



# Alias-Free Generative Adversarial Networks

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, Timo Aila

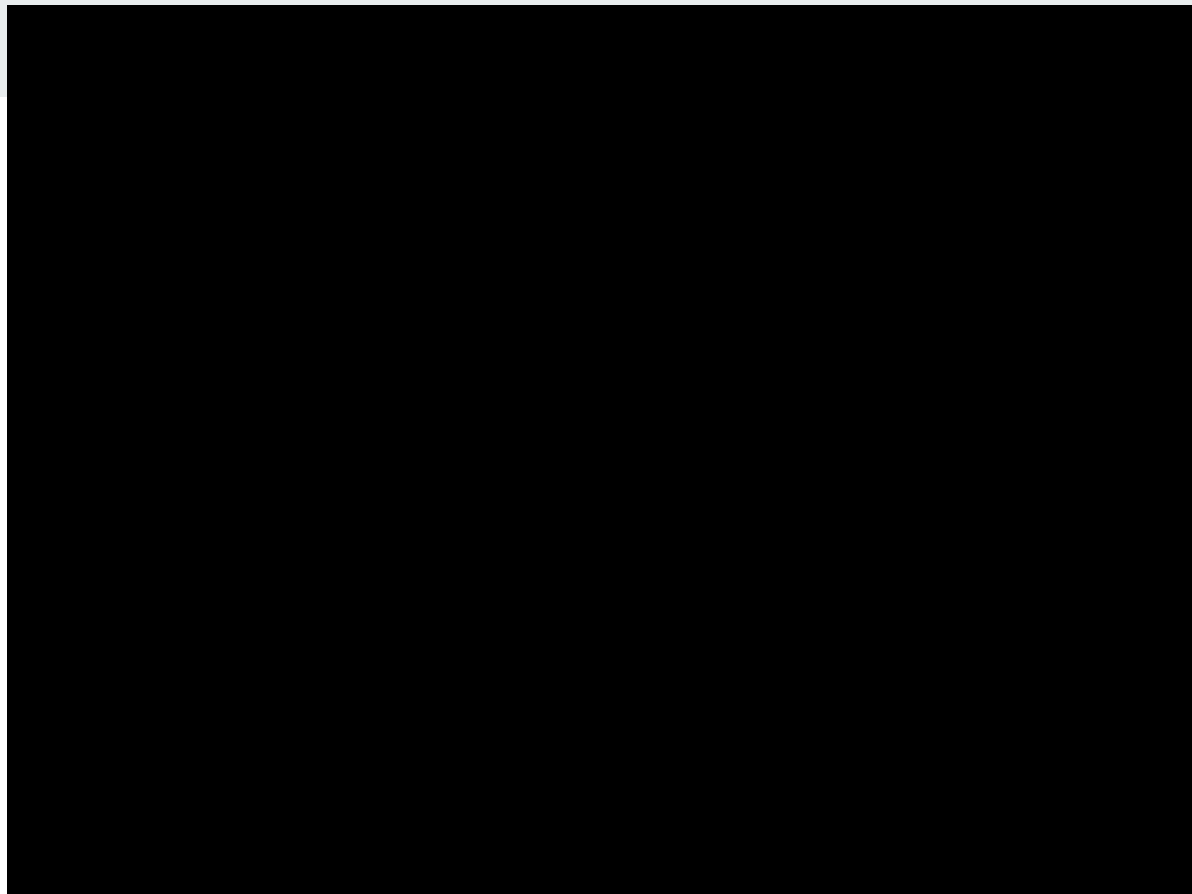
Presenter: Yixing Wang

# StyleGAN3: a translation/rotation equivariant model

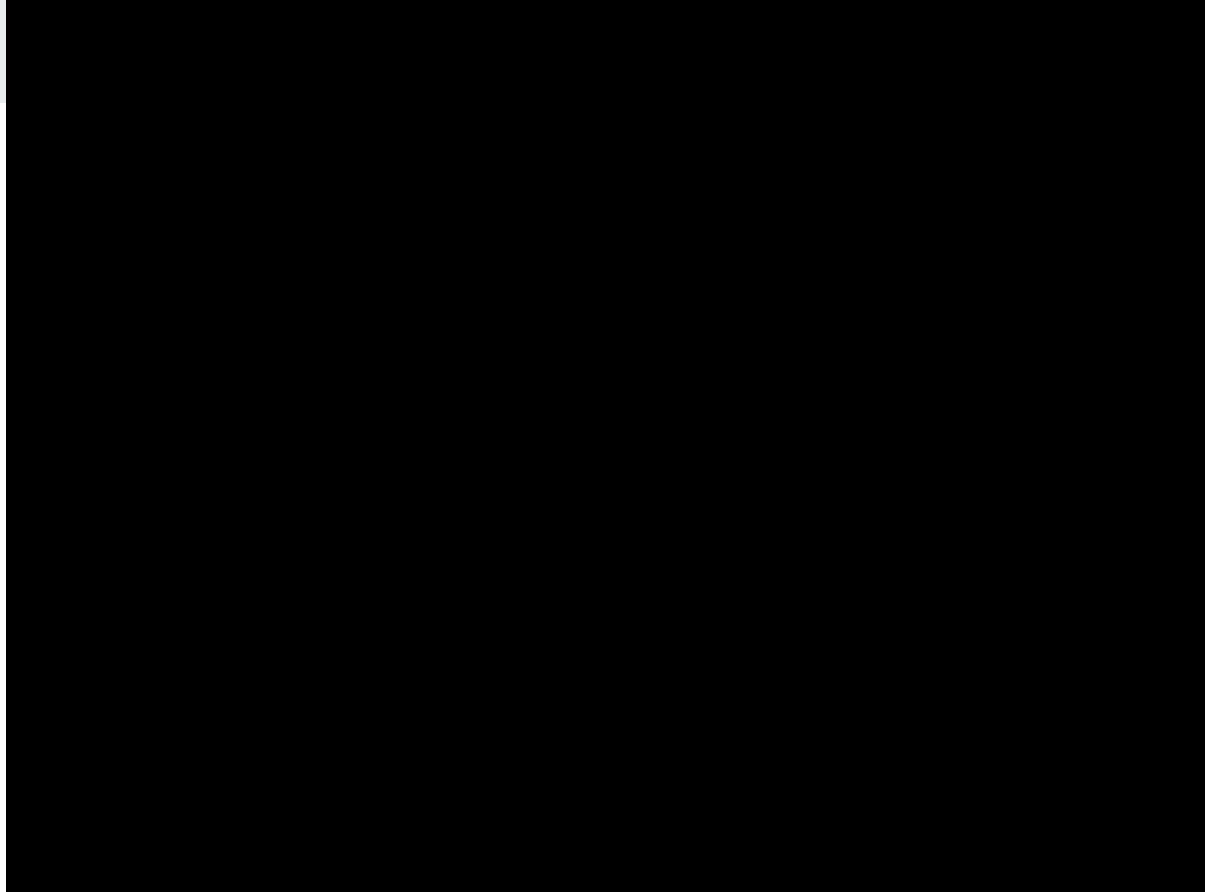


Although current SOTA GANs can generate realistic images, they are not transformation(translation/rotation) equivariant:


- Moving a head causes the nose to move, which in turn moves the skin on the nose.
- GANs: details/textures seem to stick to image coordinates instead of surfaces of parent objects.



“Texture sticking” issue: The looping video shows small random walks around a central point in the latent space. Details (hairs, wrinkles, etc.) from StyleGAN2 (left) appear to be glued to the screen coordinates while the face moves under it.



By doing latent space interpolation, extract a vertical segment of pixels from each generated image and stack them horizontally. The desired result is hairs moving in animation, but styleGAN2 creates horizontal streaks.



The reason why it happens is due to careless signal processing during generator upsampling.

- Consider a 2D feature map  $z$ , and a common operation  $f$  in NN (convolution, nonlinearity, etc), and a spatial transformation  $t$ . We don't want  $f$  to influence  $t$  when  $t$  is applied to following layers. i.e. we want  $f \circ t(z) = t \circ f(z)$ .
- Then  $f$  is an equivariant operation.
- E.g. A translation is applied to a low-resolution feature map so the coarse features (like face) is getting translated. After a series of operation in the network (convolution, upsampling, nonlinear activation, etc), the same translation gets applied to high-resolution feature maps s.t. fine features (like eyes/nose) can move with the face consistently.



## Formally...

- Nyquist-Shannon sampling theorem: a regularly sampled signal cannot represent any continuous signal containing frequencies between zero and half of the sample rate.
- Image as signals: sampling rate  $s$  = image pixel width
- Suppose  $Z$ : discrete pixel map,  $z$ : underlying continuous image intensity signal space  
 $F$ : equivariant operation on  $Z$ ,  $f$ : corresponding operation on continuous space  $z$ .
- An operation in discrete domain can be seen to perform a corresponding operation on the continuous domain.

$$\mathbf{f}(z) = \phi_{s'} * \mathbf{F}(\mathbb{I}_s \odot z),$$

$$\mathbf{F}(Z) = \mathbb{I}_{s'} \odot \mathbf{f}(\phi_s * Z),$$

Where  $\phi_s$  is interpolation filter that recovers continuous  $f(z)$  from discrete  $F[Z]$ ,

$$\phi_s(\mathbf{x}) = \text{sinc}(sx_0) \cdot \text{sinc}(sx_1)$$

$\mathbb{I}_s$  is the sampling grid.

\* denotes continuous convolution and  $\odot$  denotes pointwise multiplication.



## Formally...

- Nyquist–Shannon sampling theorem: a regularly sampled signal cannot represent any continuous signal containing frequencies between zero and half of the sample rate.
- Suppose  $s$ : sampling rate of input feature map,  $s'$ : sampling rate of output feature map. **Then an equivariant operation must not generate frequency content above the output bandlimit of  $s'/2$ .**



## Check common operations in neural networks

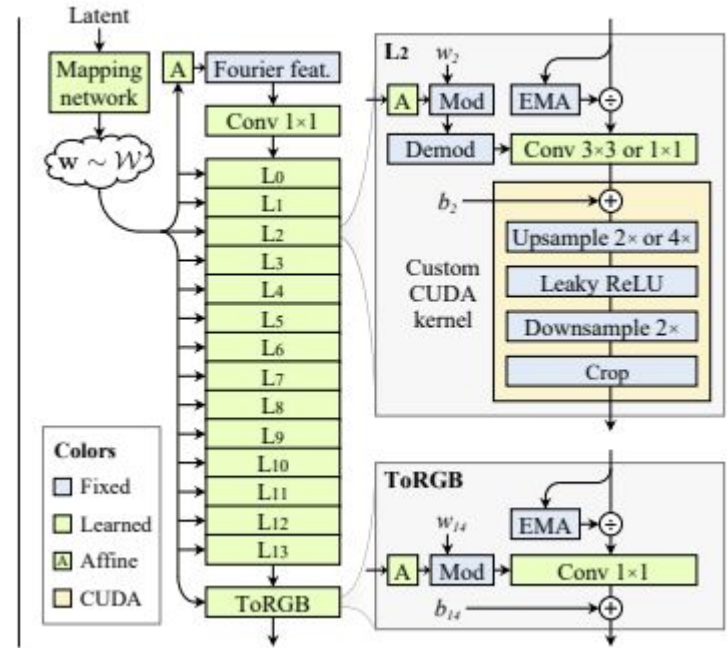
- **Convolution:**  $\mathbf{F}_{\text{conv}}(Z) = K * Z$   
$$\mathbf{f}_{\text{conv}}(z) = \phi_s * (K * (\mathbb{I}\mathbb{I}_s \odot z)) = K * (\phi_s * (\mathbb{I}\mathbb{I}_s \odot z)) = K * z$$
  - Introduces no new frequencies, bandlimit requirements fulfilled.
  - But for rotation equivariance, K needs to be radially symmetric.
- **Upsampling:**  $\mathbf{f}_{\text{up}}(z) = z$ 
  - Just increases sampling rate in discrete domain, does nothing / identity mapping in continuous domain, bandlimit requirements fulfilled.
- **Downsampling:** must low-pass filter  $z$  to remove frequencies above the output bandlimit.
- **Nonlinearity:** activation function like ReLU can introduce arbitrarily high frequencies.
  - Solution: low-pass filtering.
  - But low-pass filtering needs to be operated in a continuous space, we can approximate it by first upsampling the signal, applying nonlinearity, and then downsampling.



# Redesign every operation in styleGAN2

## StyleGAN2:

- A mapping network transforms a latent vector to a latent code  $w \sim \mathcal{W}$ .
- A synthesis network  $G$  starts from a learned constant  $Z_0$  and applies a sequence of  $N$  layers (convolutions, nonlinearities, upsampling, per-pixel noise) to produce output image  $G(Z_0; w)$ , where  $w$  controls the modulation of the convolution kernels.

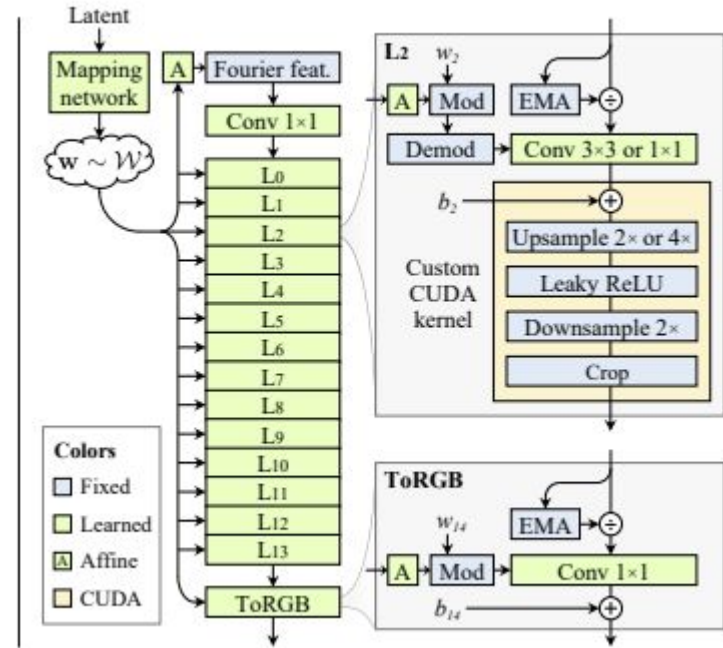


StyleGAN3

# Redesign every operation in styleGAN2

## StyleGAN3:

- Replace the learned input constant  $Z_0$  with Fourier features -> advantage of naturally defining a spatially infinite map.
- Maintaining a fixed-size margin around the target canvas, since theorem assumes an infinite continuous space.
- Replace 2x upsampling filter with an ideal low-pass filter, a windowed *sinc* filter with a large window size  $n=6$ . Filter cutoff =  $s/2$ .
- Wrap each leakyReLU with an upsampling and downsampling layer.
- For rotation equivariance, replace the 3x3 convolution kernel with 1x1 kernel.



StyleGAN3

# Results

Dataset	Config	FID ↓	EQ-T ↑	EQ-R ↑
FFHQ-U 70000 img, 1024 <sup>2</sup> Train from scratch	StyleGAN2	3.79	15.89	10.79
	StyleGAN3-T (ours)	3.67	61.69	13.95
	StyleGAN3-R (ours)	<b>3.66</b>	<b>64.78</b>	<b>47.64</b>
FFHQ 70000 img, 1024 <sup>2</sup> Train from scratch	StyleGAN2	<b>2.70</b>	13.58	10.22
	StyleGAN3-T (ours)	2.79	61.21	13.82
	StyleGAN3-R (ours)	3.07	<b>64.76</b>	<b>46.62</b>
METFACES-U 1336 img, 1024 <sup>2</sup> ADA, from FFHQ-U	StyleGAN2	18.98	18.77	13.19
	StyleGAN3-T (ours)	<b>18.75</b>	64.11	16.63
	StyleGAN3-R (ours)	<b>18.75</b>	<b>66.34</b>	<b>48.57</b>
METFACES 1336 img, 1024 <sup>2</sup> ADA, from FFHQ	StyleGAN2	15.22	16.39	12.89
	StyleGAN3-T (ours)	<b>15.11</b>	<b>65.23</b>	16.82
	StyleGAN3-R (ours)	15.33	64.86	<b>46.81</b>
AFHQV2 15803 img, 512 <sup>2</sup> ADA, from scratch	StyleGAN2	4.62	13.83	11.50
	StyleGAN3-T (ours)	<b>4.04</b>	60.15	13.51
	StyleGAN3-R (ours)	4.40	<b>64.89</b>	<b>40.34</b>
BEACHES 20155 img, 512 <sup>2</sup> ADA, from scratch	StyleGAN2	5.03	15.73	12.69
	StyleGAN3-T (ours)	<b>4.32</b>	59.33	15.88
	StyleGAN3-R (ours)	4.57	<b>63.66</b>	<b>37.42</b>

EQ-T and EQ-R measures the model's equivariance to translation and rotation, which is peak signal-to-noise ratio (PSNR) in decibels (dB) between two assets of images.



## Pros

- Generate images with details moving with coarse features when doing interpolation in latent space (equivariance to translation and rotation in latent space)
- Potential application toward GAN-based video/animation generation.

## Cons

- Computationally hard and expensive.
- It's difficult to train it on image datasets where aliasing is a feature of the aesthetic (e.g. black-and-white cartoons/low-quality jpegs)

