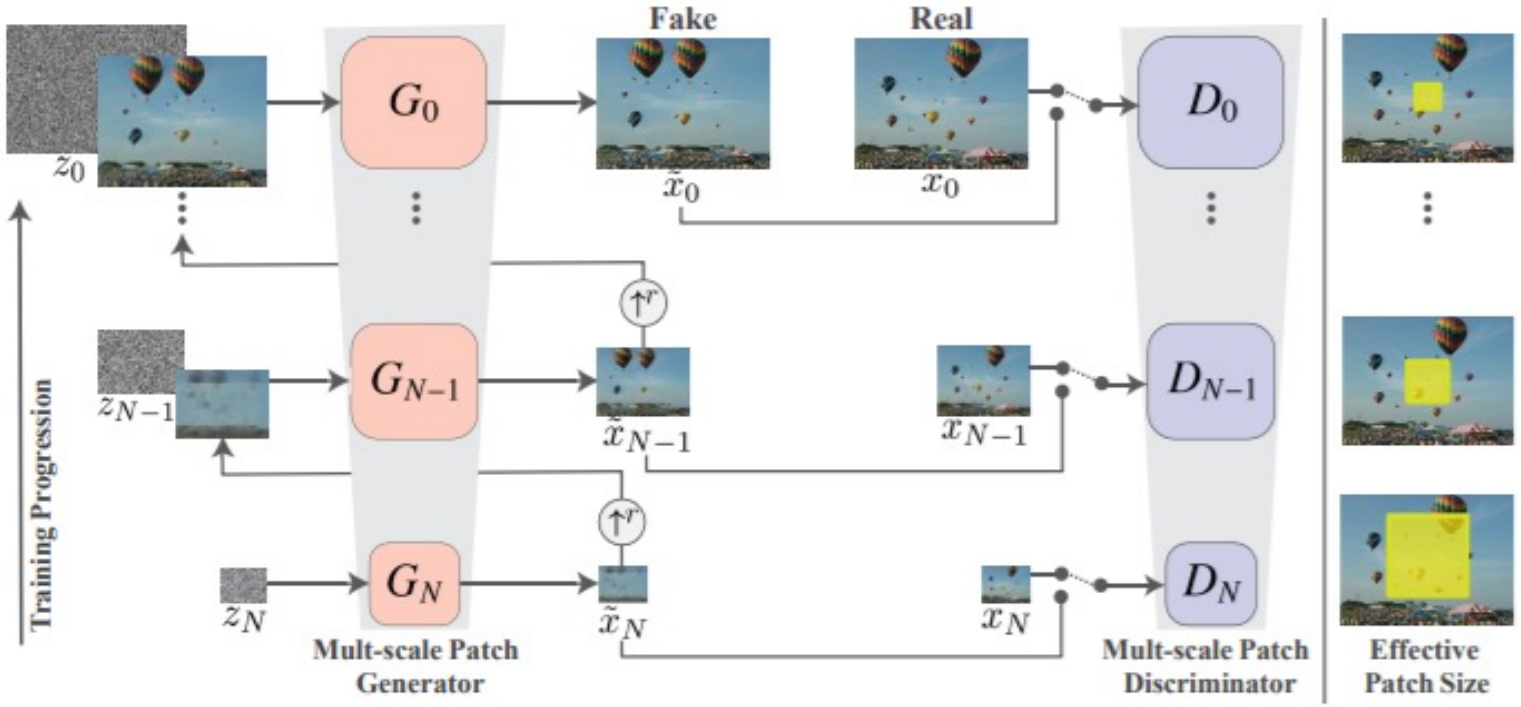
- Generative Adversarial Network
- Minmax problem
- Training data distribution of training dataset





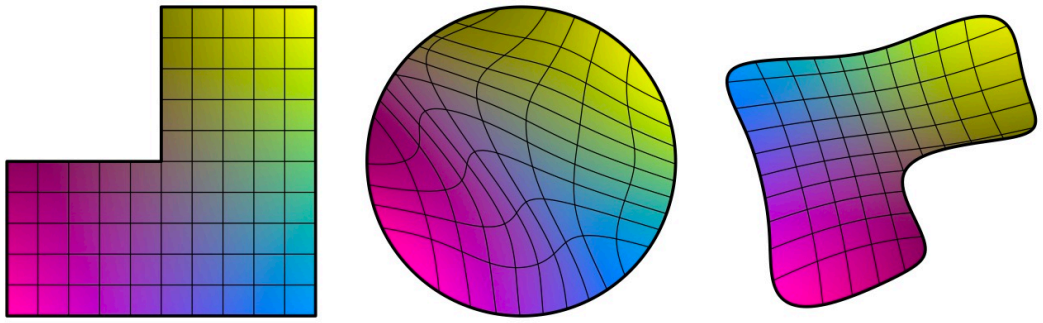
SinGAN

- Learning a Generative Model from a Single Natural Image
- Pyramid of fully convolutional GANs
- Coarse to Fine



StyleGAN

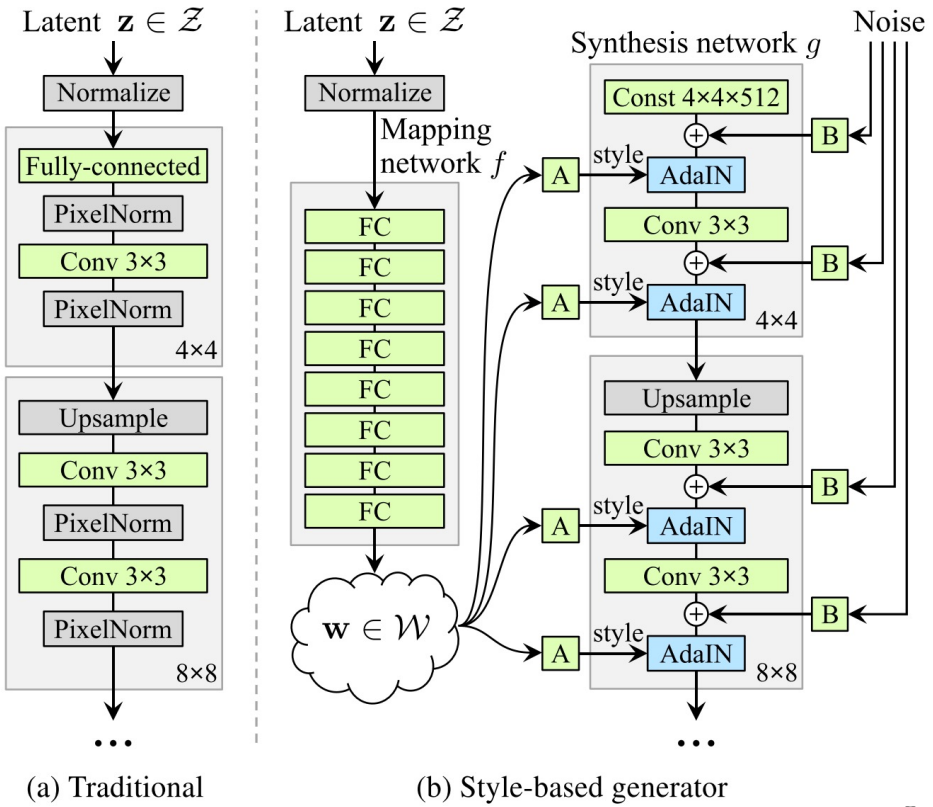
- A Style-Based Generator Architecture for GANs
- Intermediate latent vector
- More complex data distribution



(a) Distribution of features in training set

(b) Mapping from $\mathcal{Z}$ to features

(c) Mapping from $\mathcal{W}$ to features

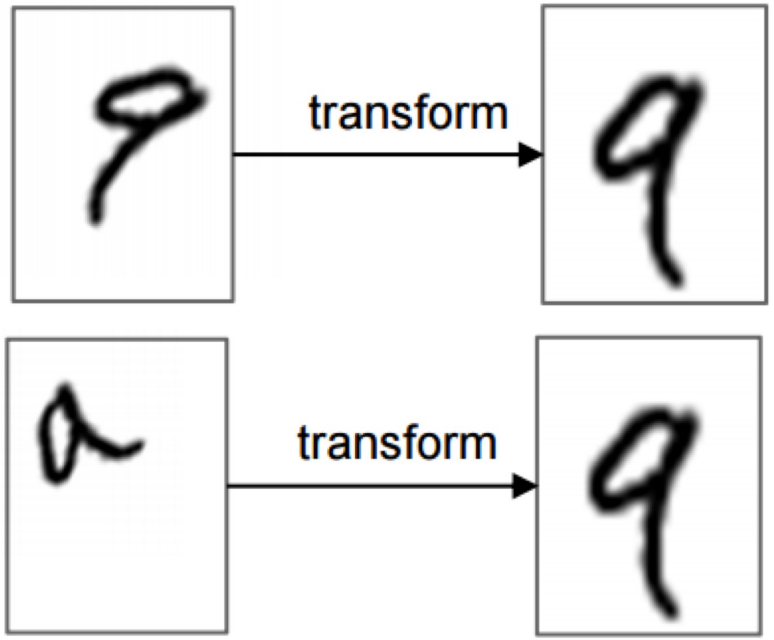


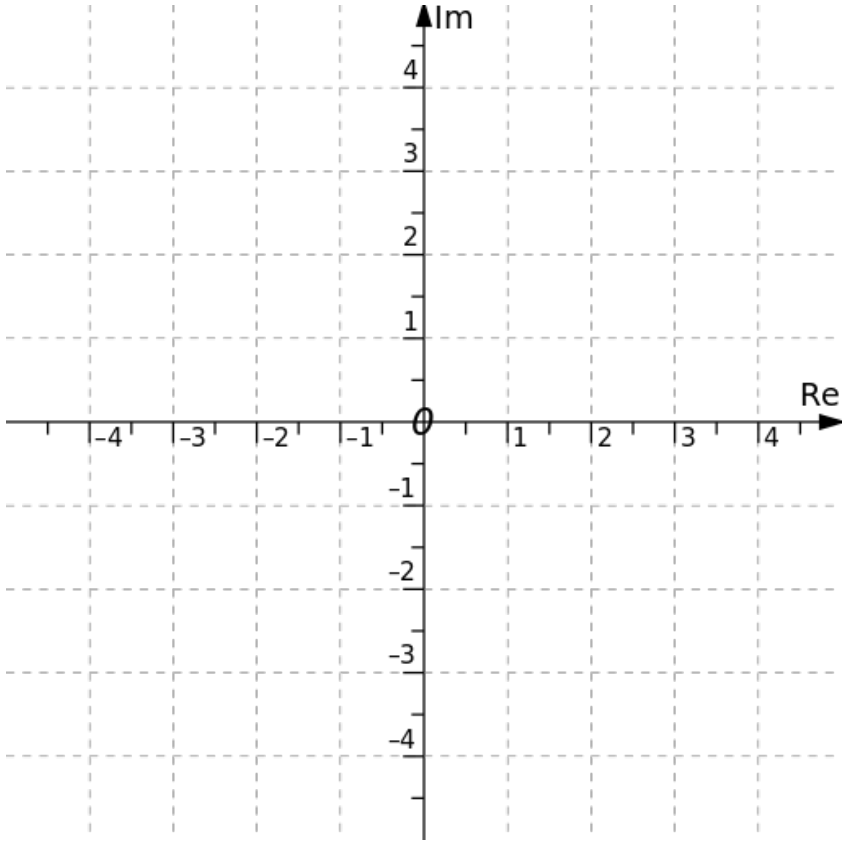
(a) Traditional

(b) Style-based generator

Cartesian Grid

- Spatial Transformer Network
- Spatially-Invariant



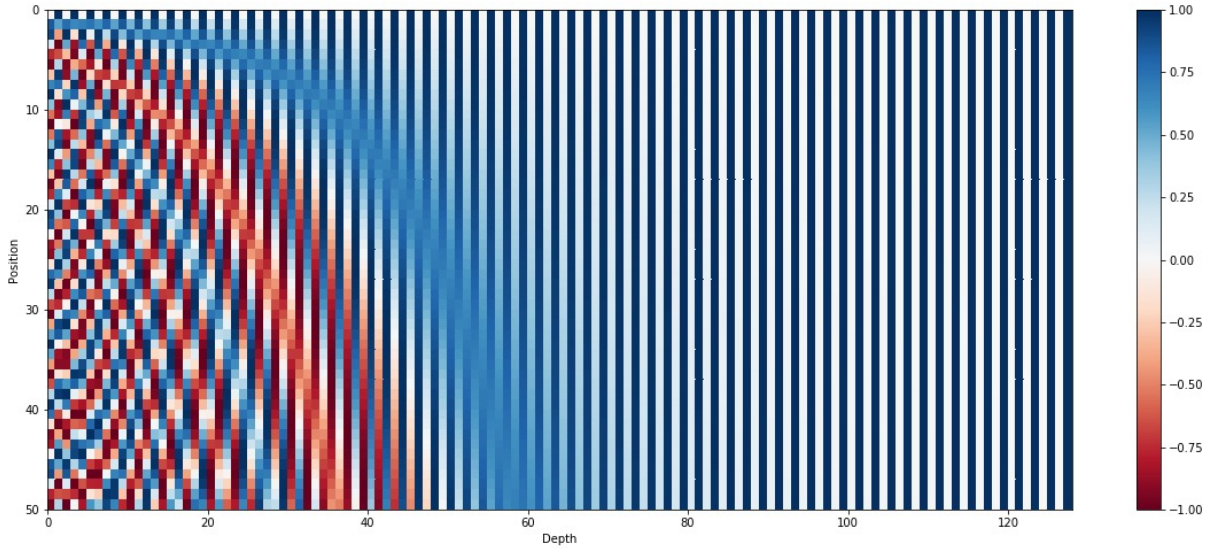


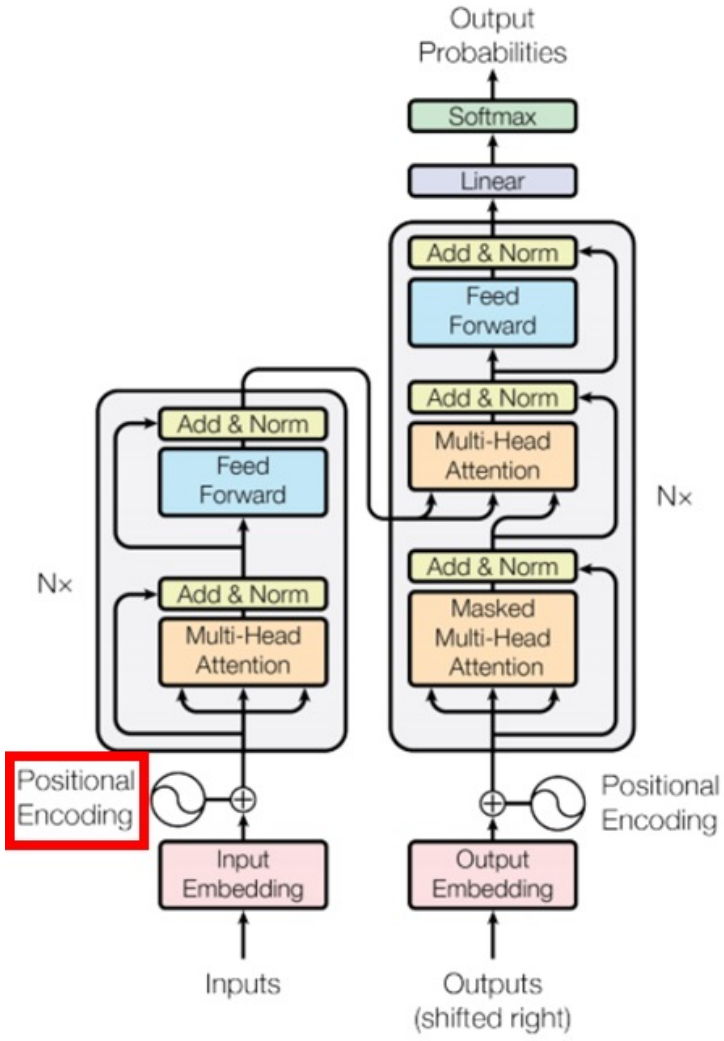
Sinusoidal Positional Encoding

- NLP – Transformer
- Sequence of the input

$$PE_{(\text{pos},\ 2\text{k})} = \sin\left(\text{pos}/10000^{2k/d}\right)$$

$$PE_{(\text{pos},2\text{k}+1)} = \cos\left(\text{pos}/10000^{2k/d}\right)$$





Motivation

Effect of zero padding, Statistical Property

GAN Results

- Small receptive field, but good performance
- Weak center region

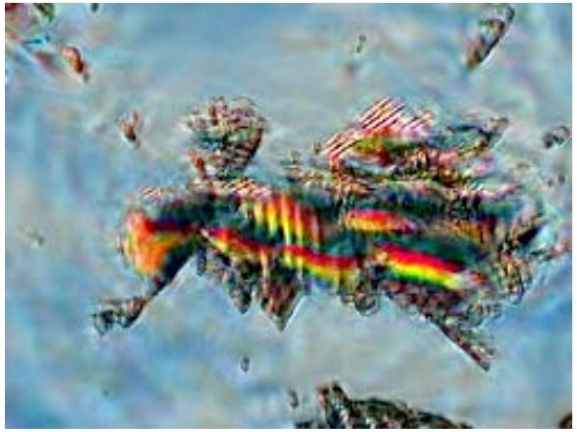
- Spatial bias spread from the **border to center** -> zero padding?

GAN without Padding

- Fail to capture **global structure**
- Zero padding works as **positional encoding**

Original Image

(a) SinGAN



(b) SinGAN w/o padding

Statistical property w/o padding

- $\mathbb{E}(y_i)$: Distributional property of each location i in the convolutional fea ture map
- $\mathrm{R}(y_i, y_j)$: Relationship between two spatial locations

Statistical property - $\mathbb{E}(y_i)$

- $\mathbb{E}(y_i)$: Distributional property of each location i in the convolutional fea ture map
- Keep spatially identical expectation value

$$\mathbb{E}(y_{\vec{\mathbf{i}}}^{(1)}) = \sum_k w_k^{(1)} \int_{-\infty}^{+\infty} x_k p(x_k) dx_k + b^{(1)}$$
$$= \sum_k w_k^{(1)} \mathbb{E}(x_k) + b^{(1)} = b^{(1)},$$

$$\mathbb{E}(y_{\vec{\mathbf{i}}}^{(2)}) = \sum_k w_k^{(2)} \int_{-\infty}^{+\infty} g(y_k^{(1)}) p(y_k^{(1)}) dy_k^{(1)} + b^{(2)}$$
$$= \sum_k w_k^{(2)} \cdot (\gamma \mathbb{C}_1 + \mathbb{C}_2) + b^{(2)},$$

$X_{\vec{\mathbf{i}}} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$

x_k : input feature
$p(\,)$: probability density function
w_k, b : parameters in the convolutional layer
g( ) : LeakyReLU

Statistical property - $\mathbf{R}(y_i, y_j)$

- R(y_i, y_j) : Relationship between two spatial locations
- Irrelevant to the absolute position vector

$$
\begin{aligned}
R(y_{\vec{\mathbf{i}}}, y_{\vec{\mathbf{j}}}) &= \mathbb{E}(y_{\vec{\mathbf{i}}} y_{\vec{\mathbf{j}}}) \\
&= \sum_{x_l \in X_{\vec{\mathbf{i}}} \cap X_{\vec{\mathbf{j}}}} w_{k_l} w_{t_l} \mathbb{E}(x_l^2) \\
&= R(\vec{\mathbf{i}} - \vec{\mathbf{j}}).
\end{aligned}
$$

$X_{\vec{\mathbf{i}}} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$

x_k : input feature
w_k, b : parameters in the convolutional layer

Statistical property w/o padding

- $\mathbb{E}(y_i)$: Distributional property of each location i in the convolutional fea ture map
- $\mathrm{R}(y_i, y_j)$: Relationship between two spatial locations

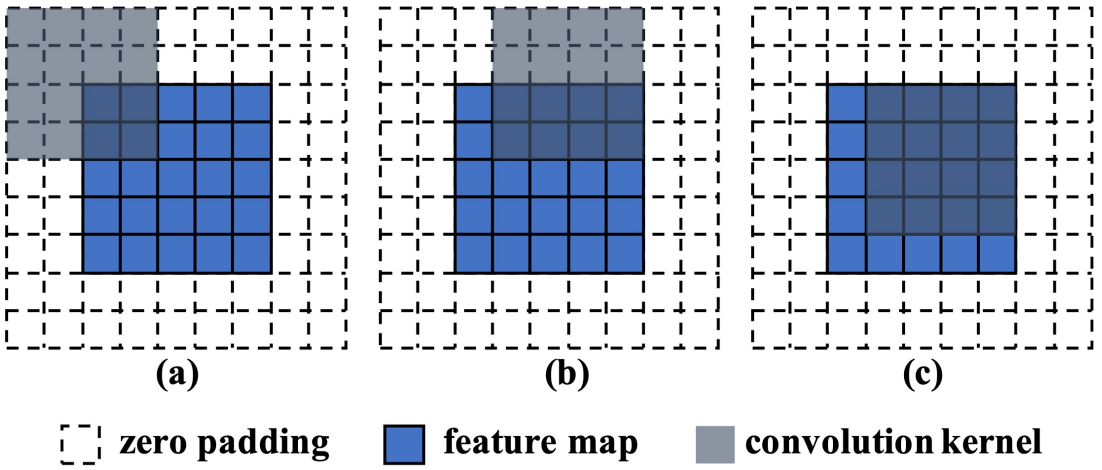
- **Weak stationary stochastic process (fail to capture global structure)**
- **Learn only small patch**

Statistical property with padding

- Unique encoding on corners

$$\mathbb{E}(y_{\vec{\mathbf{i}}}) = \sum_k w_k(\gamma\mathbb{C}_1 + \mathbb{C}_2)\mathbb{1}(x_k \notin Pad) + b,$$

$$R(y_{\vec{\mathbf{i}}}, y_{\vec{\mathbf{j}}}) = \sum_{x_l \in X_{\vec{\mathbf{i}}} \cap X_{\vec{\mathbf{j}}}} w_{k_l} w_{t_l} \mathbb{E}(x_l^2)\mathbb{1}(x_l \notin Pad),$$



Method

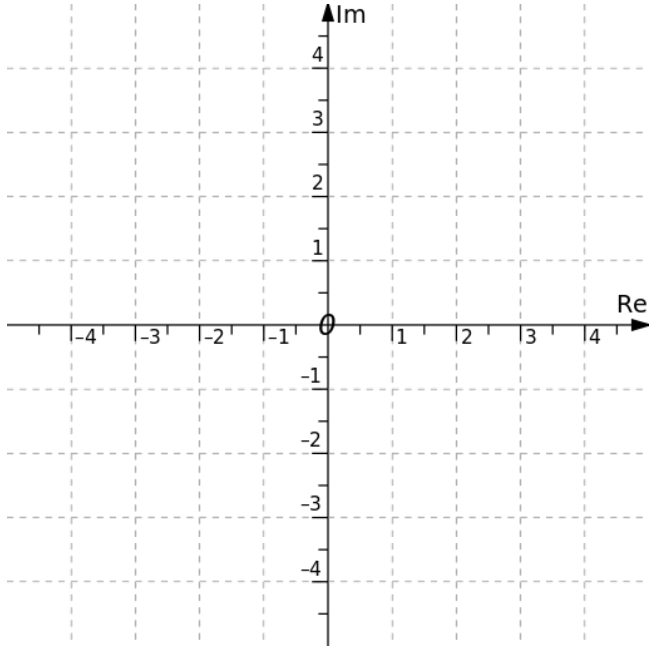
Learnable Constant Input, Cartesian Spatial Grid, Sinusoidal Positional Encoding, MS-PIE

Learnable Constant Input

- StyleGAN
- “Learnable” constant as the input of generator
- Unclear and lacks explicit prior on image space

Cartesian Spatial Grid

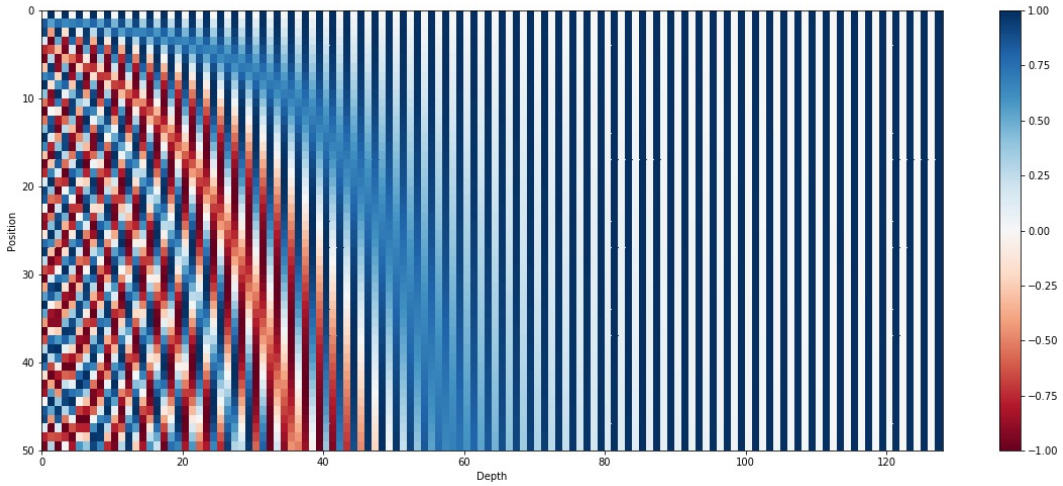
- Spatial Transformer Network (transformation)
- [-1, -1] for top left, [1, 1] for bottom right
- Larger input scale, Closer distances between adjacent points
- **Robust to align global structure**



Sinusoidal Positional Encoding

- NLP – Transformer
- Combination of sin and cos function
- **No spatial anchor**

$$[\underbrace{sin(\omega_0 i), cos(\omega_0 i), \cdots,}_{height\ dimension}\ \underbrace{sin(\omega_0 j), cos(\omega_0 j), \cdots}_{width\ dimension}],$$



Positional Encodings

Spatial Inductive Bias		Pad	Learnable Constant	Cartesian Grid	SPE
Spatial Anchors		✓		✓	
Identical Transform	SS			✓	✓
	MS				✓
Resize Mode	Interp		✓	✓	✓
	Expand				✓

Table 1: Summary of spatial inductive bias defined by different positional encodings. 'Identical Transform' with 'SS' means the transformation between locations is spatially identical at a single scale. The 'MS' row shows whether the transformation is scale-invariant. 'Interp' is the traditional interpolation while 'Expand' denotes the expansion in SPE.

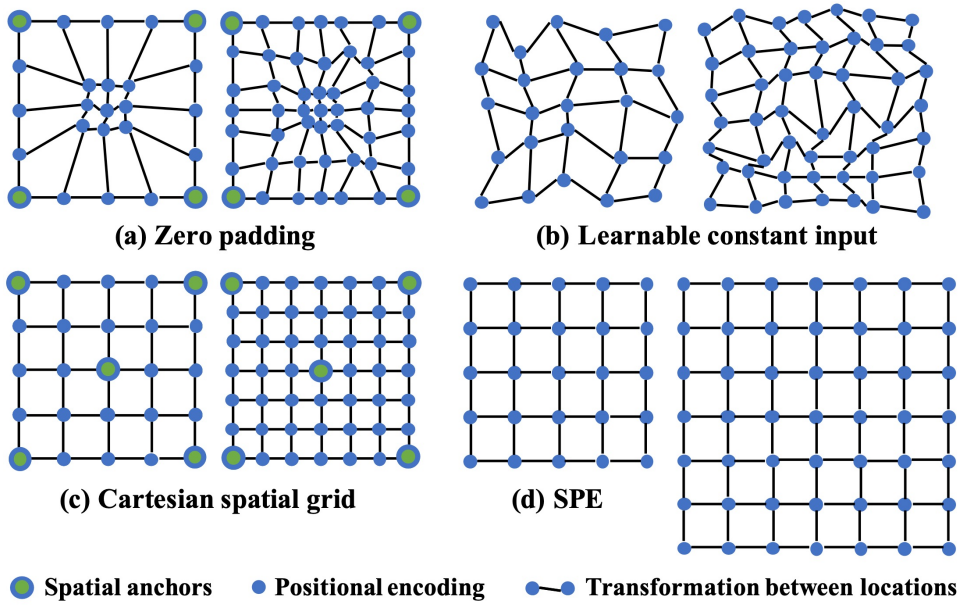


Figure 4: Illustration for the 2D spatial space defined by different positional encodings. For each encoding, we present two spatial spaces at 5×5 and 7×7 scale. (d) shows how SPE naturally expands its space with consistent transformation.

MS-PIE

- Multi-Scale training with Positional Encoding
- Two methods
  - with zero padding
  - w/o zero padding

MS-PIE – with zero padding

- Standard Model with zero padding
  - Capture global structure well
  - Weak representation on central region

- SPE (Spatial Positional Encoding)
  - Much **precise spatial inductive bias** over **dynamic input space**

MS-PIE – w/o zero padding

- Standard Model without zero padding
  - Fail to capture global structure
  - Weak representation on central region

- CSG (Cartesian Spatial Grid)
  - Fixed anchors algin global structure

Experiments

Effect of zero padding, StyleGAN

Experiments – zero padding

Training configuration	SWD ($\times 10^3$) ↓			
	128	64	32	Avg.
(a) PGGAN	**3.162**	**4.285**	**5.000**	**4.149**
(b) + Remove padding	11.169	6.945	7.488	8.534
(c) + SPE, w/o padding	4.555	6.164	6.365	5.694

Training configuration	SWD ($\times 10^3$)↓			
	128	64	32	Avg.
(a) DCGAN w/ conv	**11.82**	17.30	28.10	19.07
(b) + No padding	22.21	27.26	44.24	31.24
(c) + SPE, w/o padding	14.75	**16.52**	**22.53**	**17.93**

Experiments - StyleGAN

Table 2: Results based on 256^2 StyleGAN2 with a channel multiplier of two [22] in FFHQ dataset [21, 24]. The Fréchet inception distance (FID) are reported on two best snapshots before the discriminator has been shown with 10M and 20M images, respectively. The precision and recall are reported on the training snapshot with best FID. ↑ indicates higher is better, and ↓ indicates lower is better.

Training configuration	FID@256↓		Precision	Recall
	10M	20M	(%)↑	(%)↑
(a) StyleGAN2-C2-256	6.32	**5.62**	**76.85**	50.41
(b) Deconv → Up-Conv	**5.79**	5.68	76.78	50.46
(c) + Remove padding	6.45	6.13	73.71	51.73
(d) + Cartesian grid	6,31	6.07	73.01	**52.90**
(e) + SPE	6.40	5.86	72.94	**52.87**

Experiments – MS-PIE

Table 5: Main results for our MS-PIE with 256^2 StyleGAN2 in the FFHQ dataset. The precision and recall are calculated at the same scale as the Fréchet inception distance (FID). 'C2' indicates the channel multiplier in the generator is two.

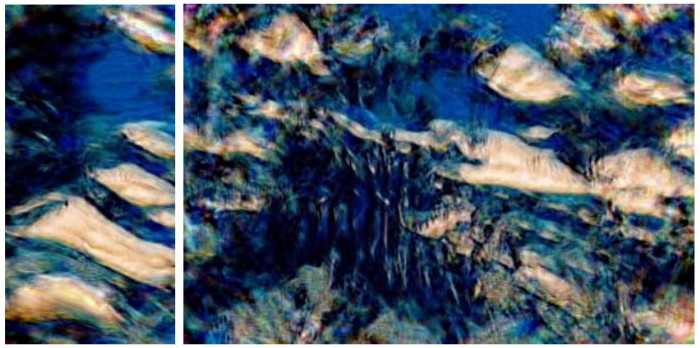
Training configuration		Resize	FID@512↓ 20M	FID@512↓ 25M	Precision (%)↑	Recall (%)↑	FID@256↓ 20M	FID@256↓ 25M	Precision (%)↑	Recall (%)↑
(a) StyleGAN2-C2-256		—	—	—	—	—-	5.62	5.56	75.92	51.24
(b) StyleGAN2-C2-512		—	3.47	3.41	75.88	54.61	5.00	4.91	75.65	54.58
MS-PIE w/ padding	(c) Leanable constant input	Interp	3.46	3.35	73.84	55.77	4.82	4.50	72.75	55.42
	(d) Cartesian spatial grid	Interp	3.59	3.50	73.28	56.16	**4.74**	4.71	73.34	55.07
	(e) SPE-interp	Interp	3.41	3.15	**74.13**	56.88	4.99	4.73	73.28	**56.93**
	(f) SPE-expand	Expand	**3.31**	**2.93**	73.51	**57.32**	4.79	**4.27**	**73.48**	55.69
	(g) SPE-expand-C1	Expand	3.65	3.40	73.05	56.45	5.54	4.83	73.59	54.29
MS-PIE w/o padding	(h) Learnable constant input	Interp	4.99	4.01	72.81	54.35	5.67	5.11	72.37	55.52
	(i) Cartesian spatial grid	Interp	**3.96**	**3.76**	**73.26**	**54.71**	**5.30**	**5.09**	70.74	56.06
	(j) SPE-interp	Interp	4.80	4.23	73.11	54.63	5.89	5.38	71.21	**56.21**
	(k) SPE-expand	Expand	4.46	4.17	73.05	51.07	6.08	5.59	**72.65**	49.74

Original Image **(a) SinGAN** **(b) SinGAN w/ Cartesian grid** **(c) SinGAN w/ SPE**

Results

Original Image

(c) SinGAN w/ padding



(d) SinGAN w/o padding

(e) SinGAN w/ Cartesian Grid

(f) SinGAN w/ SPE

QA