Recap.

判别难度：如何有效地区分两个分布？

- 怎么判别人脸是真实的？
  - 这并不容易…
  - 判别器无法有效为生成器提供优化信息

真实！

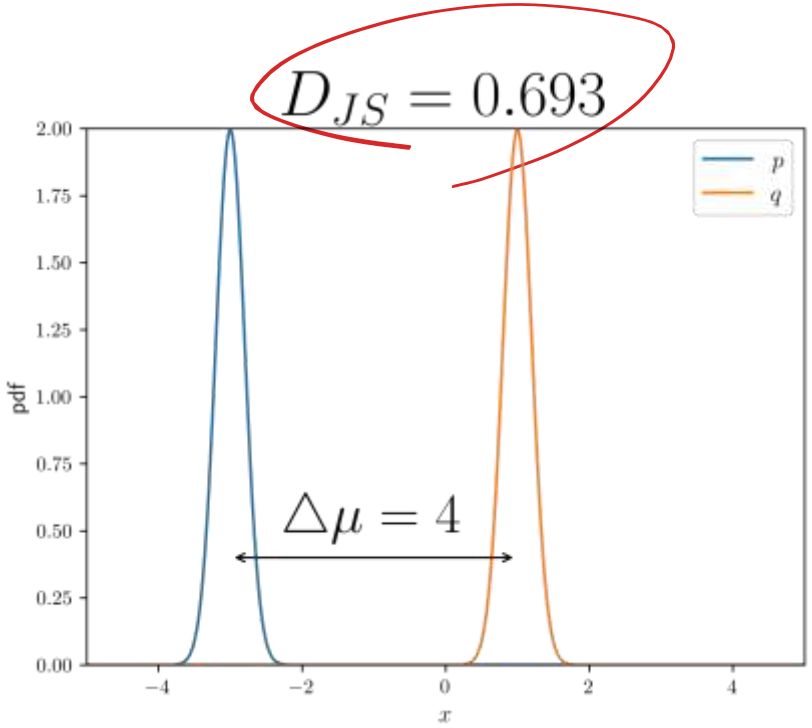
生成！

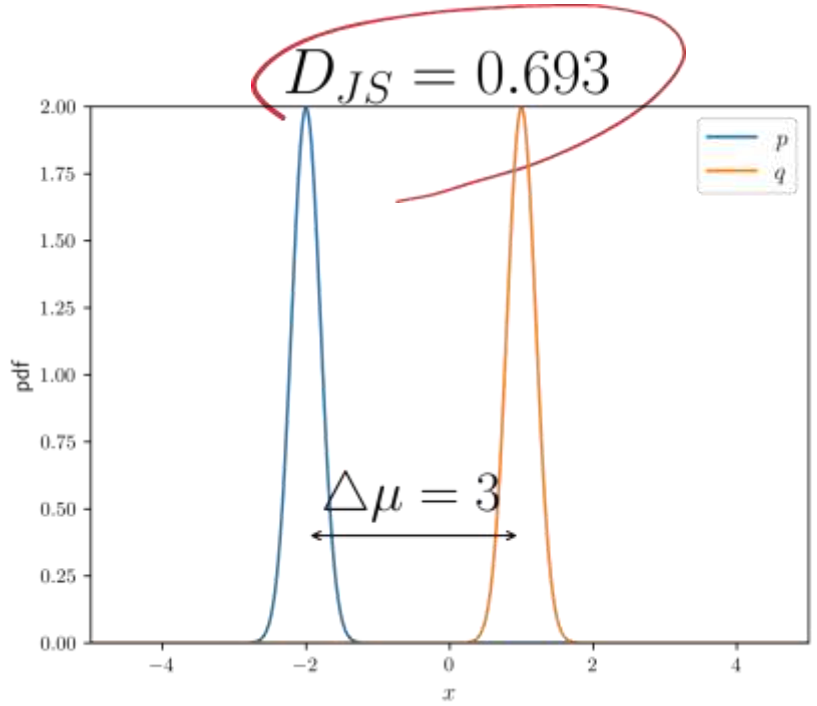
Mode Collapse

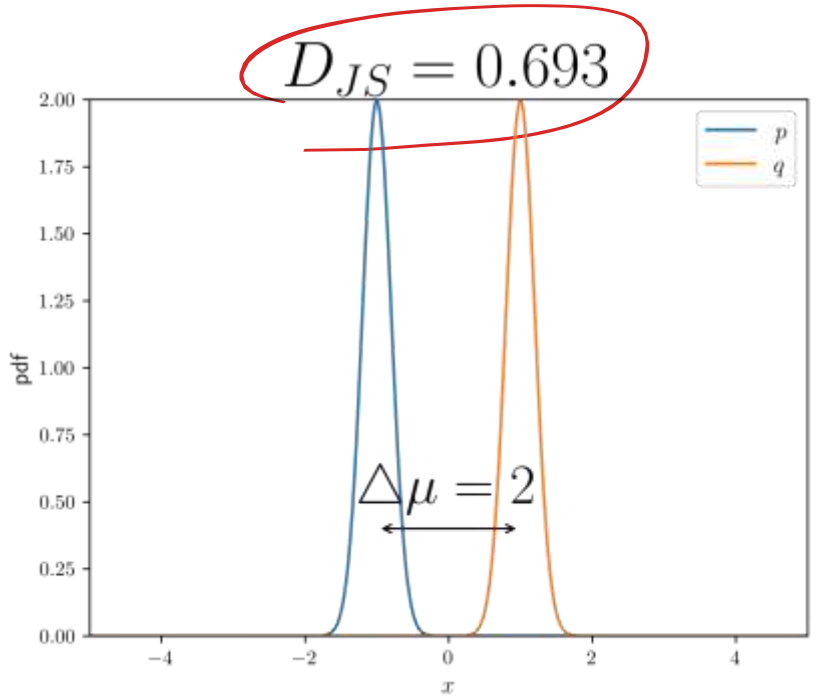
https://arxiv.org/pdf/1701.07875.pdf

Problems of D_{JS}

If p and q don't overlap, D_{JS} is a constant (log2), i.e., no gradient



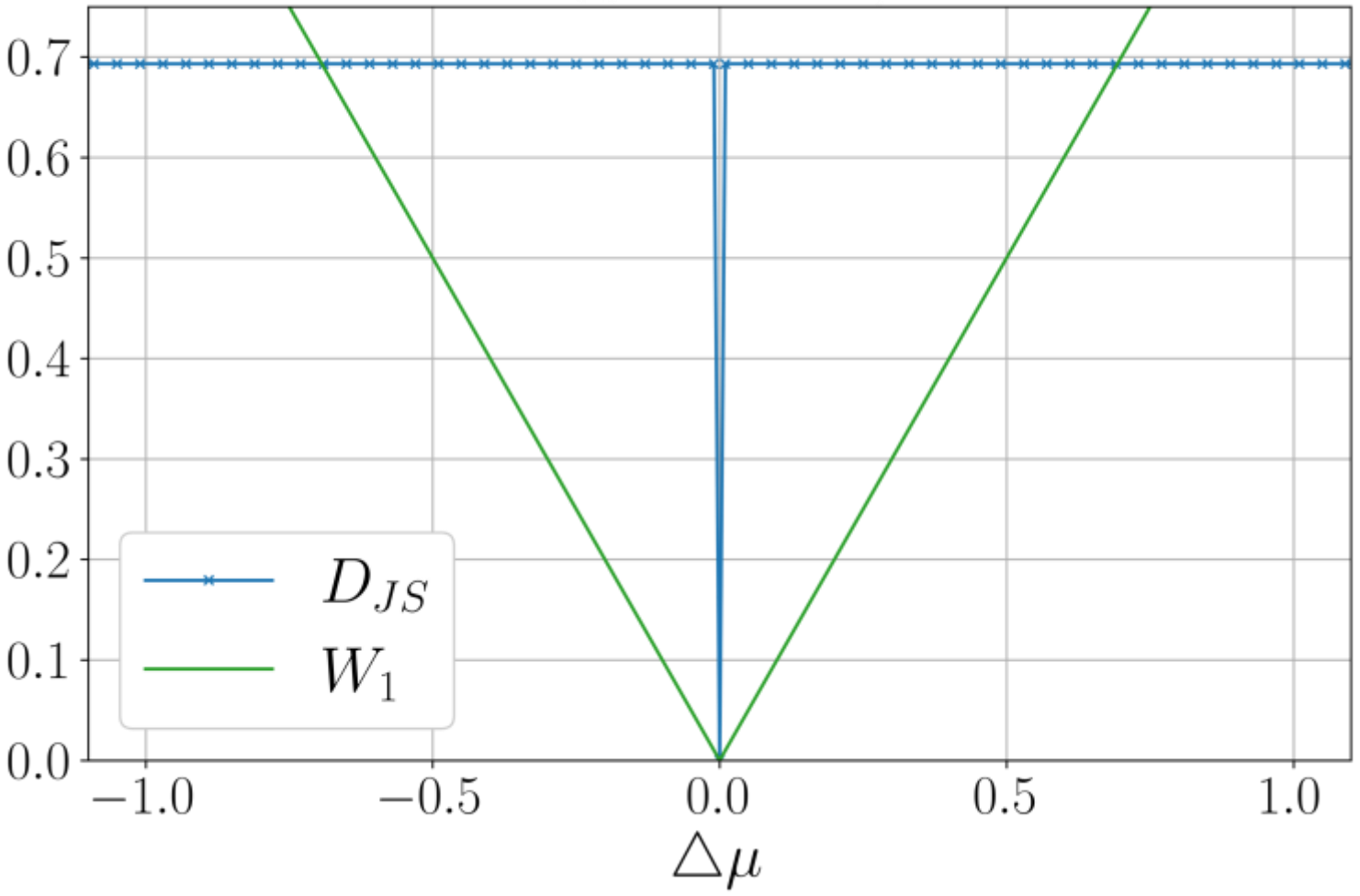


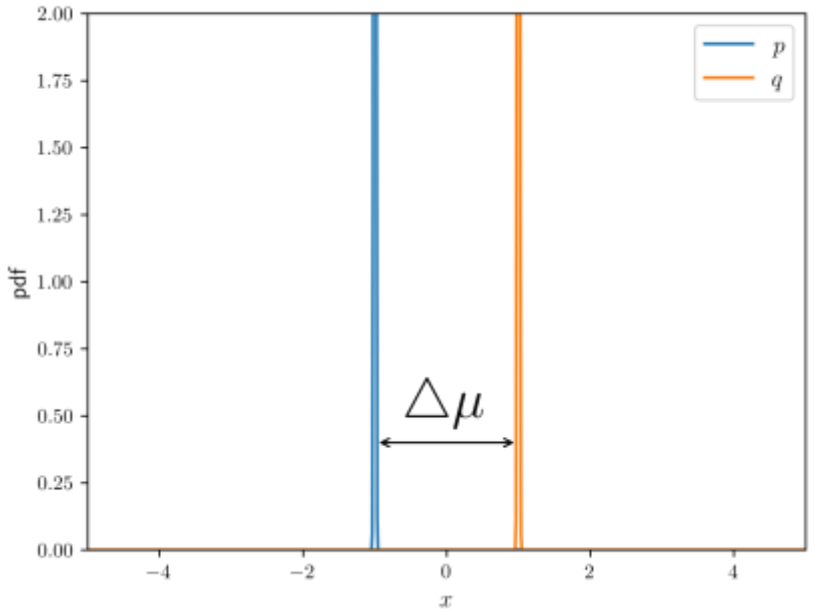


Wasserstein Distance

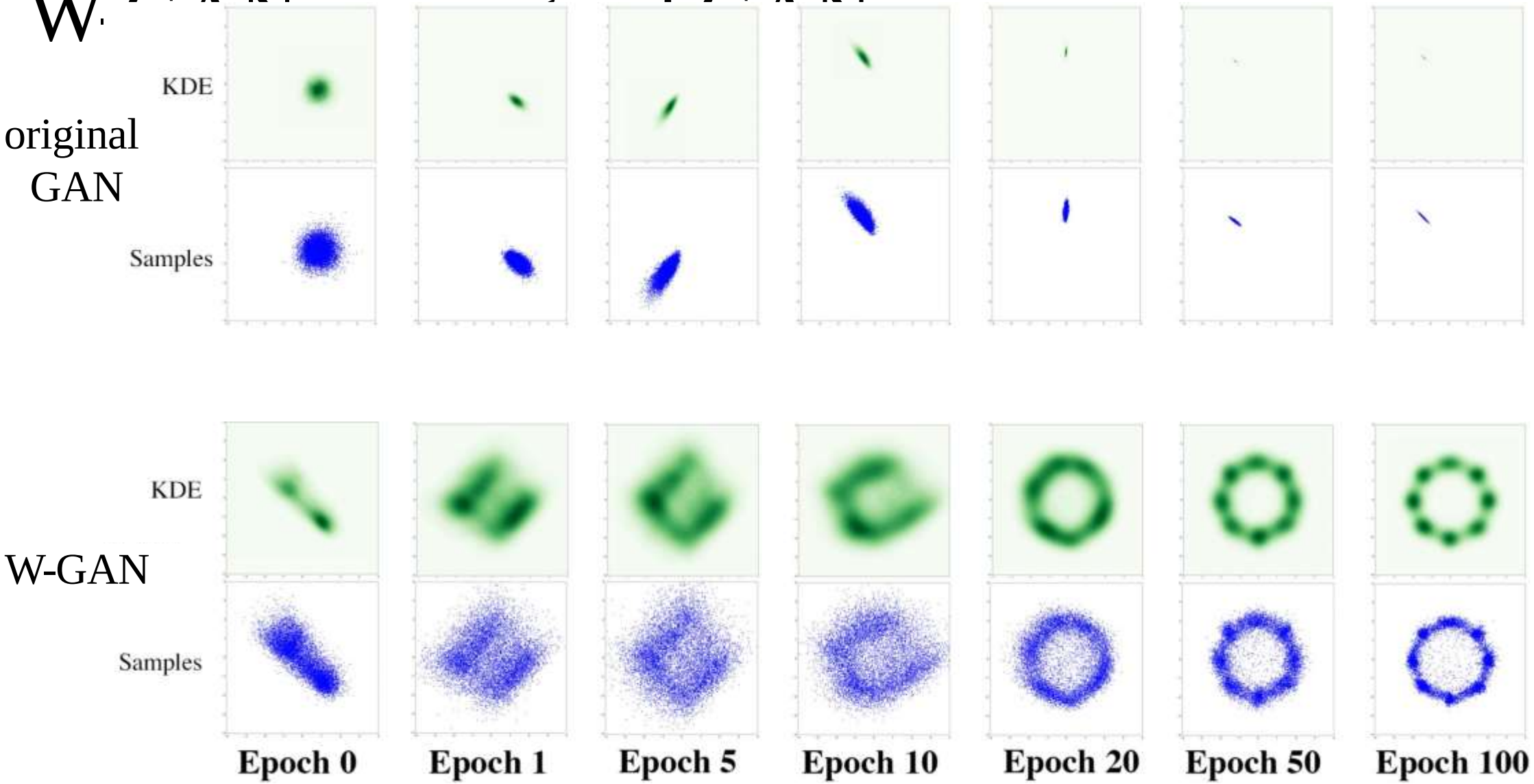
- when p and q are delta functions:

$$W_1(p, q) = |\mu_p - \mu_q|$$





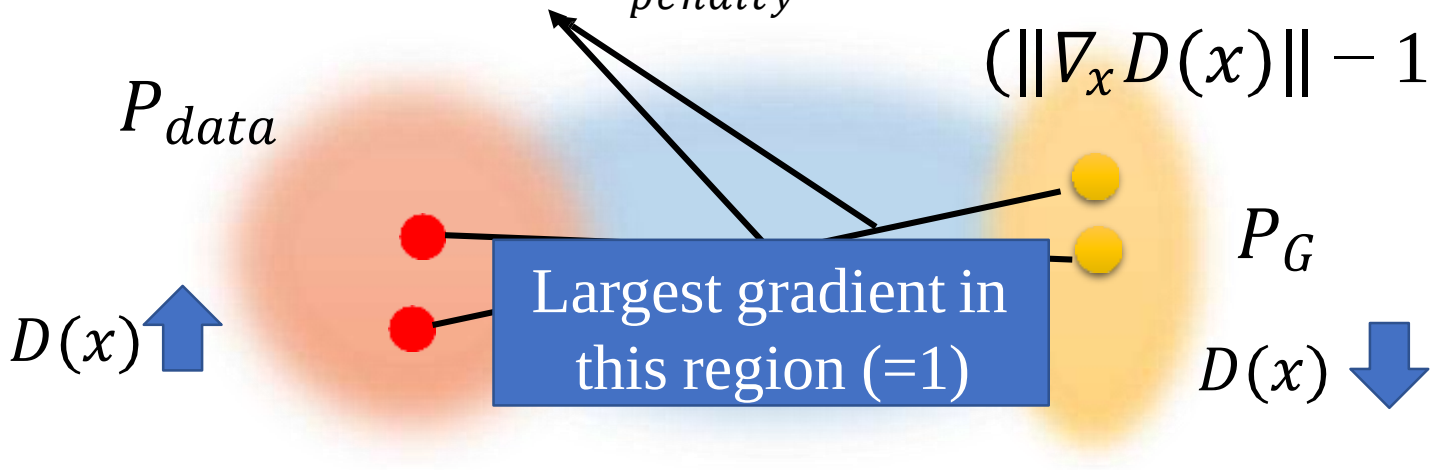
W- GAN vs original GAN



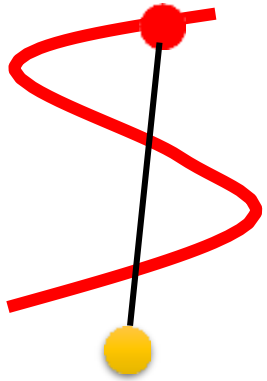
Improved WGAN (WGAN-GP)

$$V(G, D) \approx \max_{D}\{E_{x\sim P_{data}}[D(x)] - E_{x\sim P_G}[D(x)] - \lambda E_{x\sim P_{penalty}}[\cancel{max(0, \|\nabla_x D(x)\| - 1)}]\}$$

$$(\|\nabla_x D(x)\| - 1)^2$$



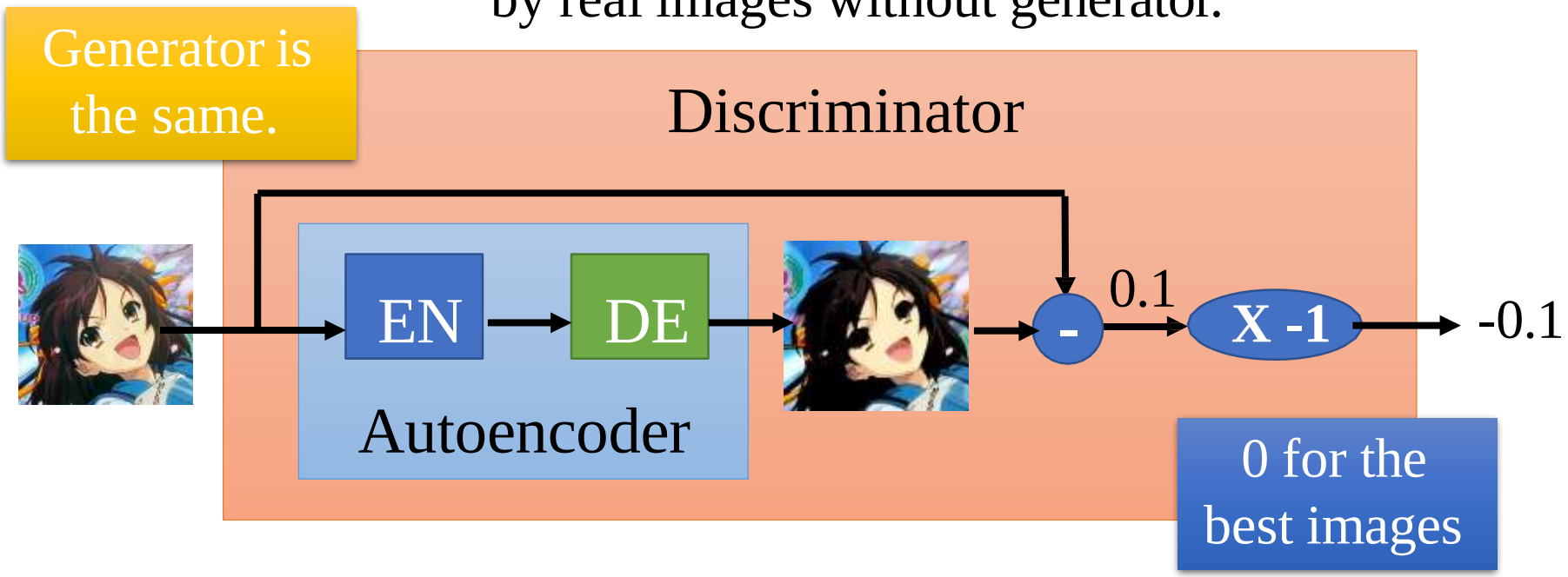
"Simply penalizing overly large gradients also works in theory, but experimentally we found that this approach converged faster and to better optima."



[Junbo Zhao, et al., arXiv, 2016]

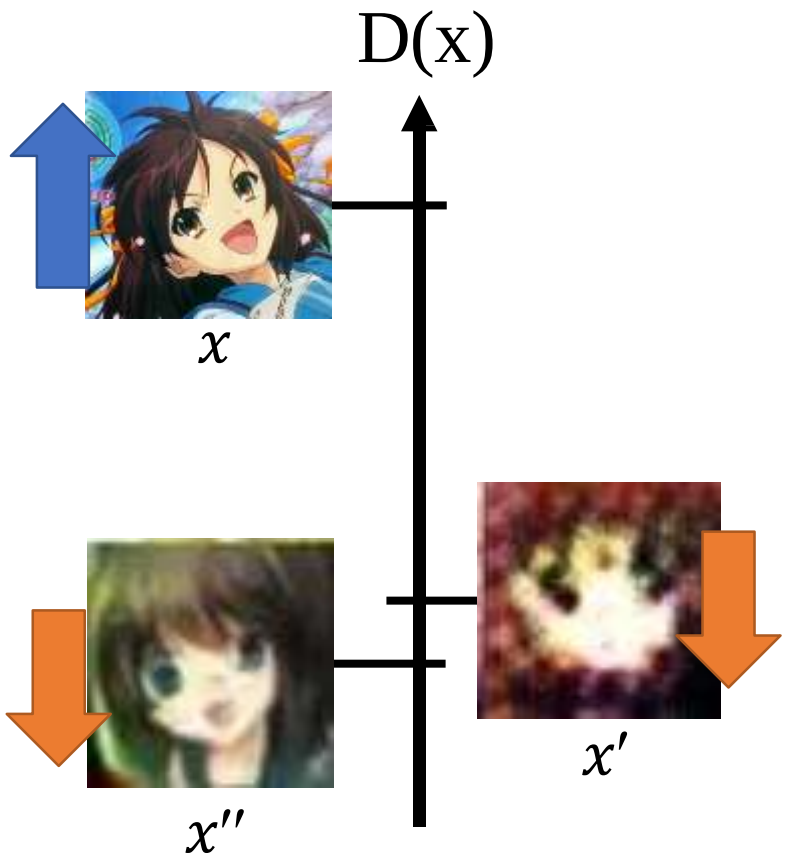
Energy-based GAN (EBGAN)

- Using an autoencoder as discriminator D
  - ➢ Using the negative reconstruction error of auto-encoder to determine the goodness
  - ➢ **Benefit**: The auto-encoder can be pre-train by real images without generator.

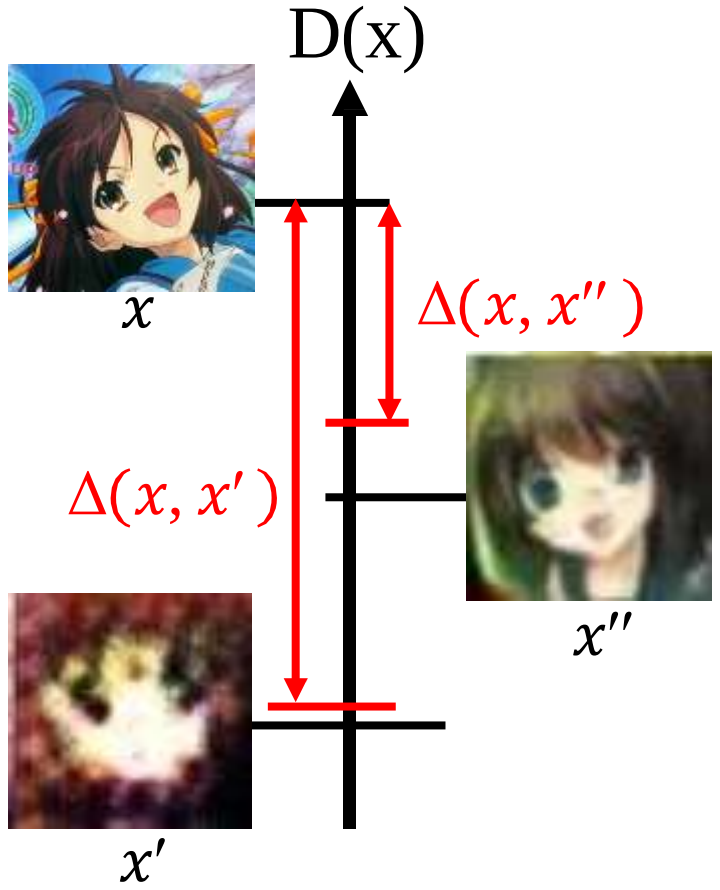


Loss-sensitive GAN (LSGAN)

WGAN



LSGAN



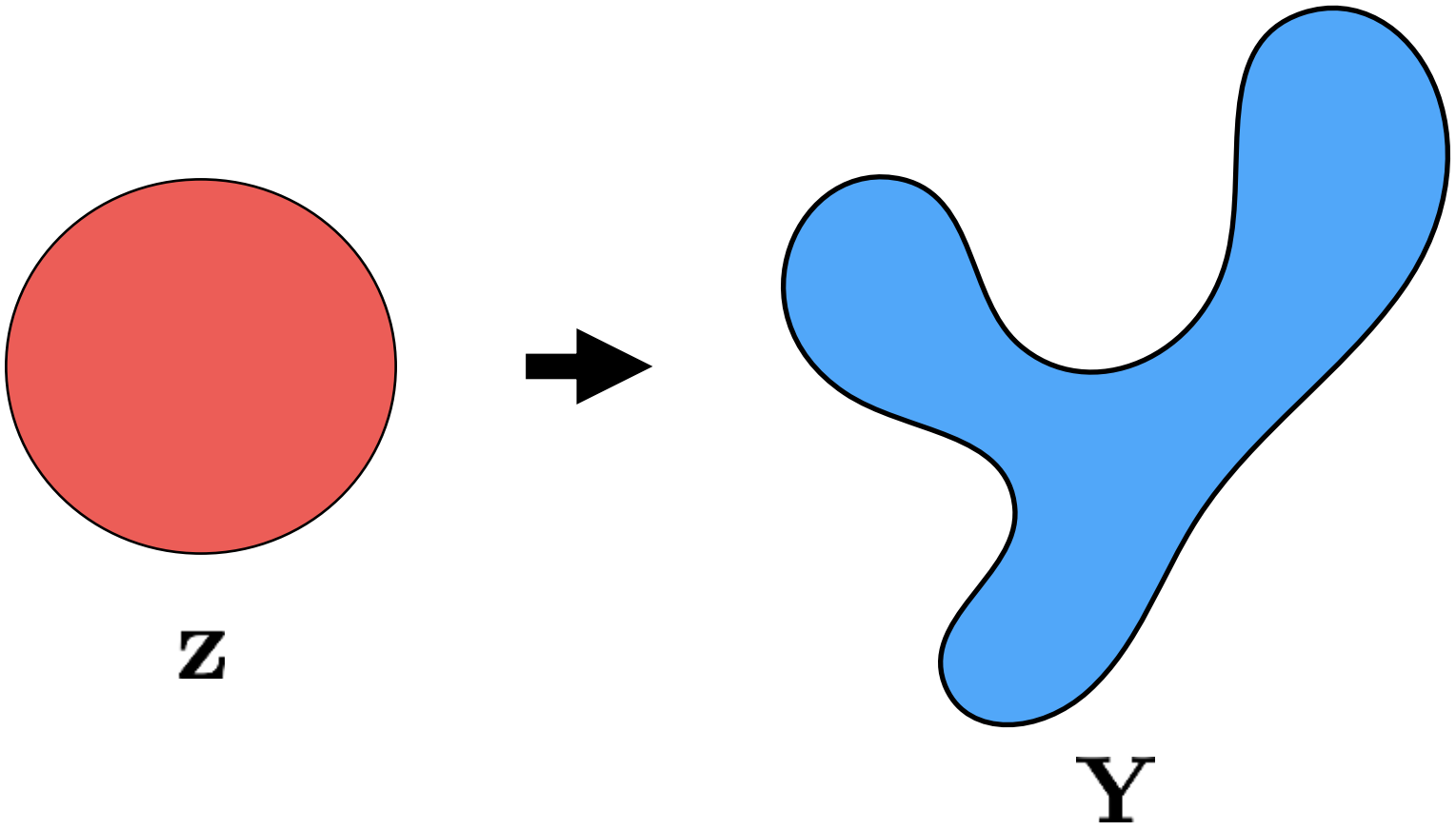
More GANs

- BigGAN
- PGGAN
- StyleGAN Series
- VQGAN

Recap: GANs

Gaussian

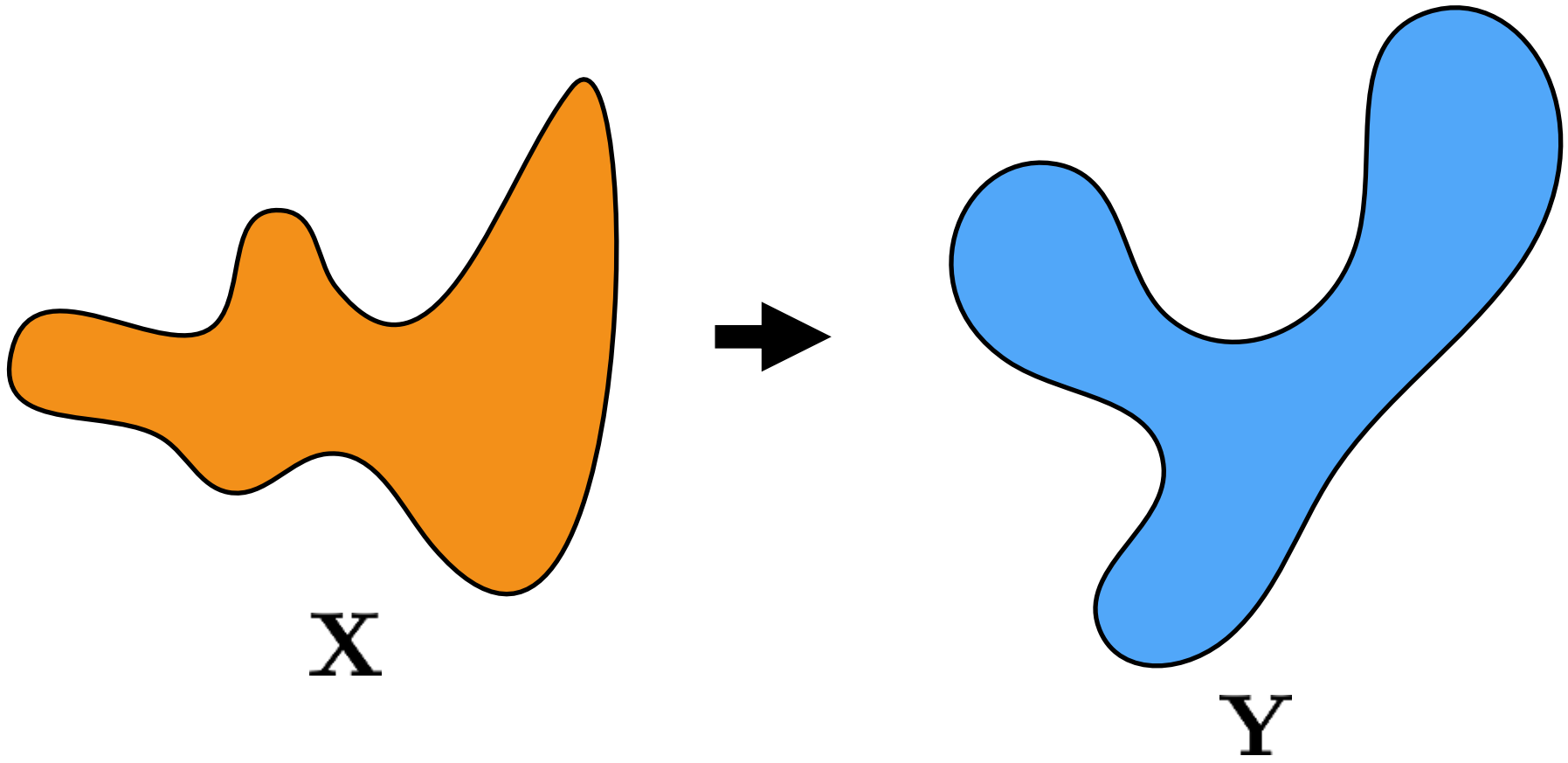
Target distribution



Recap: CycleGAN

Horses

Zebras



X

Y

BigGAN: LARGE SCALE GAN TRAINING FOR HIGH FIDELITY NATURAL IMAGE SYNTHESIS

- Bag of tricks for GANs




GANs benefit dramatically from scaling, and train models with two to four times as many parameters and eight times the batch size compared to prior art.

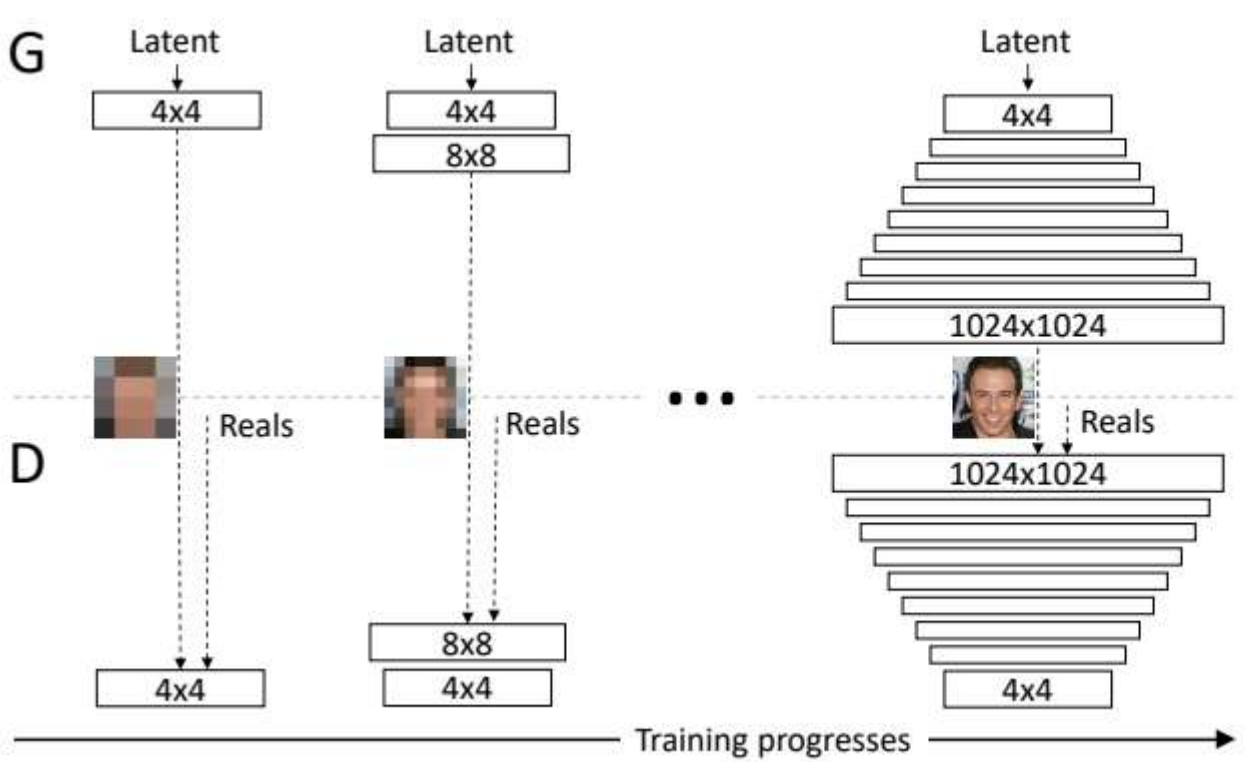
Batch	Ch.	Param (M)	Shared	Skip-z	Ortho.	Itr $\times 10^3$	FID	IS
256	64	81.5	SA-GAN Baseline			1000	18.65	52.52
512	64	81.5	✗	✗	✗	1000	15.30	58.77($\pm$1.18)
1024	64	81.5	✗	✗	✗	1000	14.88	63.03($\pm$1.42)
2048	64	81.5	✗	✗	✗	732	12.39	76.85($\pm$3.83)
2048	96	173.5	✗	✗	✗	295($\pm$18)	9.54($\pm$0.62)	92.98($\pm$4.27)
2048	96	160.6	✓	✗	✗	185($\pm$11)	9.18($\pm$0.13)	94.94($\pm$1.32)
2048	96	158.3	✓	✓	✗	152($\pm$7)	8.73($\pm$0.45)	98.76($\pm$2.84)
2048	96	158.3	✓	✓	✓	165($\pm$13)	8.51($\pm$0.32)	99.31($\pm$2.10)
2048	64	71.3	✓	✓	✓	371($\pm$7)	10.48($\pm$0.10)	86.90($\pm$0.61)

$$R_\beta(W) = \beta \|W^\top W \odot (\mathbf{1} - I)\|_{\mathrm{F}}^2,$$

https://arxiv.org/pdf/1809.11096

PGGAN: PROGRESSIVE GROWING OF GANS FOR IMPROVED QUALITY, STABILITY, AND VARIATION

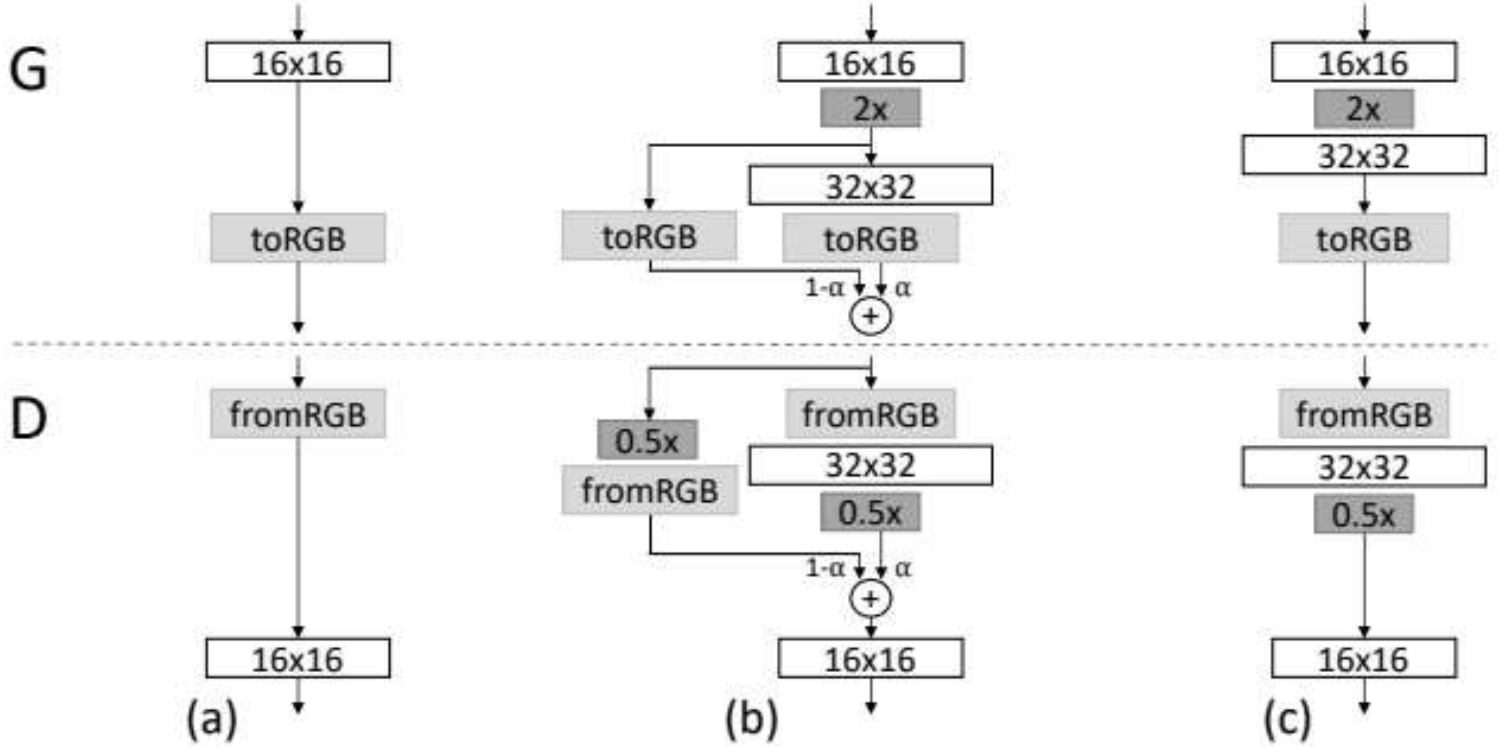
- Multi-scale framework.



https://arxiv.org/pdf/1710.10196

PGGAN: PROGRESSIVE GROWING OF GANS FOR IMPROVED QUALITY, STABILITY, AND VARIATION

- Multi-scale framework.



https://arxiv.org/pdf/1710.10196

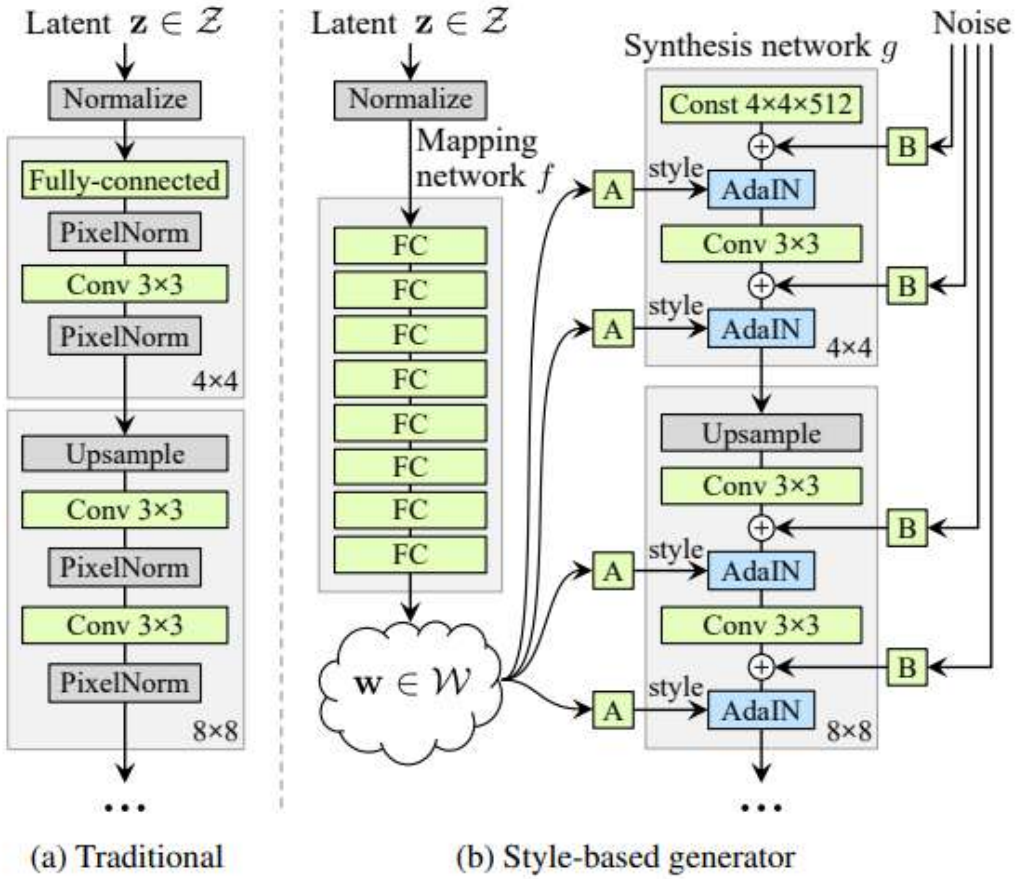
PGGAN: PROGRESSIVE GROWING OF GANS FOR IMPROVED QUALITY, STABILITY, AND VARIATION

- CelebA-HQ

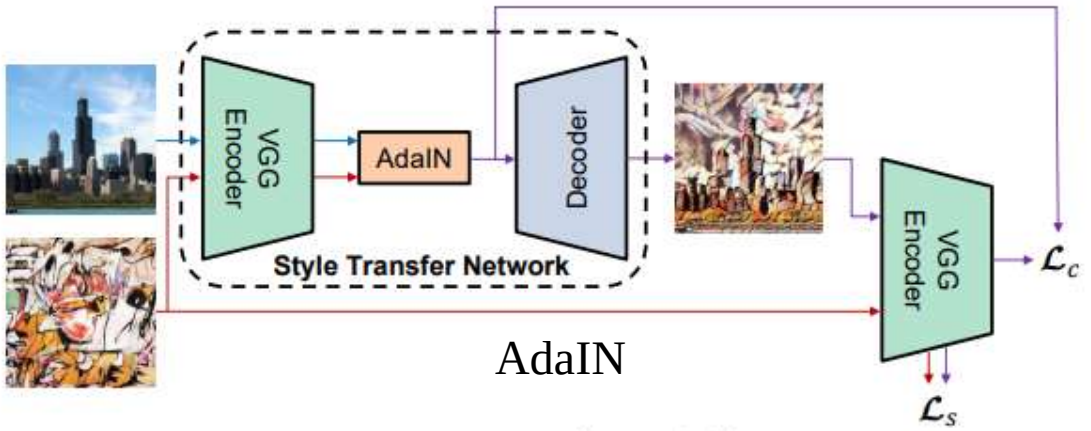
https://arxiv.org/pdf/1710.10196

STYLEGAN: A Style-Based Generator Architecture for Generative Adversarial Networks

- SoTA GAN before tokenizers.



(a) Traditional (b) Style-based generator



AdaIN

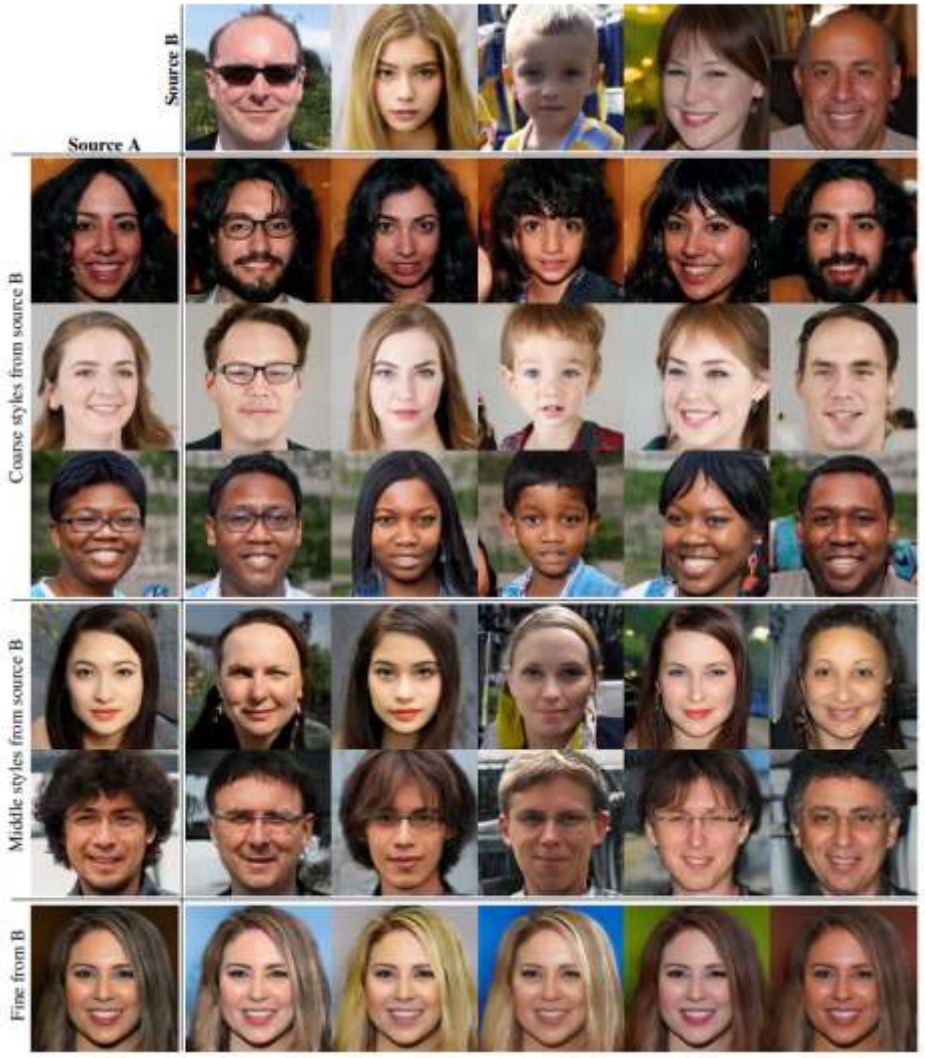
$$\text{AdaIN}(x, y) = \sigma(y)\left(\frac{x - \mu(x)}{\sigma(x)}\right) + \mu(y)$$

(a) Distribution of features in training set (b) Mapping from $\mathcal{Z}$ to features (c) Mapping from $\mathcal{W}$ to features

https://arxiv.org/pdf/1812.04948

STYLEGAN: A Style-Based Generator Architecture for Generative Adversarial Networks

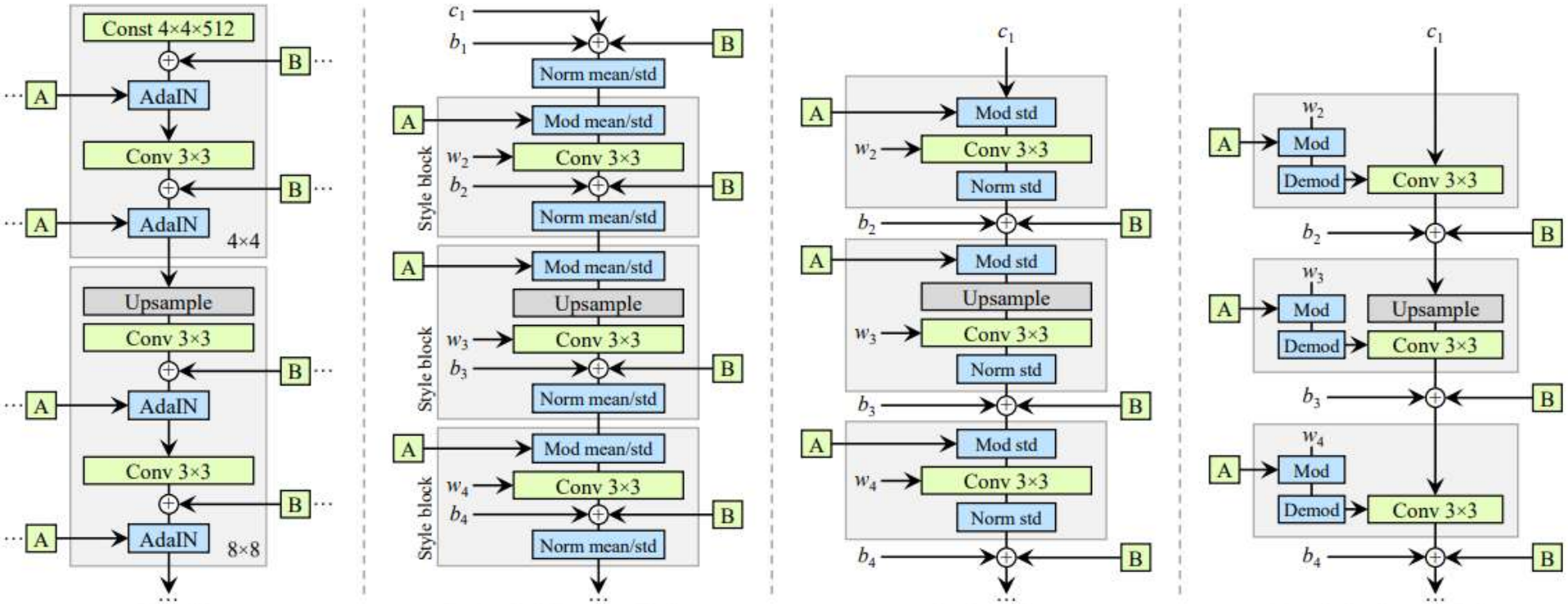
- Good editability.



https://arxiv.org/pdf/1812.04948

STYLEGANv2

- Analyses and minor updates.



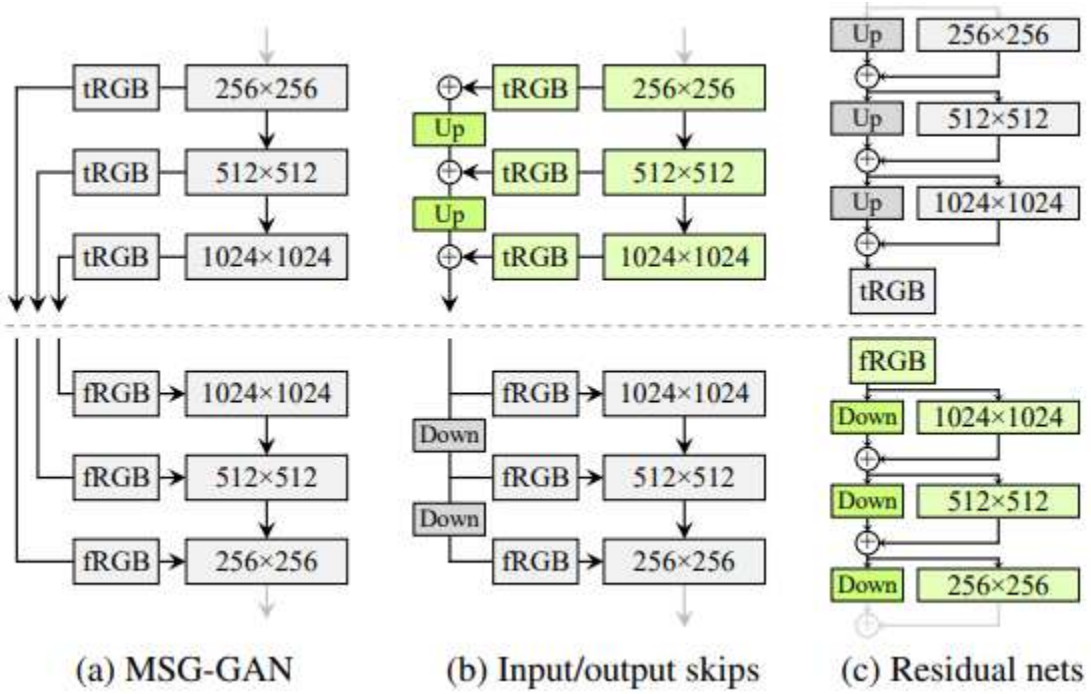
(a) StyleGAN (b) StyleGAN (detailed) (c) Revised architecture (d) Weight demodulation

Configuration	FFHQ, 1024×1024 FID ↓	Path length ↓	Precision ↑	Recall ↑	LSUN Car, 512×384 FID ↓	Path length ↓	Precision ↑	Recall ↑
A Baseline StyleGAN [24]	4.40	212.1	**0.721**	0.399	3.27	1484.5	**0.701**	0.435
B + Weight demodulation	4.39	175.4	0.702	0.425	3.04	862.4	0.685	0.488
C + Lazy regularization	4.38	158.0	0.719	0.427	2.83	981.6	0.688	0.493
D + Path length regularization	4.34	**122.5**	0.715	0.418	3.43	651.2	0.697	0.452
E + No growing, new G & D arch.	3.31	124.5	0.705	0.449	3.19	471.2	0.690	0.454
F + Large networks (StyleGAN2)	**2.84**	145.0	0.689	**0.492**	**2.32**	**415.5**	0.678	**0.514**
Config A with large networks	3.98	199.2	0.716	0.422	–	–	–	–

https://arxiv.org/pdf/1912.04958

STYLEGANv2

- Analyses and minor updates.



(a) MSG-GAN (b) Input/output skips (c) Residual nets

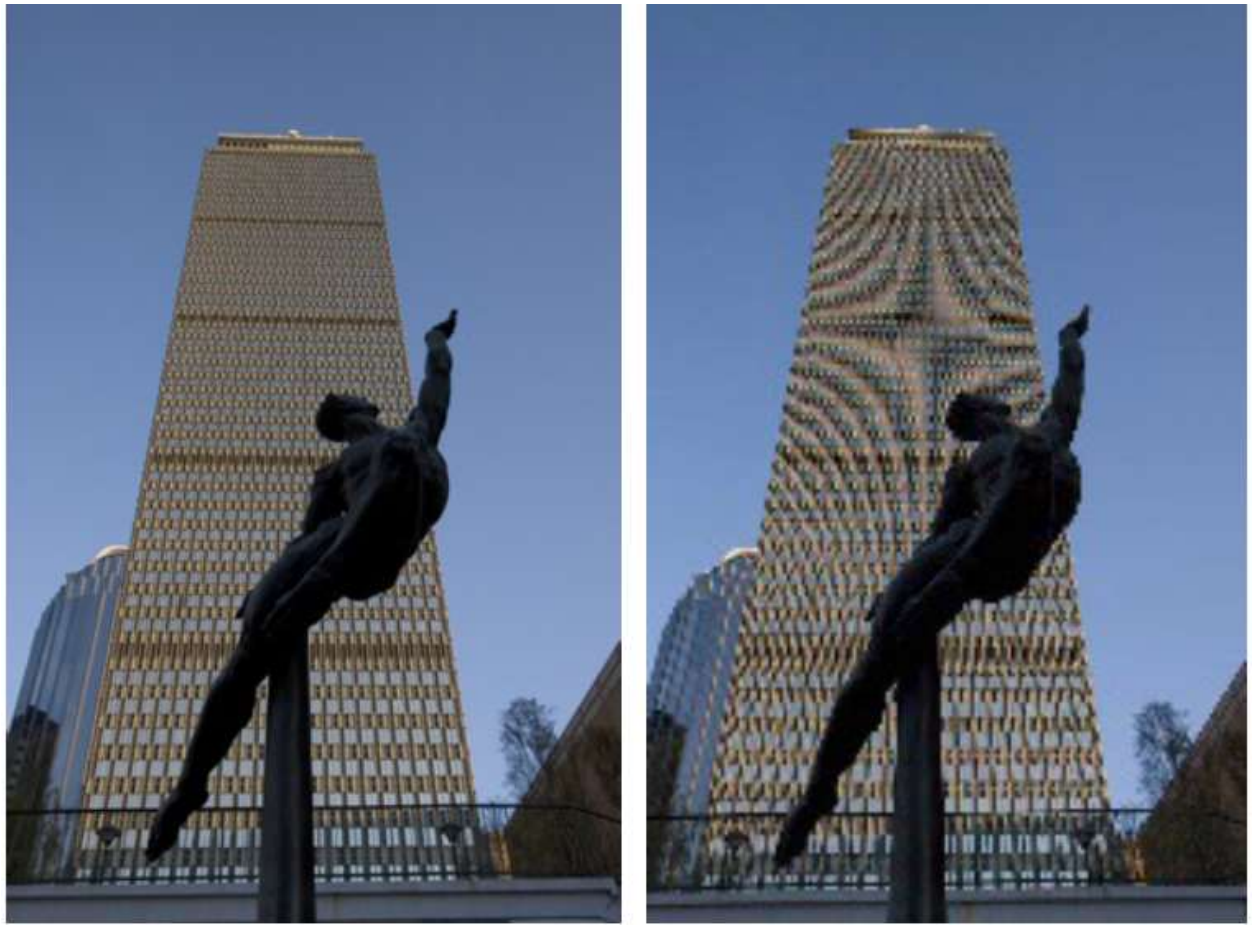
FFHQ	D original		D input skips		D residual	
	FID	PPL	FID	PPL	FID	PPL
G original	4.32	265	4.18	235	3.58	269
G output skips	4.33	169	3.77	127	**3.31**	**125**
G residual	4.35	203	3.96	229	3.79	243

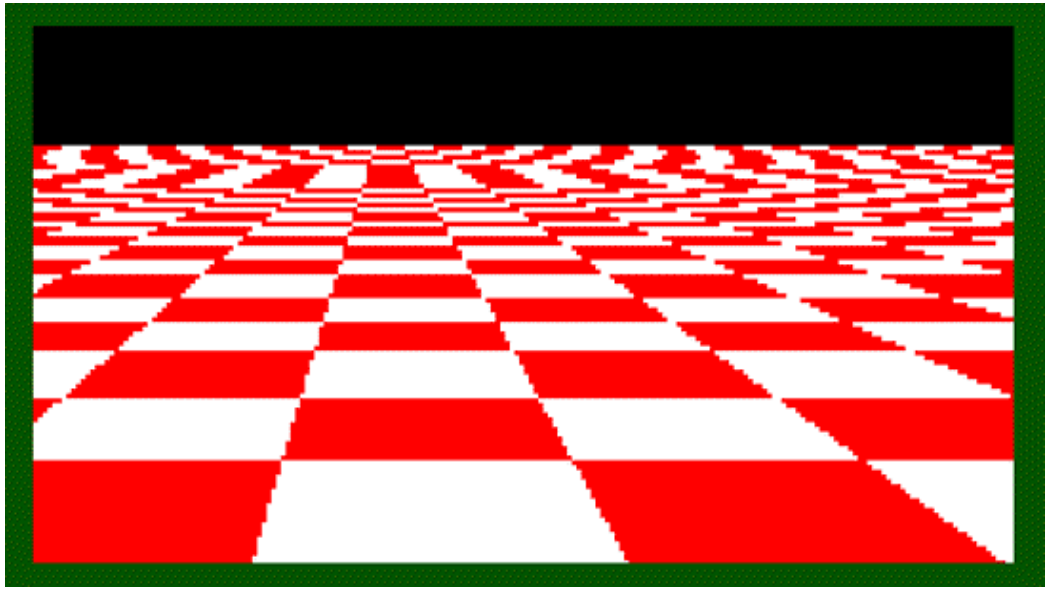
LSUN Car	D original		D input skips		D residual	
	FID	PPL	FID	PPL	FID	PPL
G original	3.75	905	3.23	758	3.25	802
G output skips	3.77	544	3.86	**316**	3.19	471
G residual	3.93	981	3.40	667	**2.66**	645

https://arxiv.org/pdf/1912.04958

STYLEGANv3: Alias-Free Generative Adversarial Networks

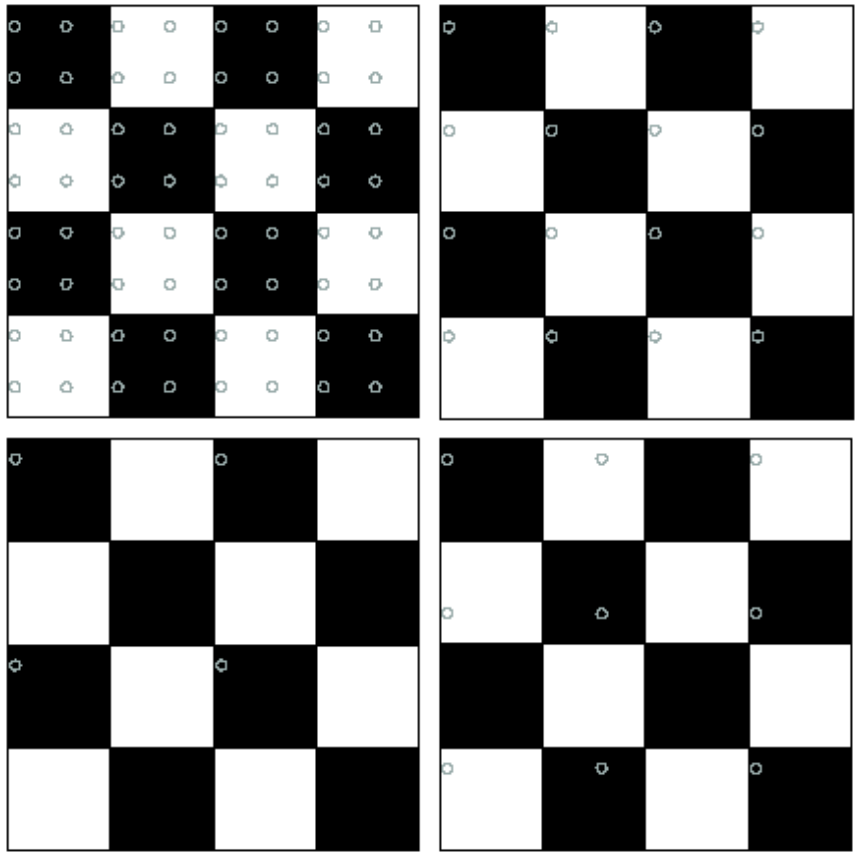
- Aliasing.

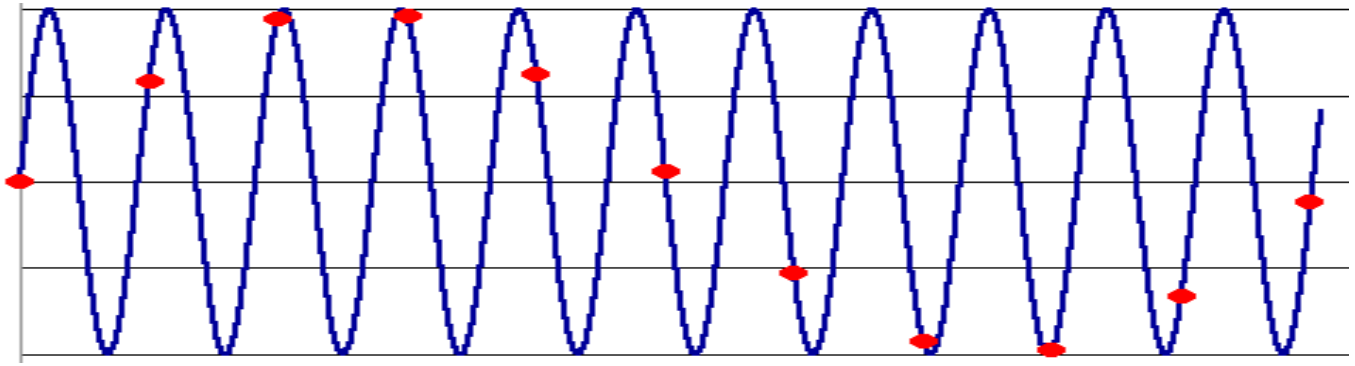




STYLEGANv3: Alias-Free Generative Adversarial Networks

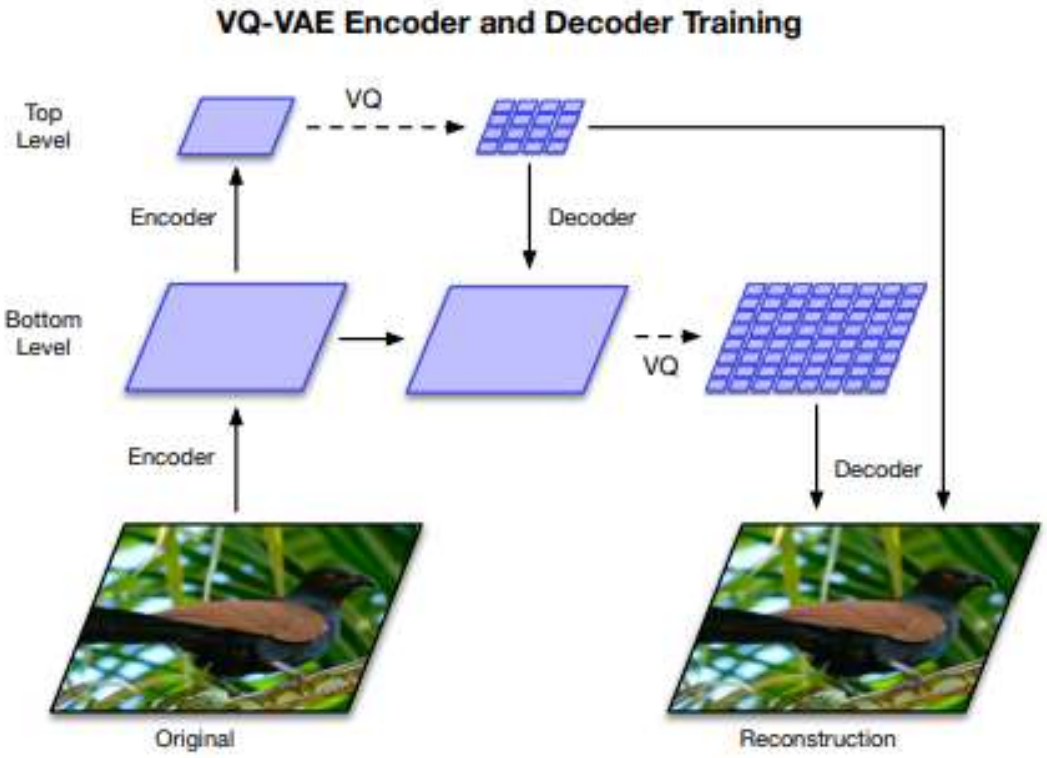
- Aliasing.

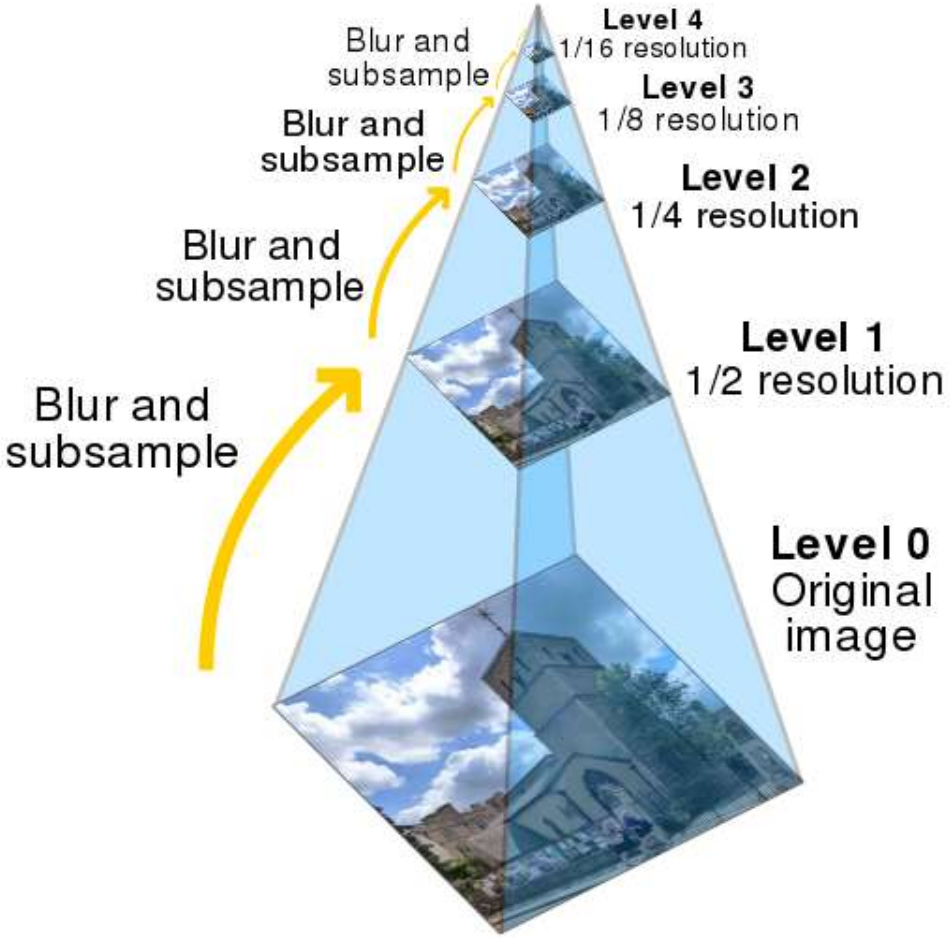




STYLEGANv3: Alias-Free Generative Adversarial Networks

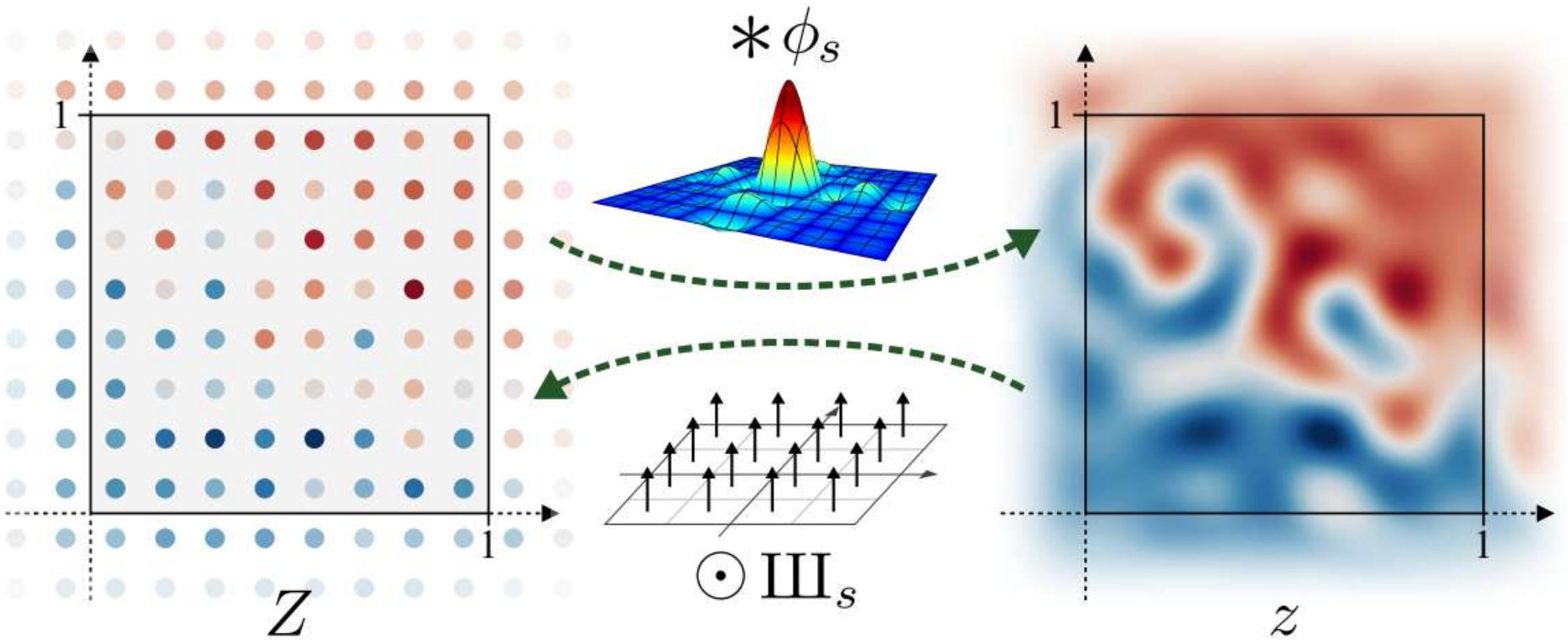
- Aliasing.





STYLEGANv3: Alias-Free Generative Adversarial Networks

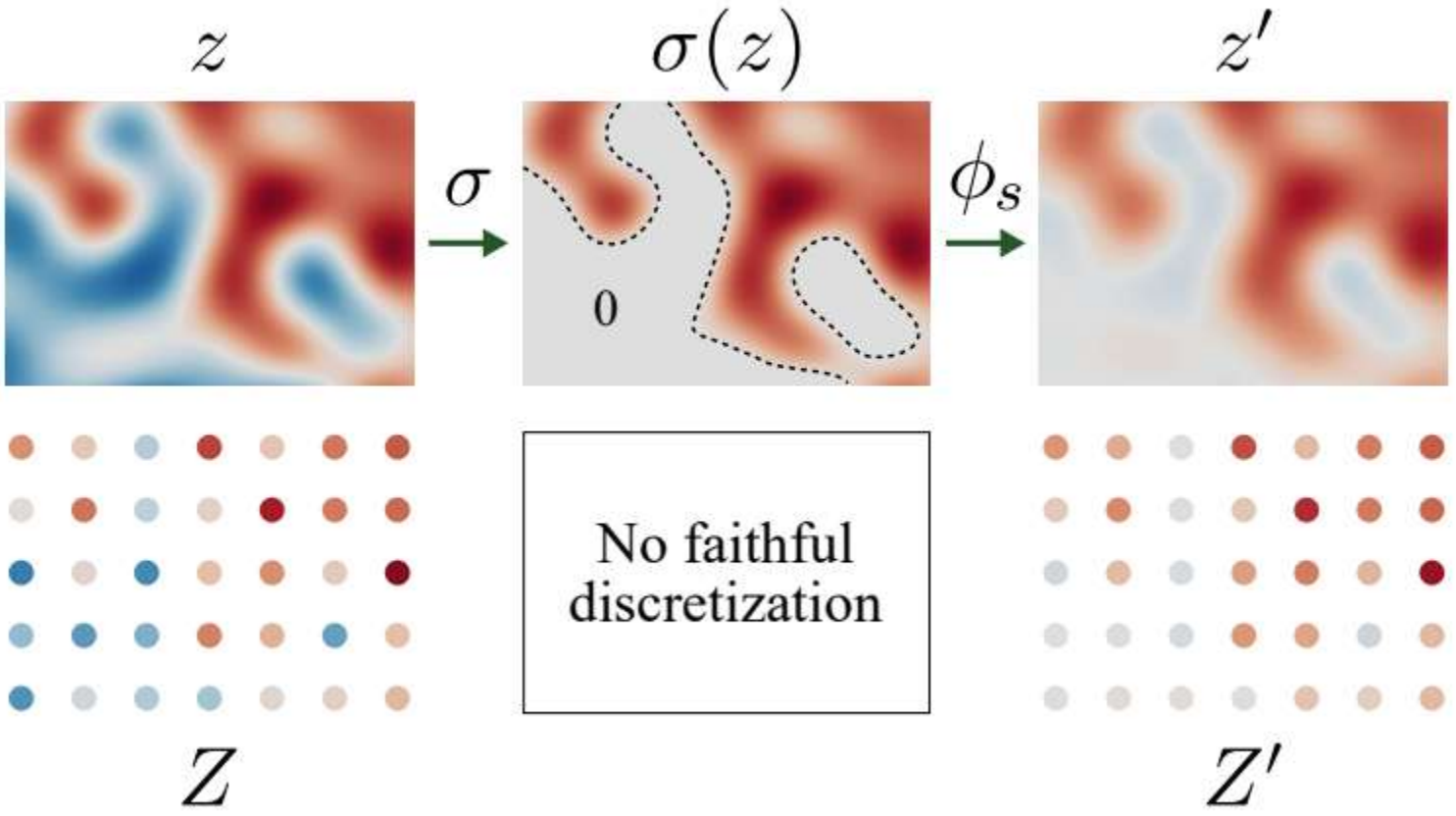
- Aliasing.



https://arxiv.org/pdf/2106.12423

STYLEGANv3: Alias-Free Generative Adversarial Networks

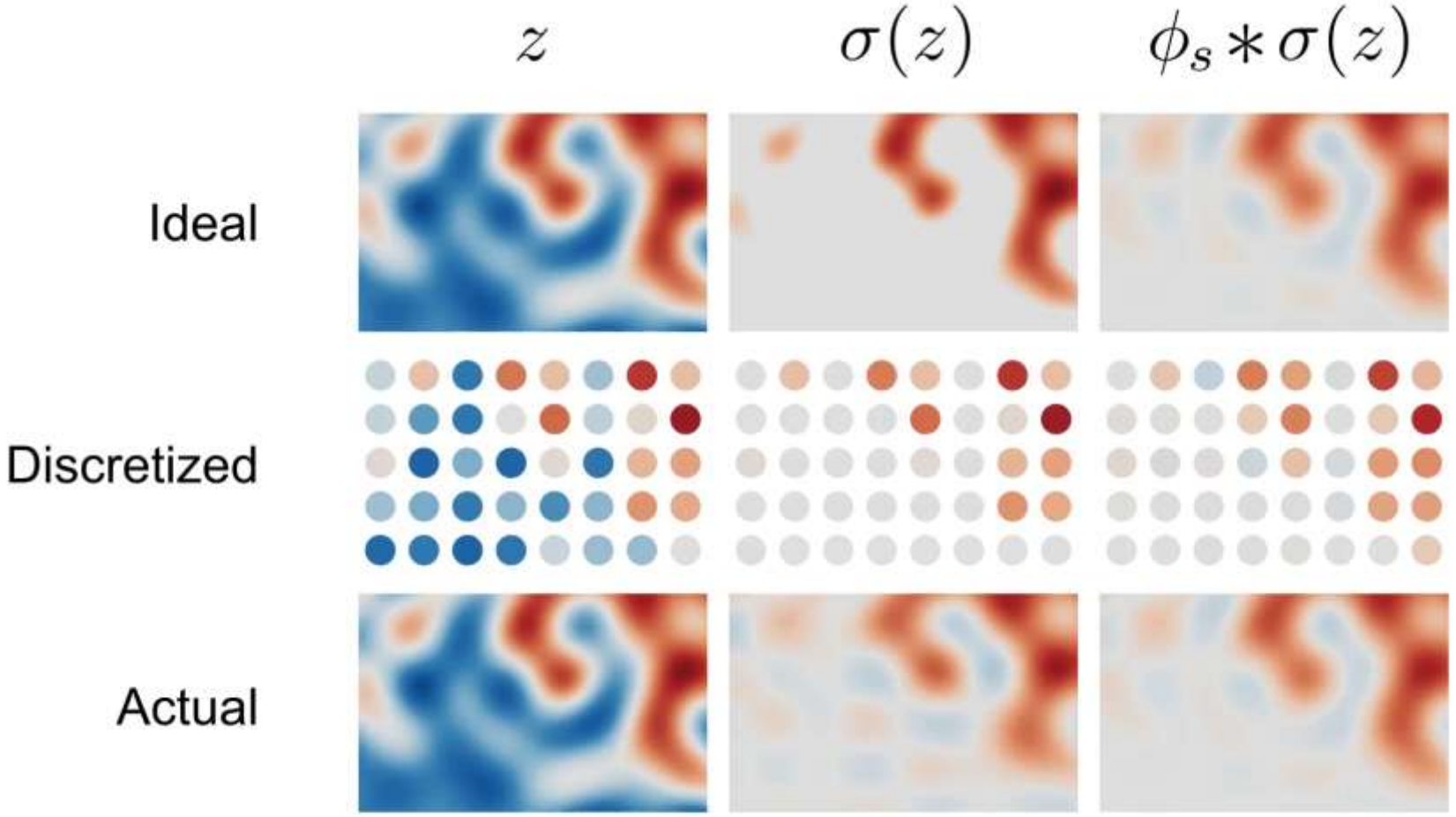
- Aliasing.



https://arxiv.org/pdf/2106.12423

STYLEGANv3: Alias-Free Generative Adversarial Networks

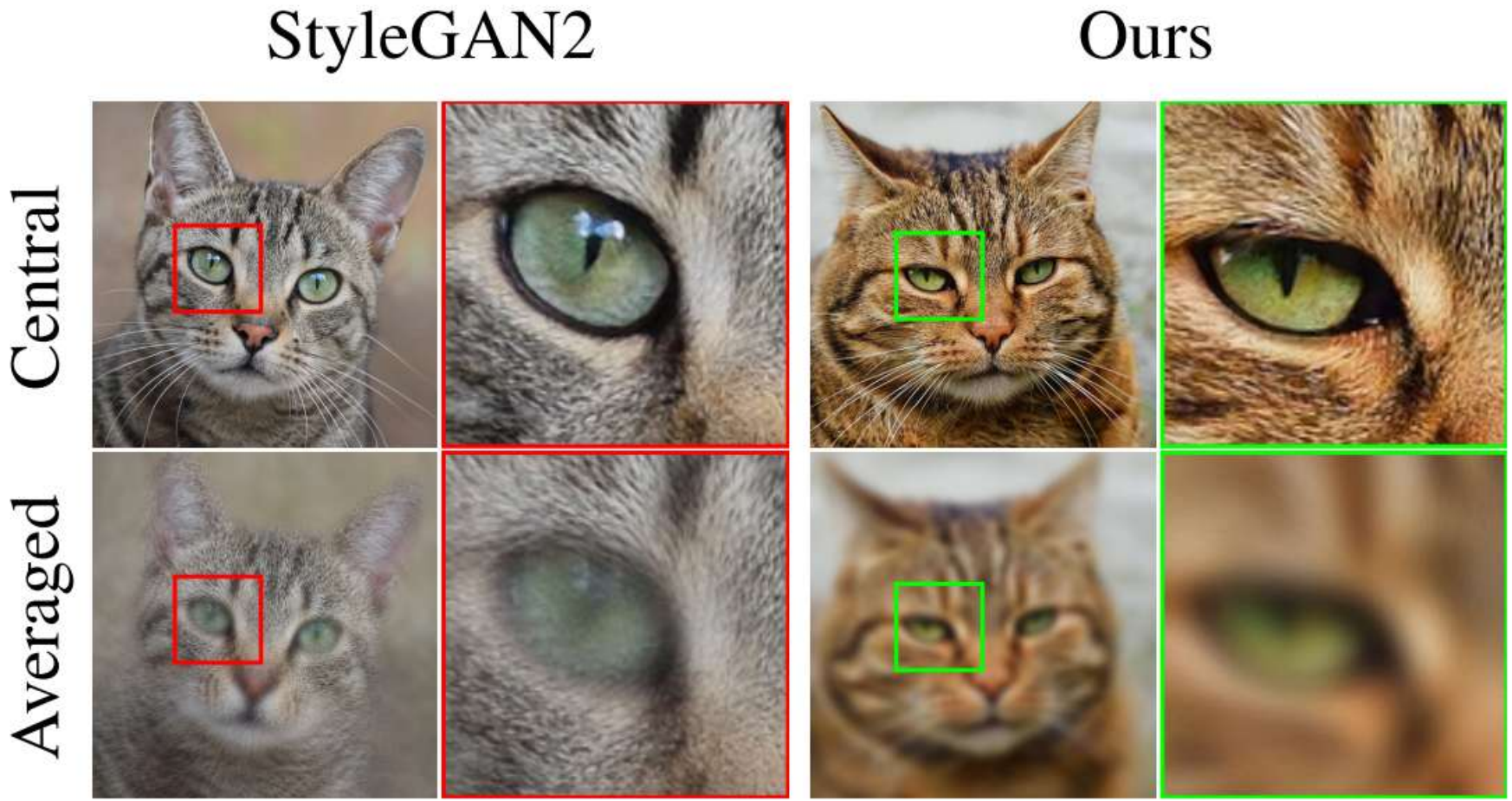
- Aliasing.



https://arxiv.org/pdf/2106.12423

STYLEGANv3: Alias-Free Generative Adversarial Networks

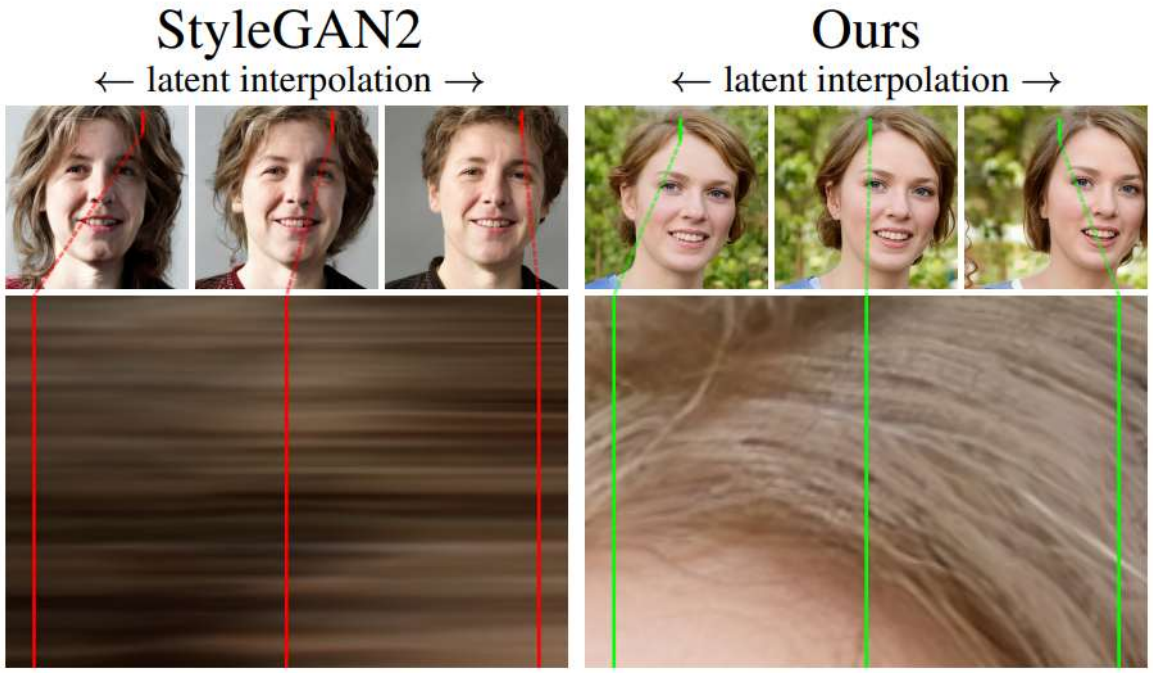
- Aliasing.



https://arxiv.org/pdf/2106.12423

STYLEGANv3: Alias-Free Generative Adversarial Networks

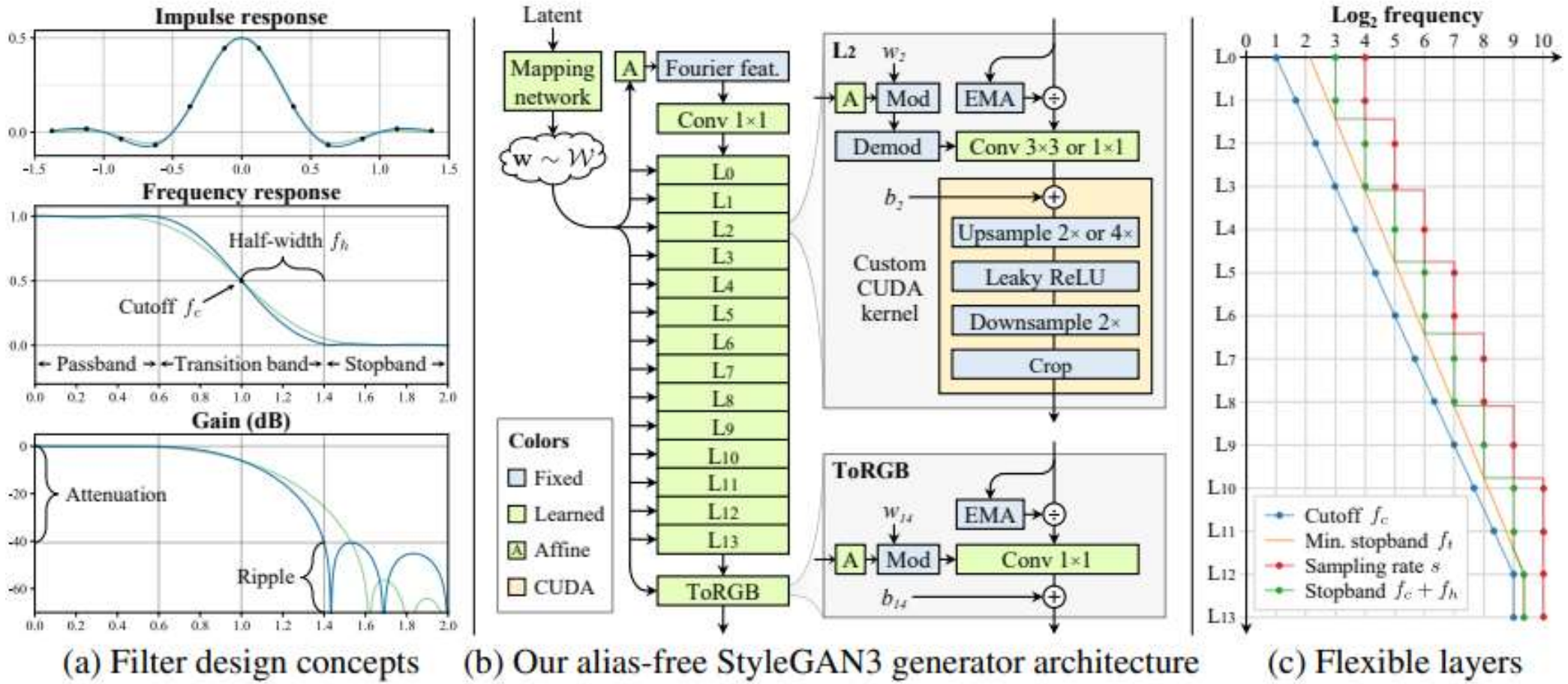
- Aliasing.



https://arxiv.org/pdf/2106.12423

STYLEGANv3: Alias-Free Generative Adversarial Networks

- Aliasing.

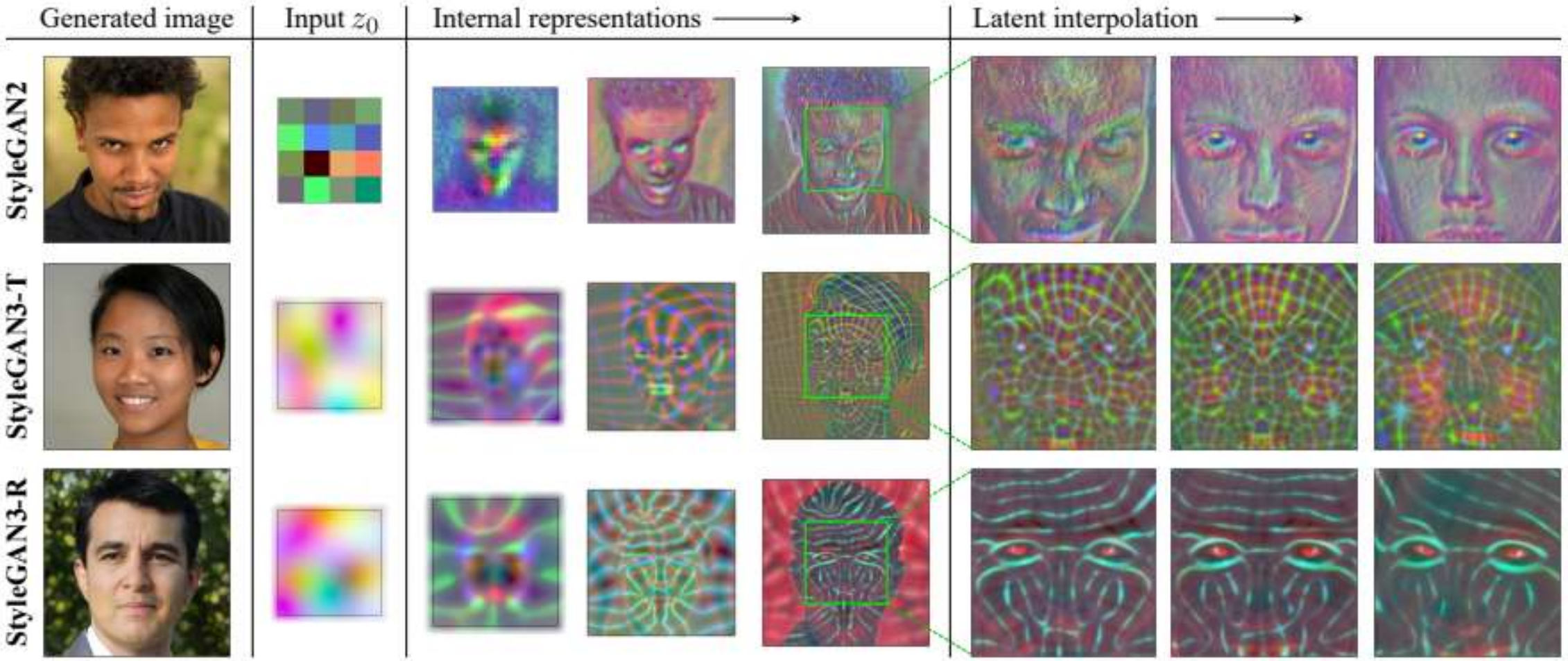


(a) Filter design concepts (b) Our alias-free StyleGAN3 generator architecture (c) Flexible layers

https://arxiv.org/pdf/2106.12423

STYLEGANv3: Alias-Free Generative Adversarial Networks

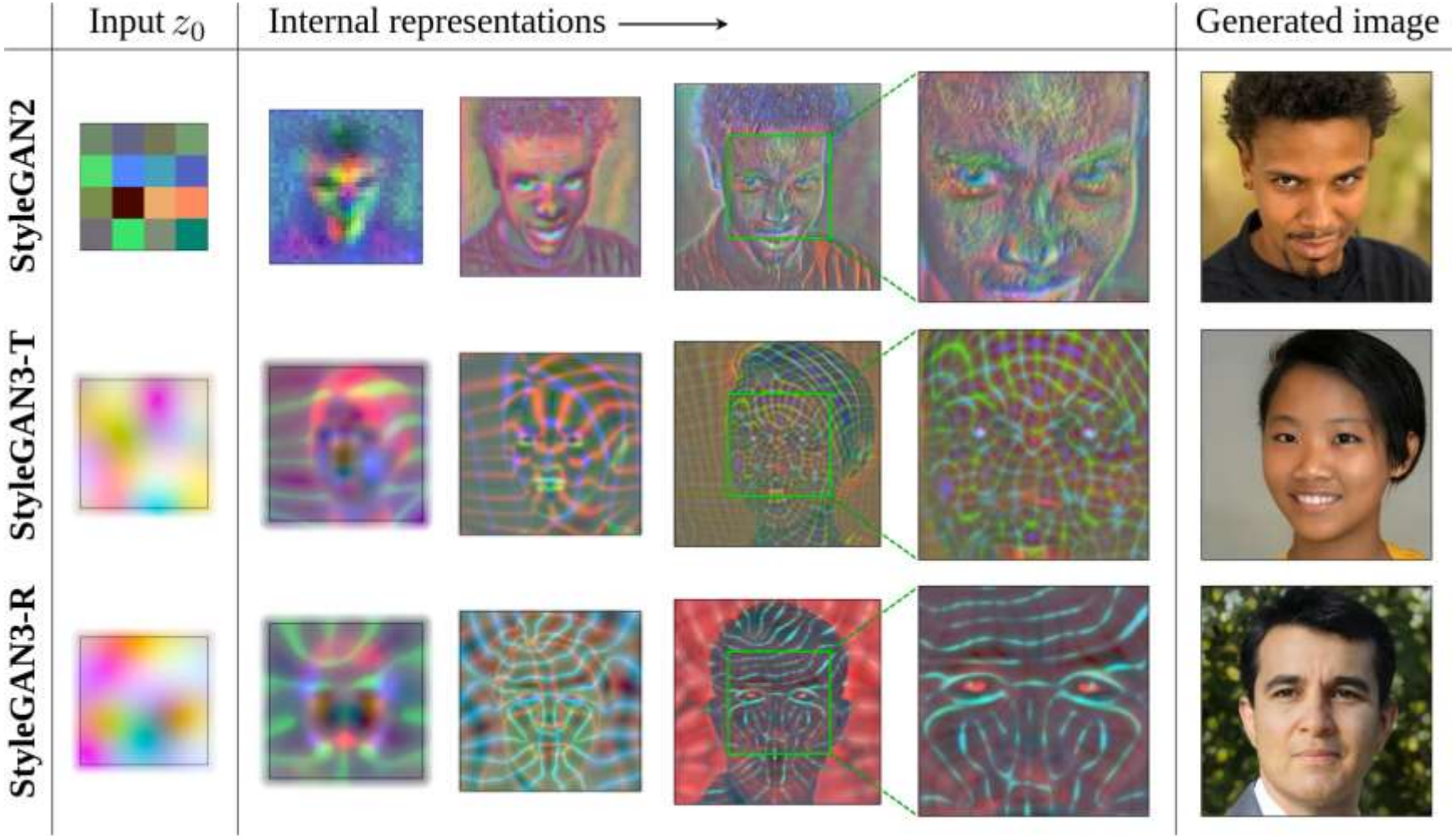
- “Coordinate system”



https://arxiv.org/pdf/2106.12423

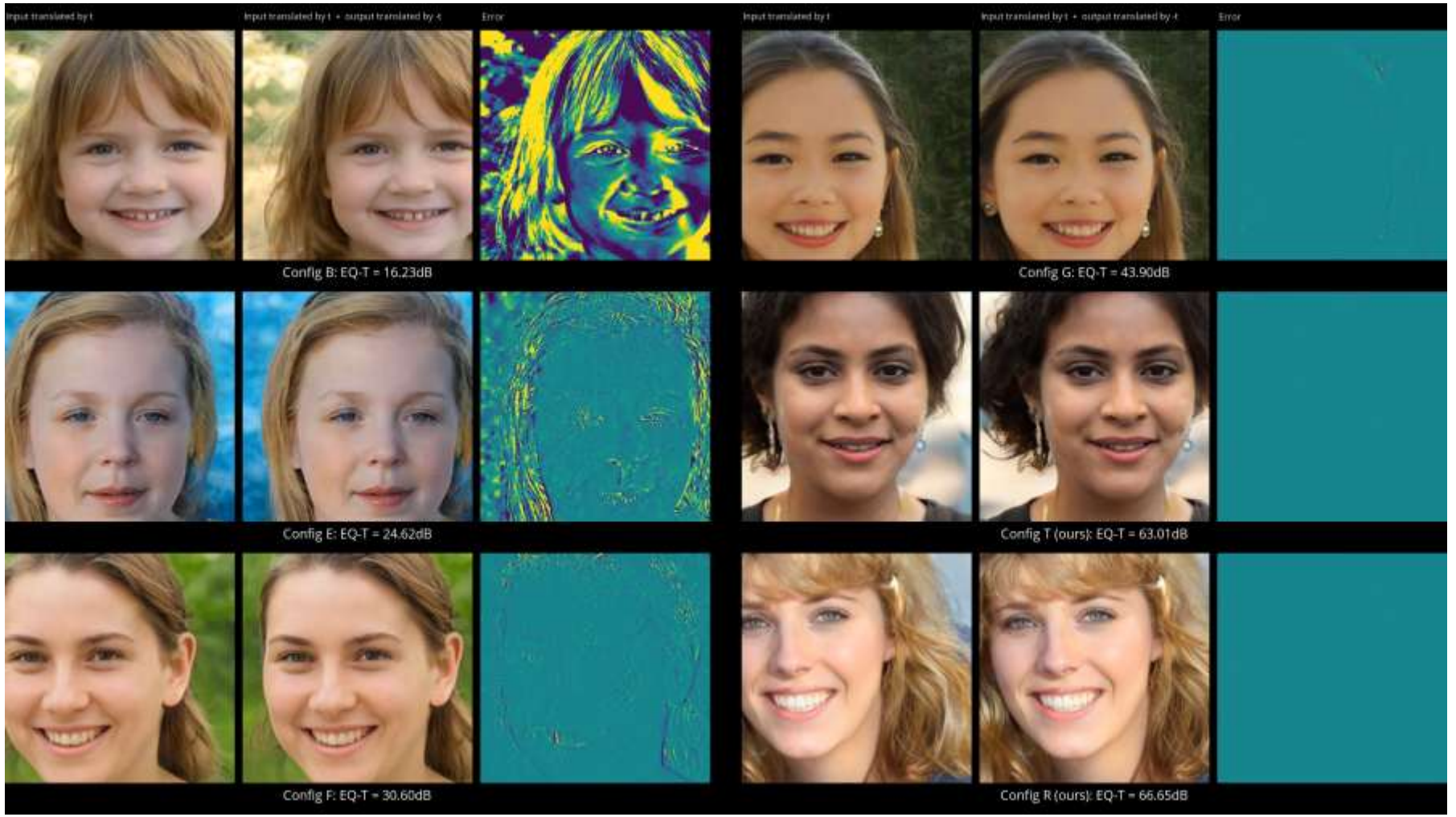
STYLEGANv3: Alias-Free Generative Adversarial Networks

- “Coordinate system”



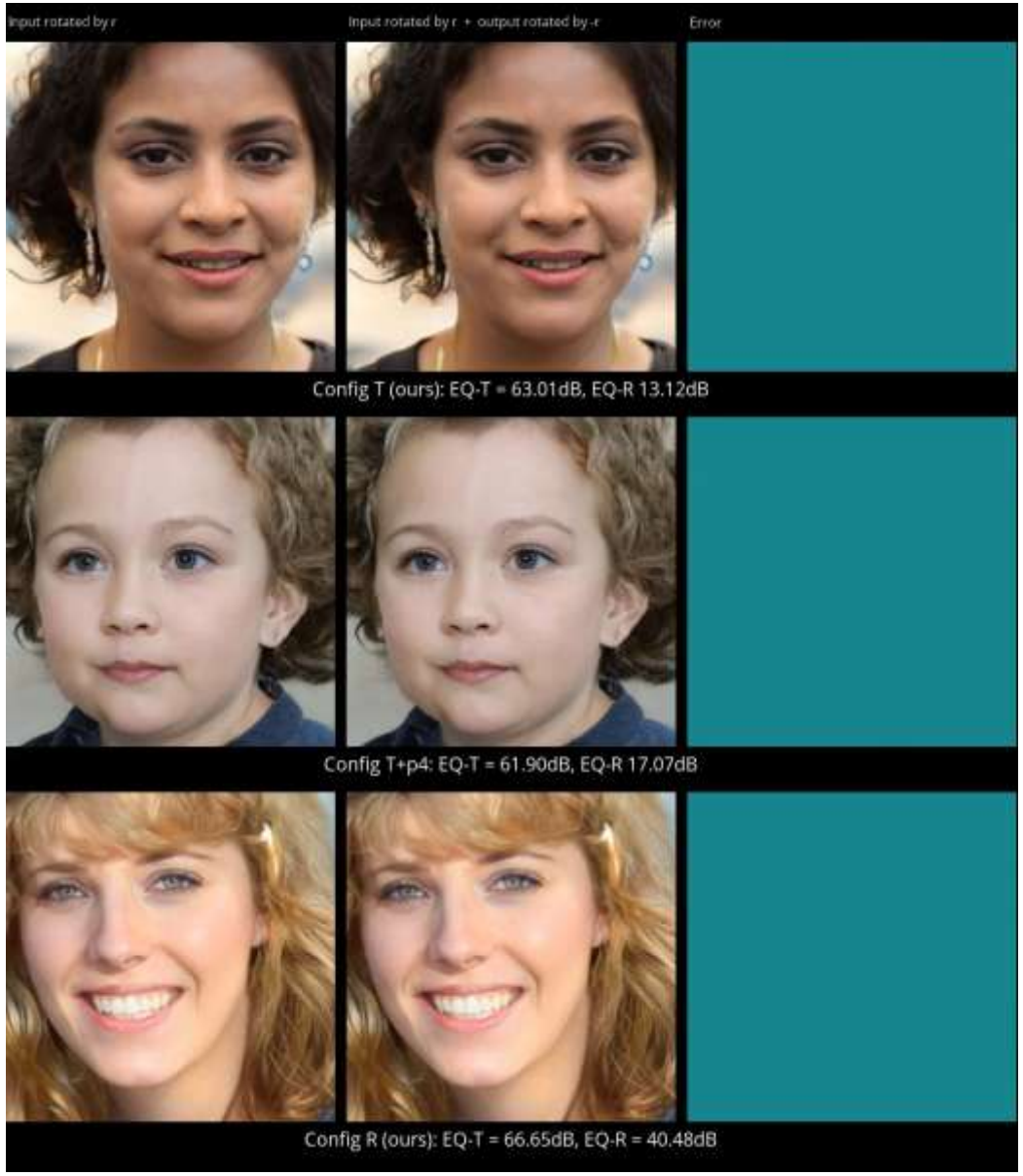
https://arxiv.org/pdf/2106.12423

STYLEGANv3: Alias-Free Generative Adversarial Networks



https://arxiv.org/pdf/2106.12423

STYLEGANv3: Alias-Free Generative Adversarial Networks



https://arxiv.org/pdf/2106.12423

STYLEGANv3: Alias-Free Generative Adversarial Networks

- Excellent interpolation.

https://arxiv.org/pdf/2106.12423

STYLEGANv3: Alias-Free Generative Adversarial Networks

- Excellent interpolation.

https://arxiv.org/pdf/2106.12423

STYLEGANv3: Alias-Free Generative Adversarial Networks

- Excellent interpolation.

https://arxiv.org/pdf/2106.12423

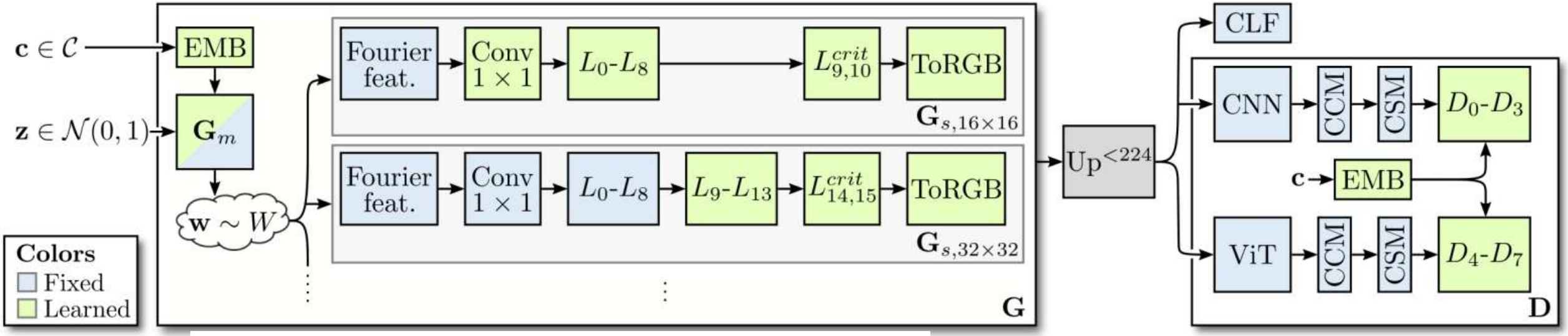
STYLEGANv3: Alias-Free Generative Adversarial Networks

- Excellent interpolation.

https://arxiv.org/pdf/2106.12423

StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets

- Training improvements.



Configuration		FID ↓	IS ↑
A	StyleGAN3	53.57	15.30
B	+ Projected GAN & small **z**	22.98	57.62
C	+ Pretrained embeddings	20.91	35.79
D	+ Progressive growing	19.51	35.74
E	+ ViT & CNN as $\mathbf{F}_{1,2}$	12.43	56.72
F	+ CLF guidance (StyleGAN-XL)	**12.24**	**86.21**

https://arxiv.org/pdf/2202.00273

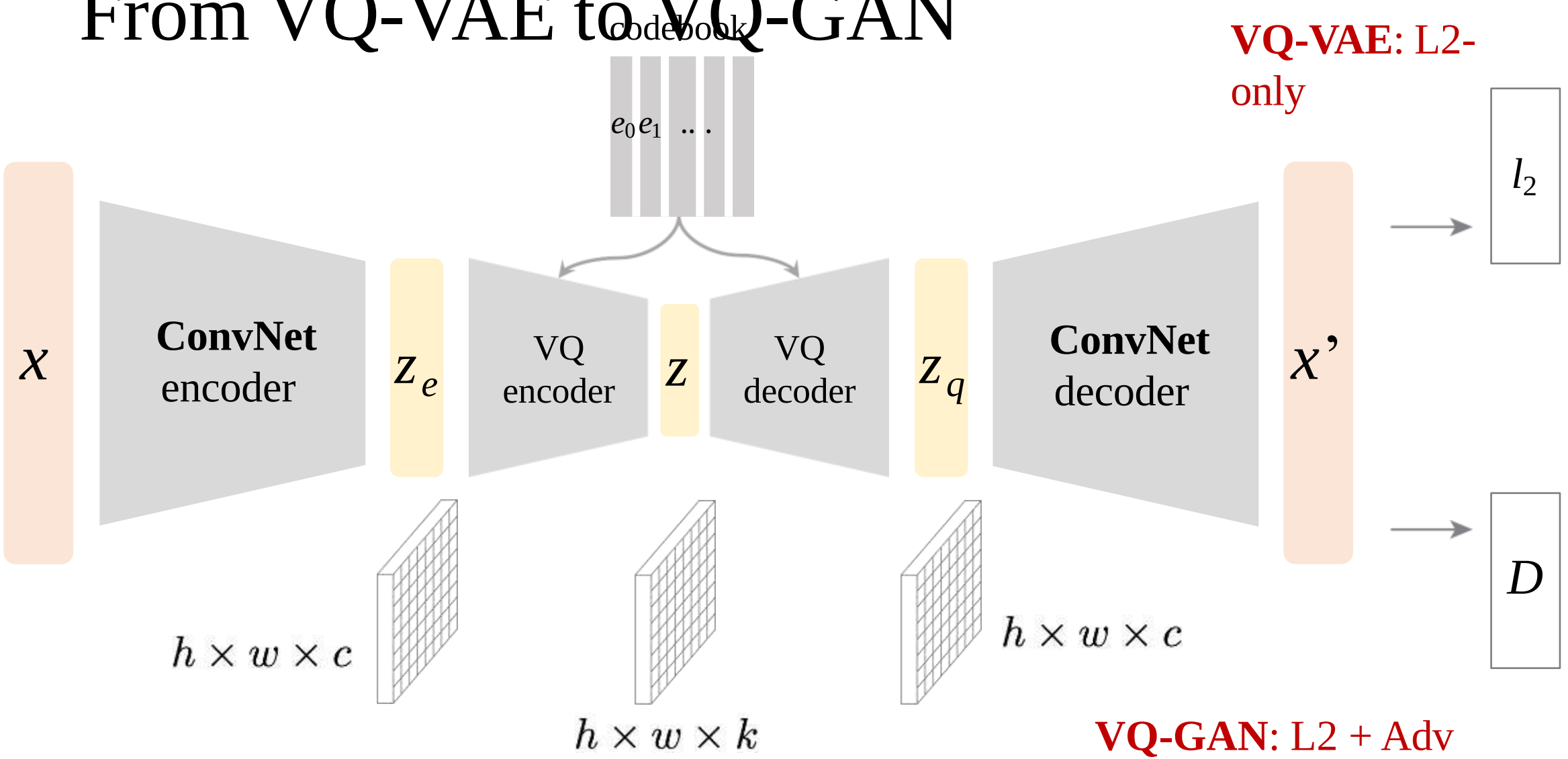
StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets

- Training improvements.

https://sites.google.com/view/stylegan-xl/

https://arxiv.org/pdf/2202.00273

From VQ-VAE to VQ-GAN

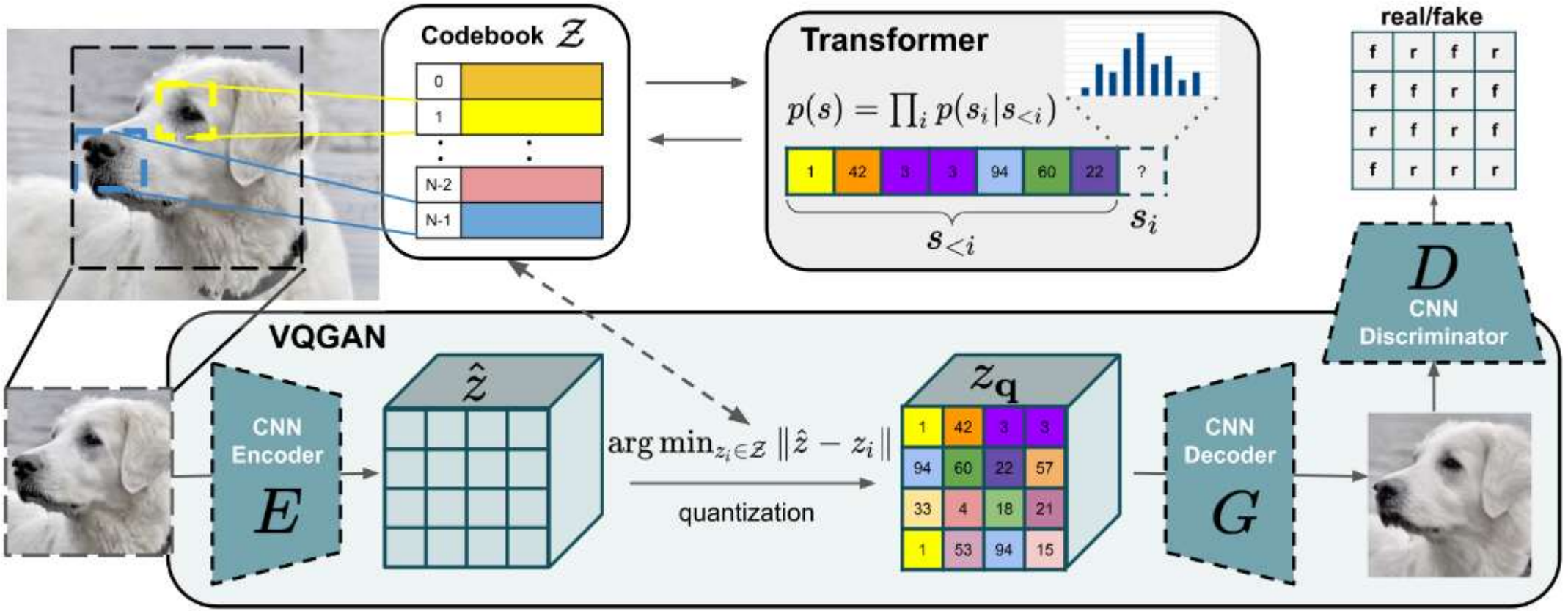


VQ-VAE: L2-only

VQ-GAN: L2 + Adv

Esser, et al., "Taming Transformers for High-Resolution Image Synthesis", CVPR 2021

VQGAN



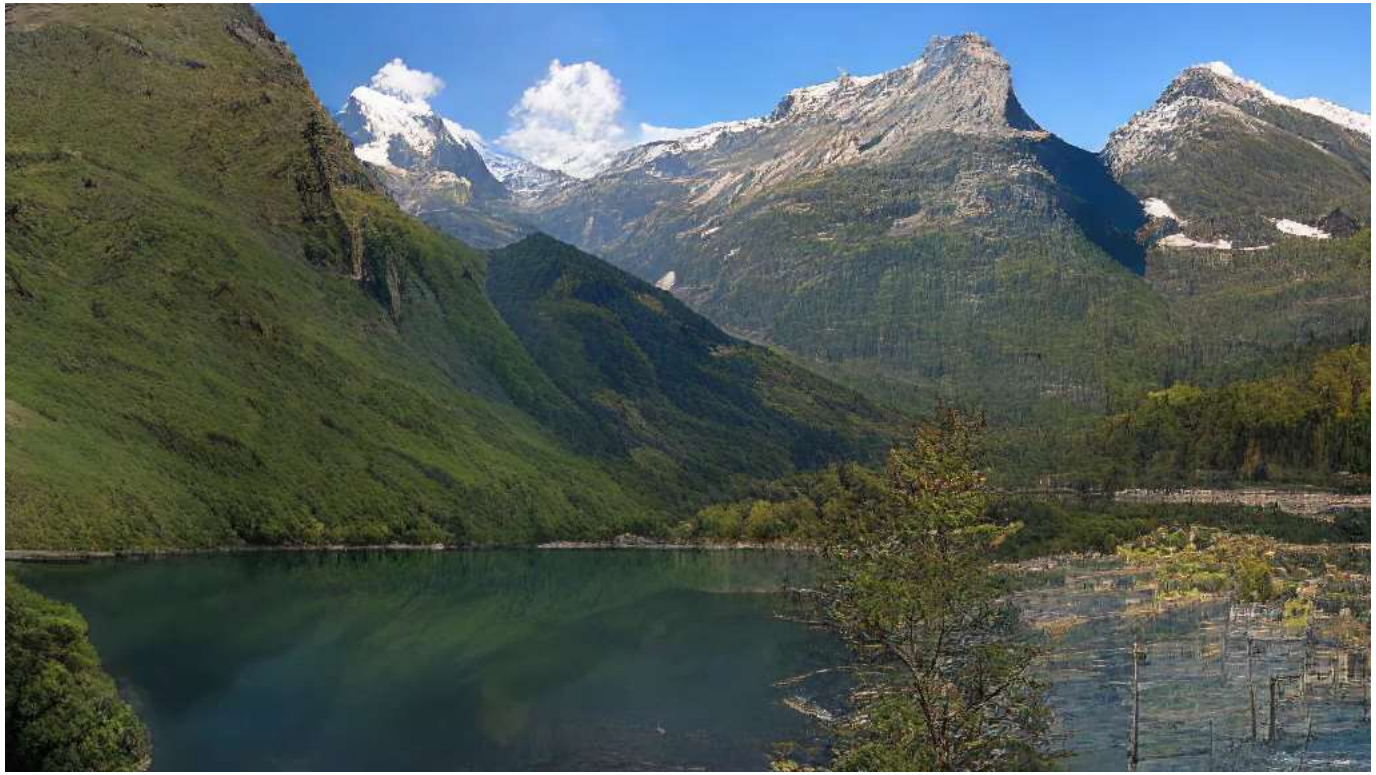
Esser, et al., "Taming Transformers for High-Resolution Image Synthesis", CVPR 2021

From VQ-VAE to VQ-GAN

VQ-VAE **VQ-GAN**

Esser, et al., “Taming Transformers for High-Resolution Image Synthesis”, CVPR 2021

From VQ-VAE to VQ-GAN



Esser, et al., “Taming Transformers for High-Resolution Image Synthesis”, CVPR 2021

Discussion

- To be precise: **VQ-GAN** = **VQ-VAE** + Adv Loss + Perceptual Loss
- w/o VQ, it's **VAE** + Adv Loss + Perceptual Loss
- Both are the *de facto* **tokenizers** in image generation
  - w/ VQ: e.g., Autoregressive Models
  - w/o VQ: e.g., Diffusion Models
- Commercial models (e.g., Stable Diffusion, Sora) use these tokenizers

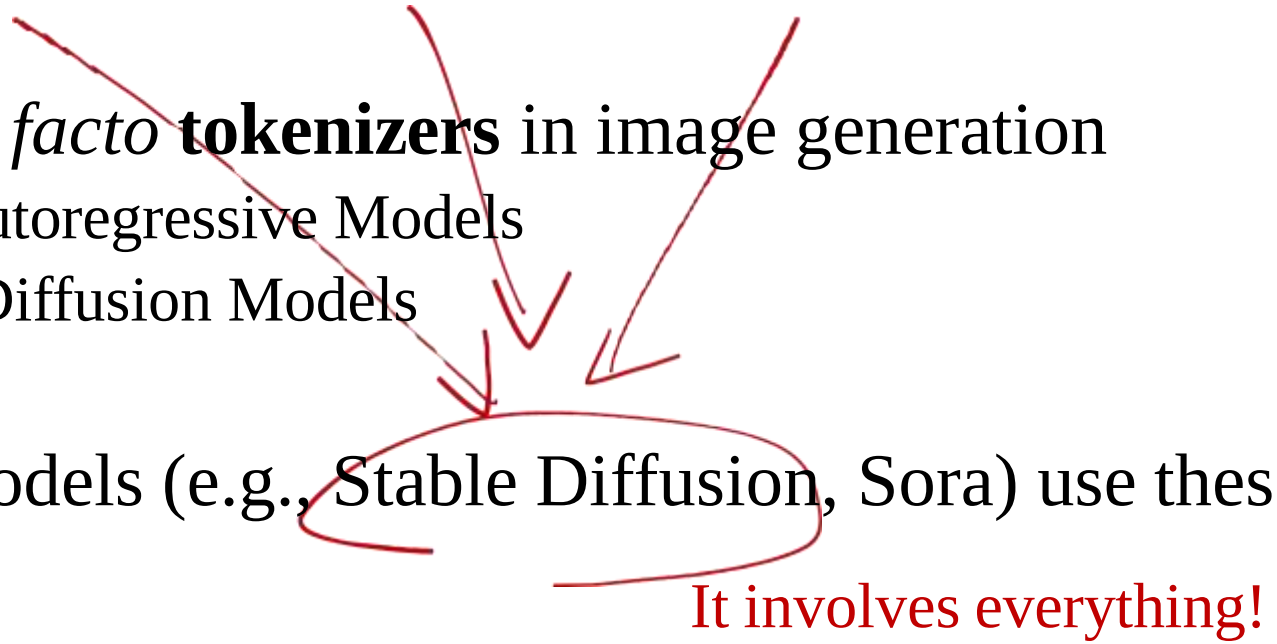
Discussion

- To be precise: **VQ-GAN** = **VQ-VAE** + Adv Loss + Perceptual Loss

- w/o VQ, it's **VAE** + Adv Loss + Perceptual Loss

- Both are the *de facto* **tokenizers** in image generation
  - w/ VQ: e.g., Autoregressive Models
  - w/o VQ: e.g., Diffusion Models

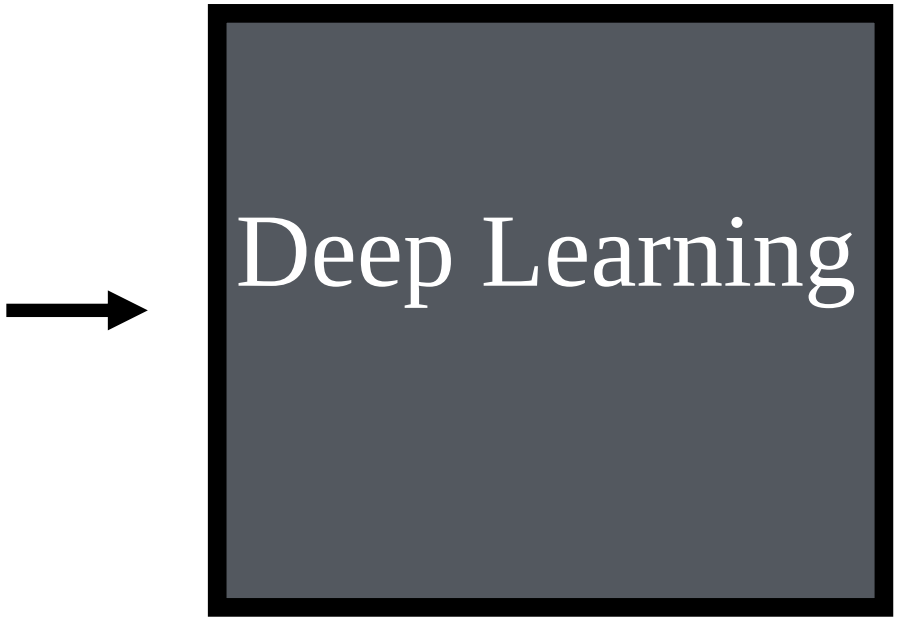
- Commercial models (e.g., Stable Diffusion, Sora) use these tokenizers



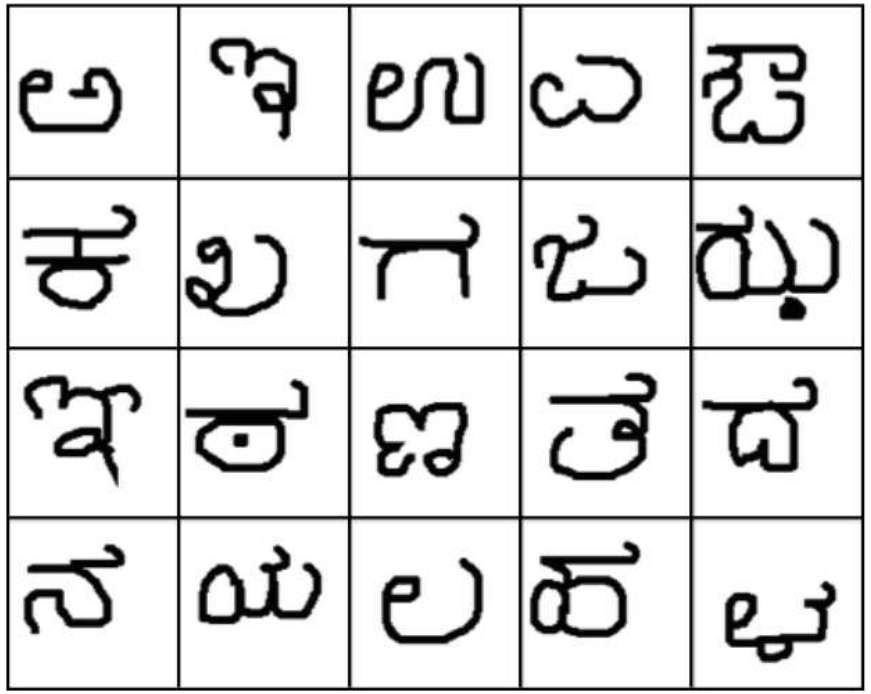
It involves everything!

Text Tokenizers Meet Vision Generative Centent

With enough data, deep learning can solve pretty much anything



Few-shot Learning



This is a “dax”.

Which of the below symbols are also daxes?

[Lake, Salakhutdinov, Tenenbaum, 2015]

Few-shot Learning

Which of these is an example of the same concept as the item in the box?

[Lake, Salakhutdinov, Tenenbaum, 2015]

The point of deep learning is to enable learning with little data

Deep learning

Representations
(encoders)

Models
(decoders)

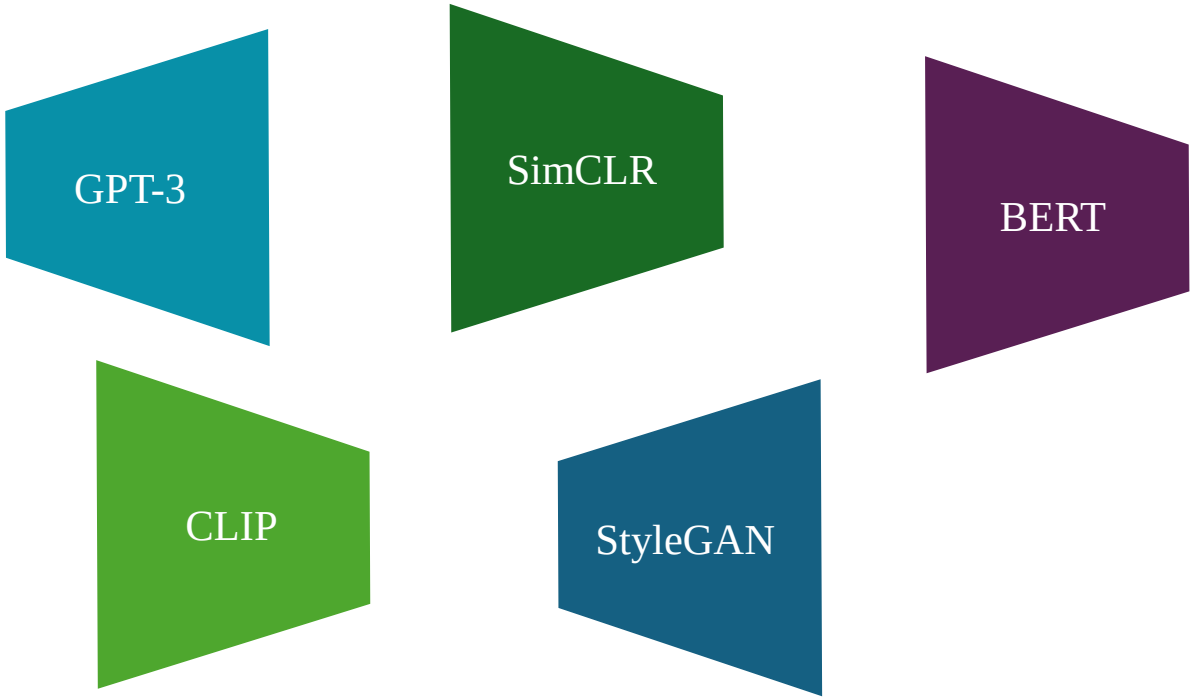
Foundation models

[Bommasani et al. 2021] https://arxiv.org/pdf/2108.07258.pdf

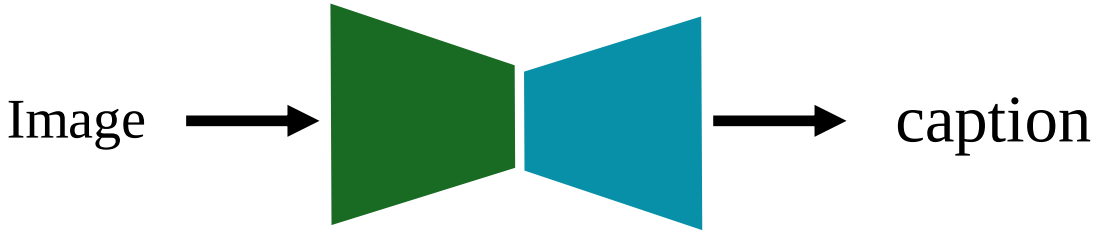
“If I have seen further it is by standing on the shoulders of Giants”
— Newton

[Blind Orion Searching for the Rising Sun by Nicolas Poussin, 1658]

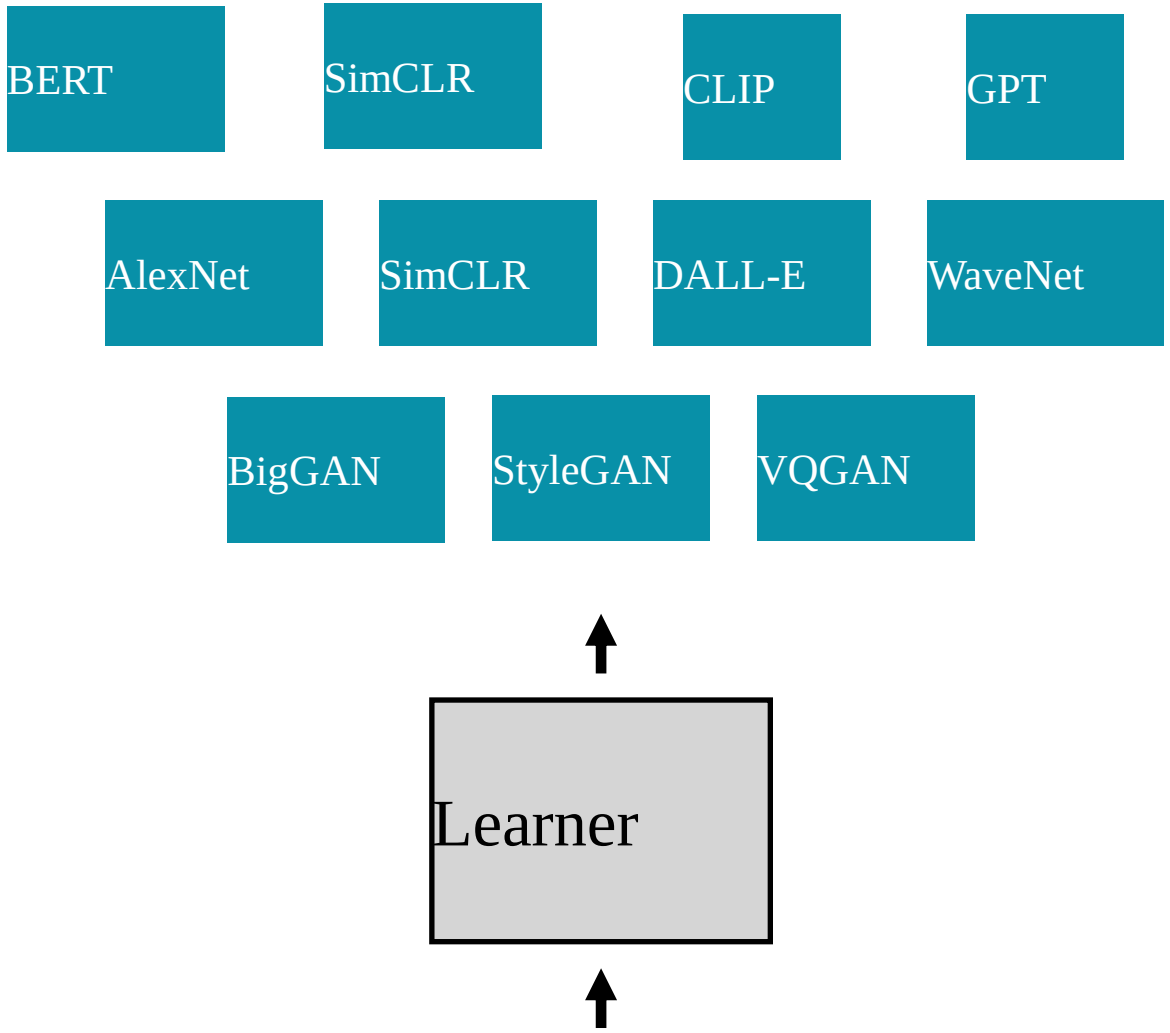
1. Learn foundation model encoders and decoders for each domain



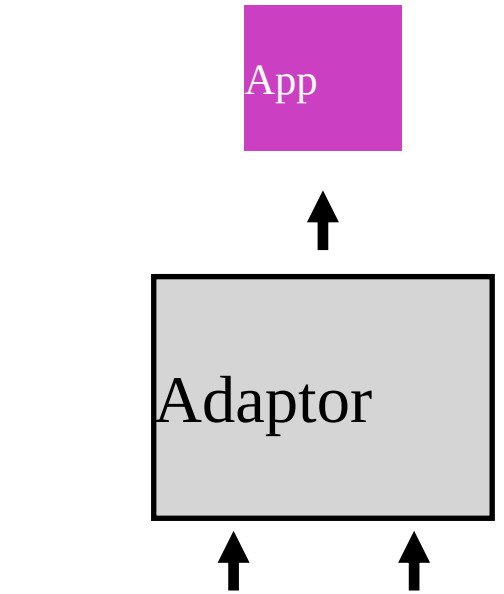
2. Plug them together to translate between modalities
(may require finetuning)



Learn foundation models

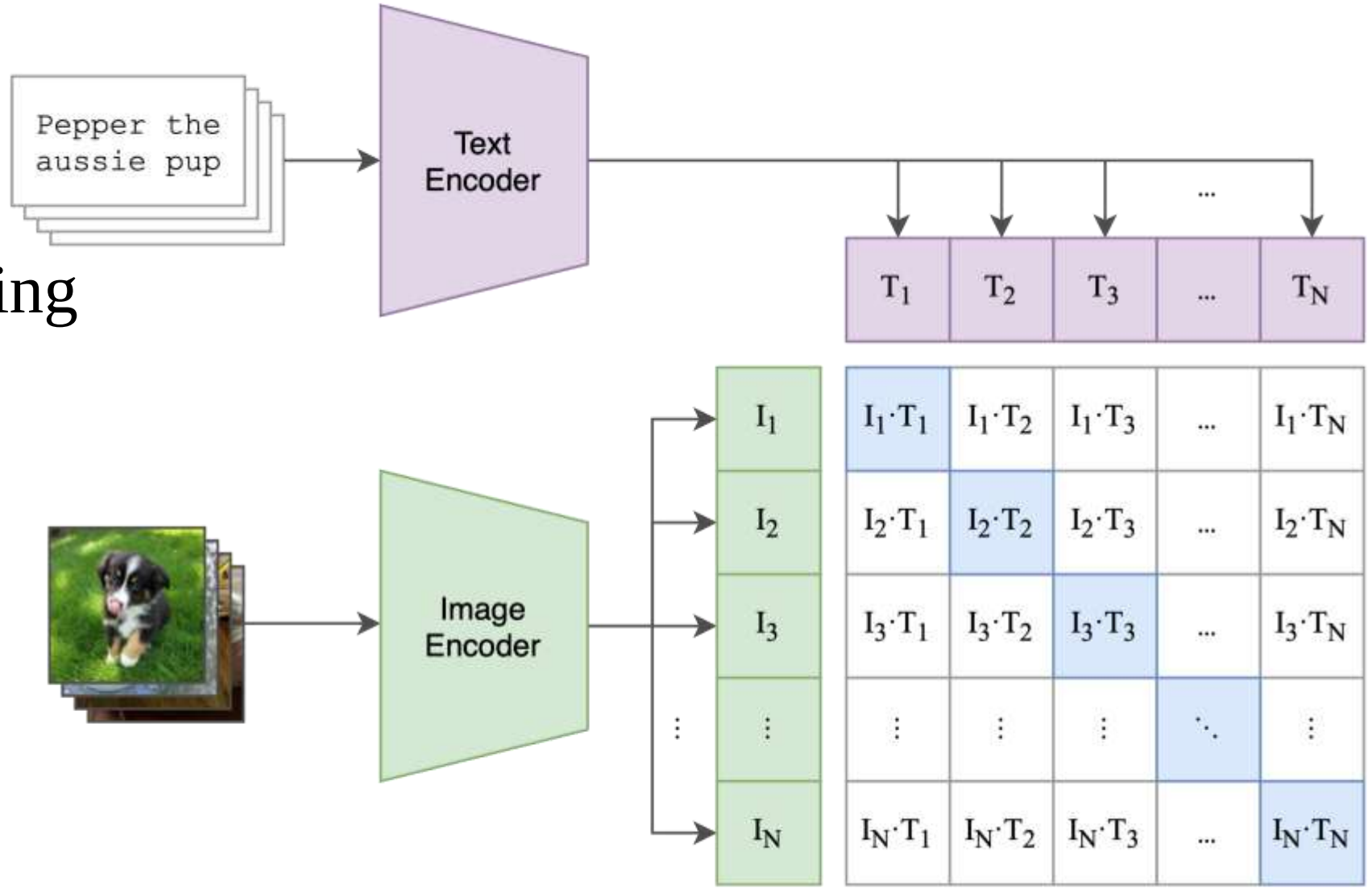


Use/adapt foundations to solve new problems



CLIP [Radford et al., 2021] https://arxiv.org/pdf/2103.00020.pdf

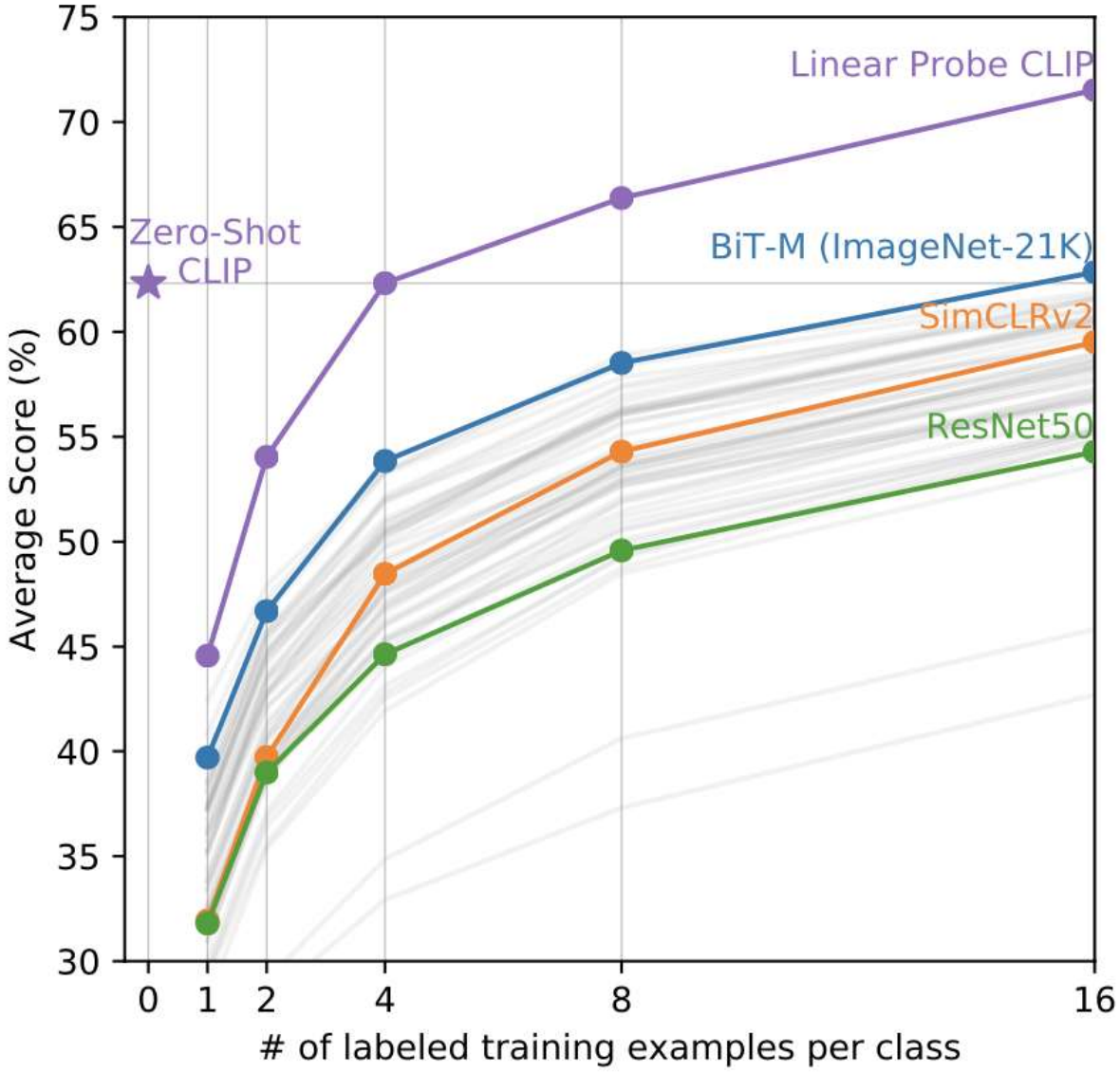
1. Multi-Modal Training: Matching Text-Image representations with Contrastive Learning



[https://openai.com/blog/clip/]

CLIP [Radford et al., 2021] https://arxiv.org/pdf/2103.00020.pdf

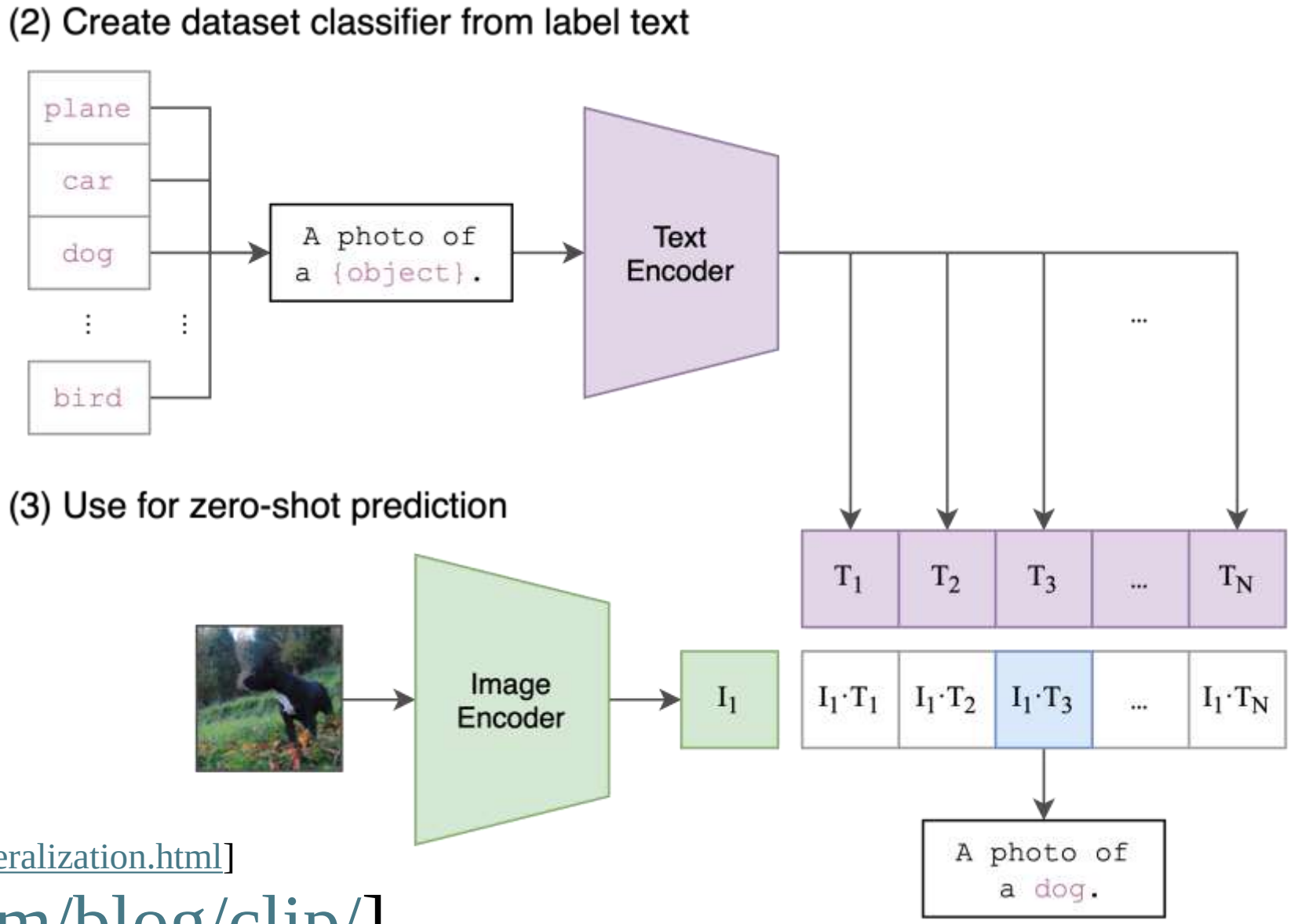
2. Adaptor: Linear classifer on top of image encodings



[https://openai.com/blog/clip/]

CLIP [Radford et al., 2021] https://arxiv.org/pdf/2103.00020.pdf

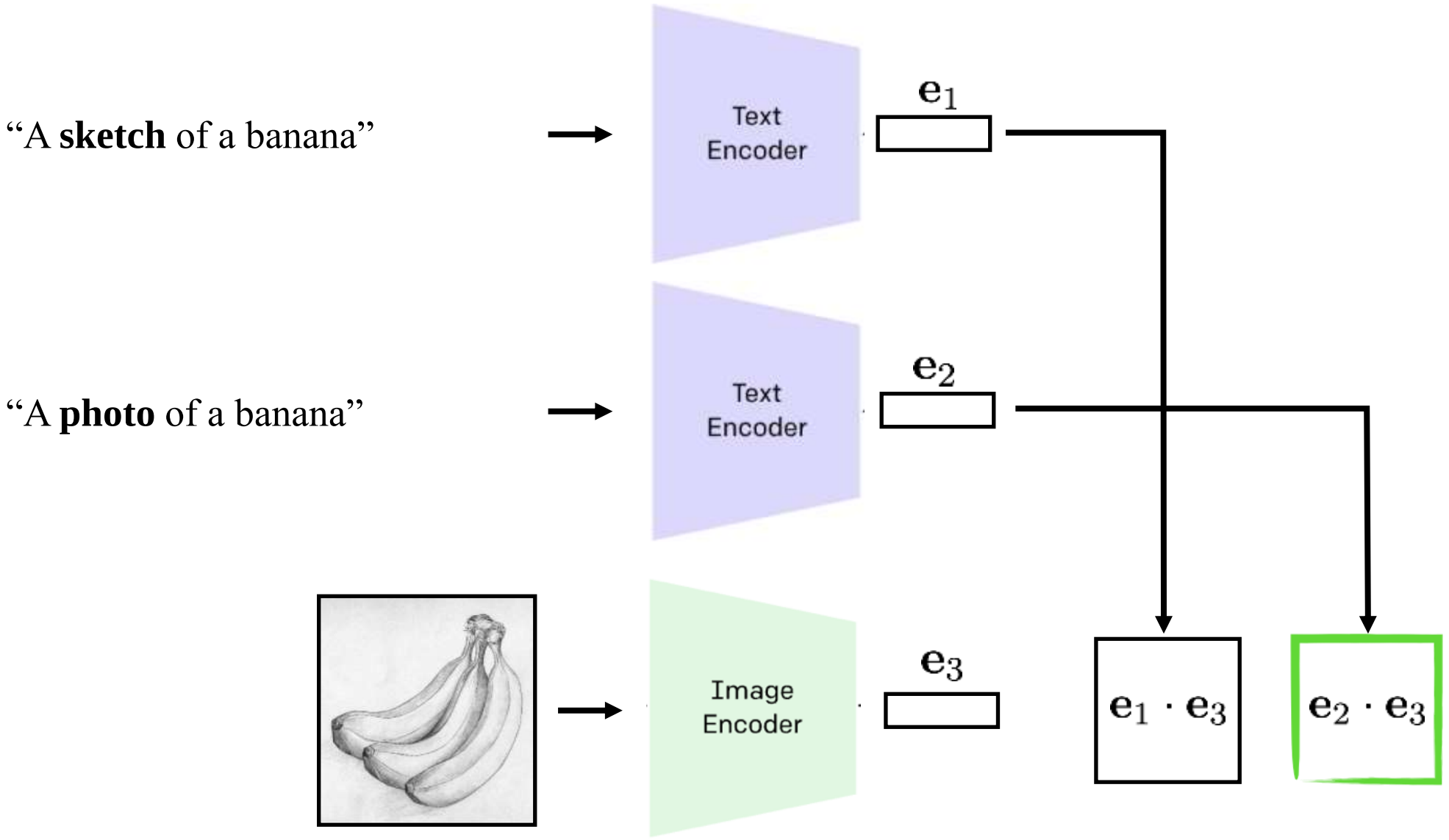
2. Adaptor: Just ask



[https://evjang.com/2021/10/23/generalization.html]

[https://openai.com/blog/clip/]

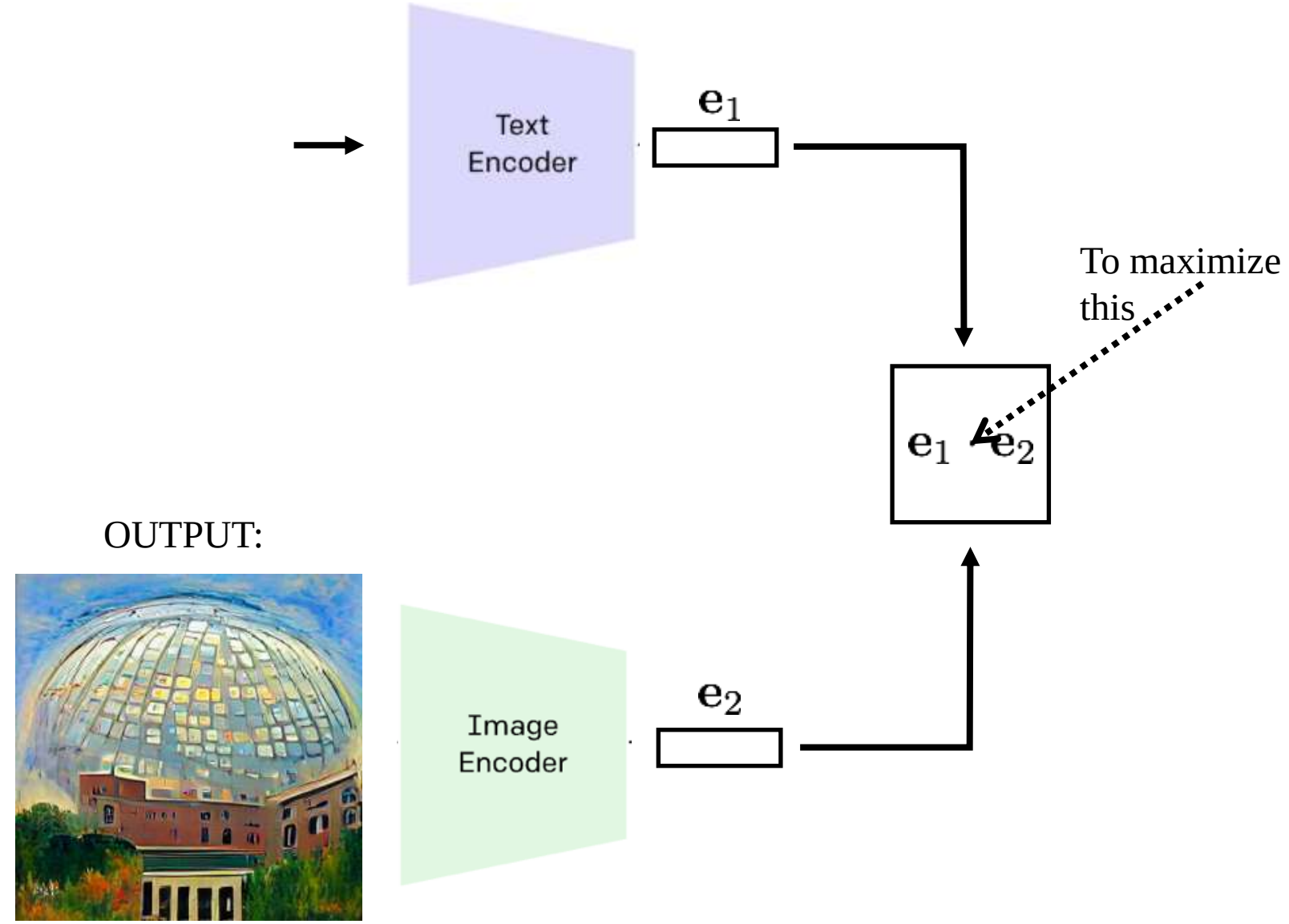
New capabilities by just asking



New capabilities by plugging pretrained models together: CLIP+GAN

INPUT:

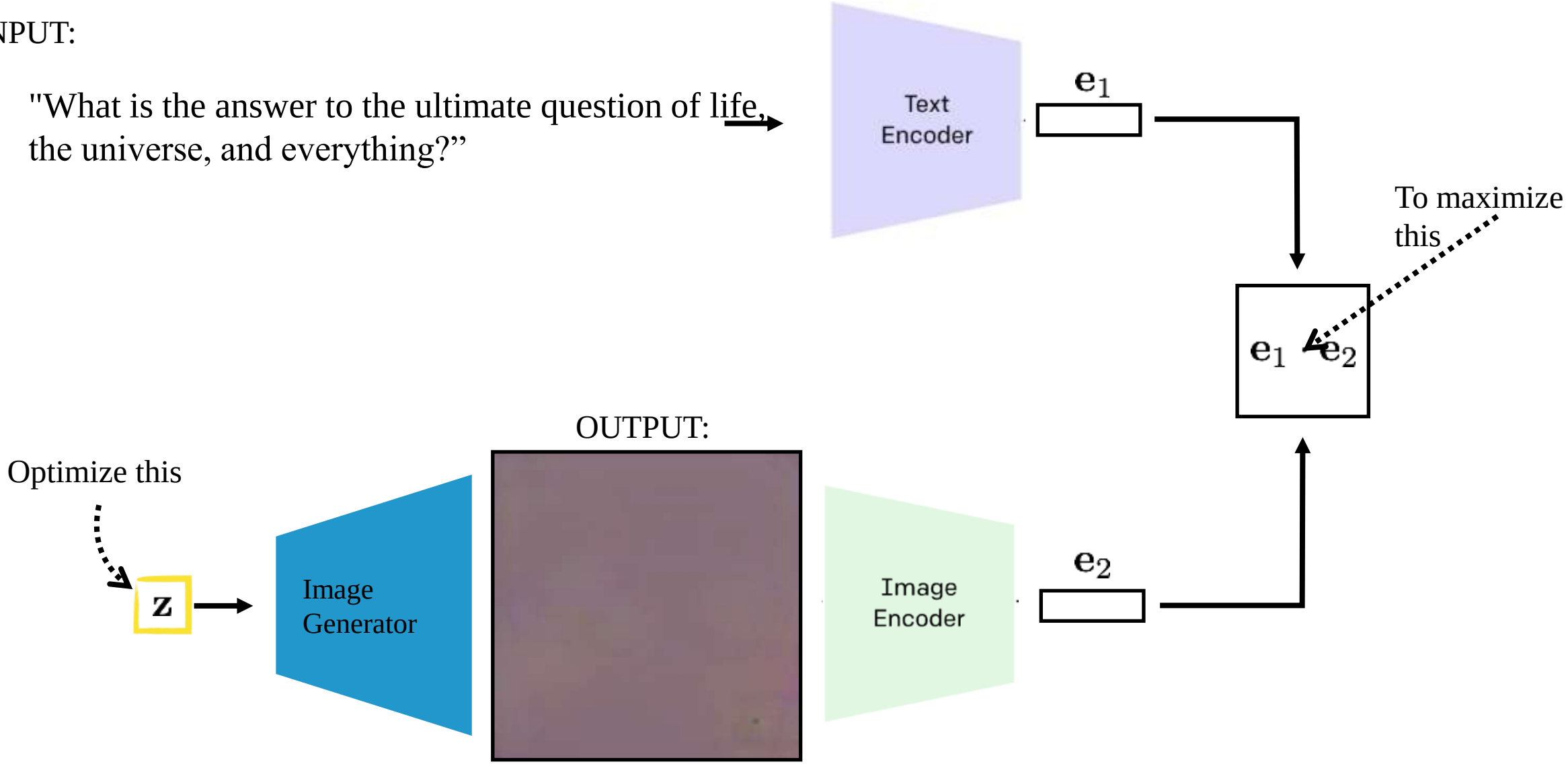
“A Monet painting of the Dome”



New capabilities by plugging pretrained models together: CLIP+GAN

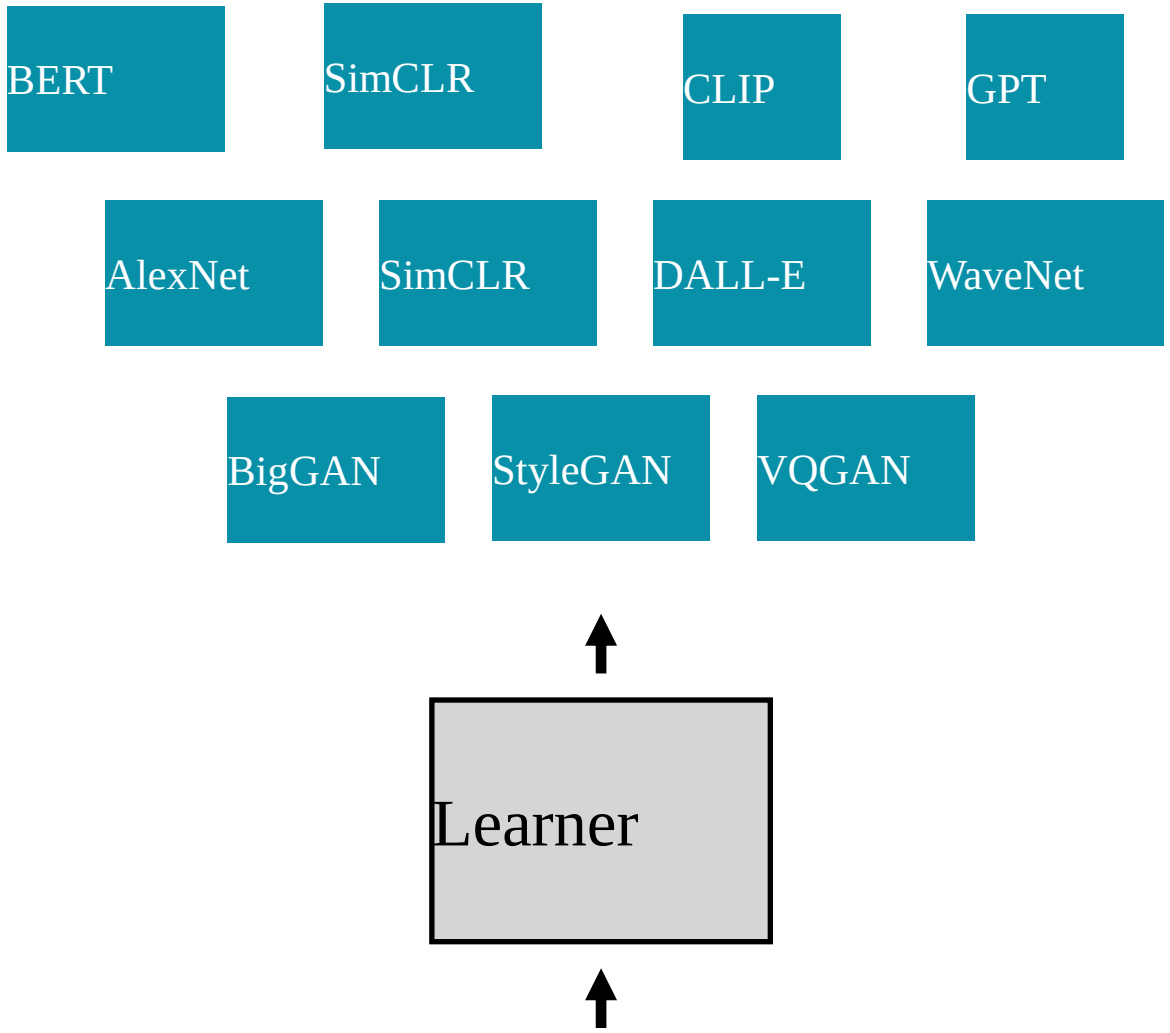
INPUT:

"What is the answer to the ultimate question of life, the universe, and everything?"

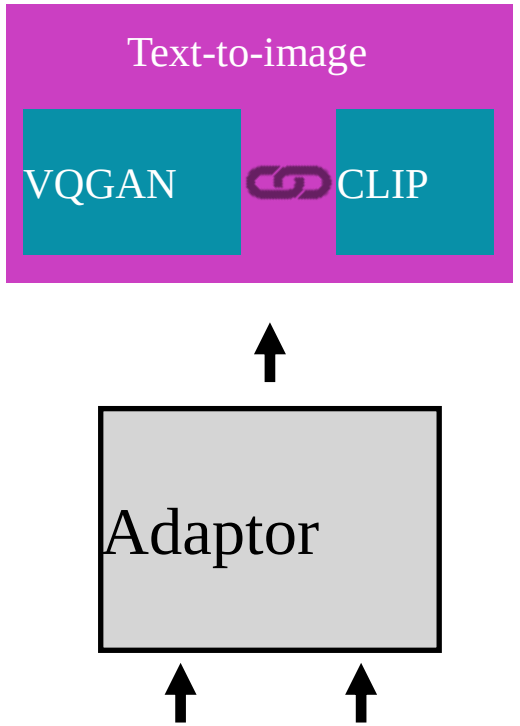


Code: https://colab.research.google.com/drive/1_4PQqzM_0KKytCzWtn-ZPi4cCa5bwK2F?usp=

Learn foundation models



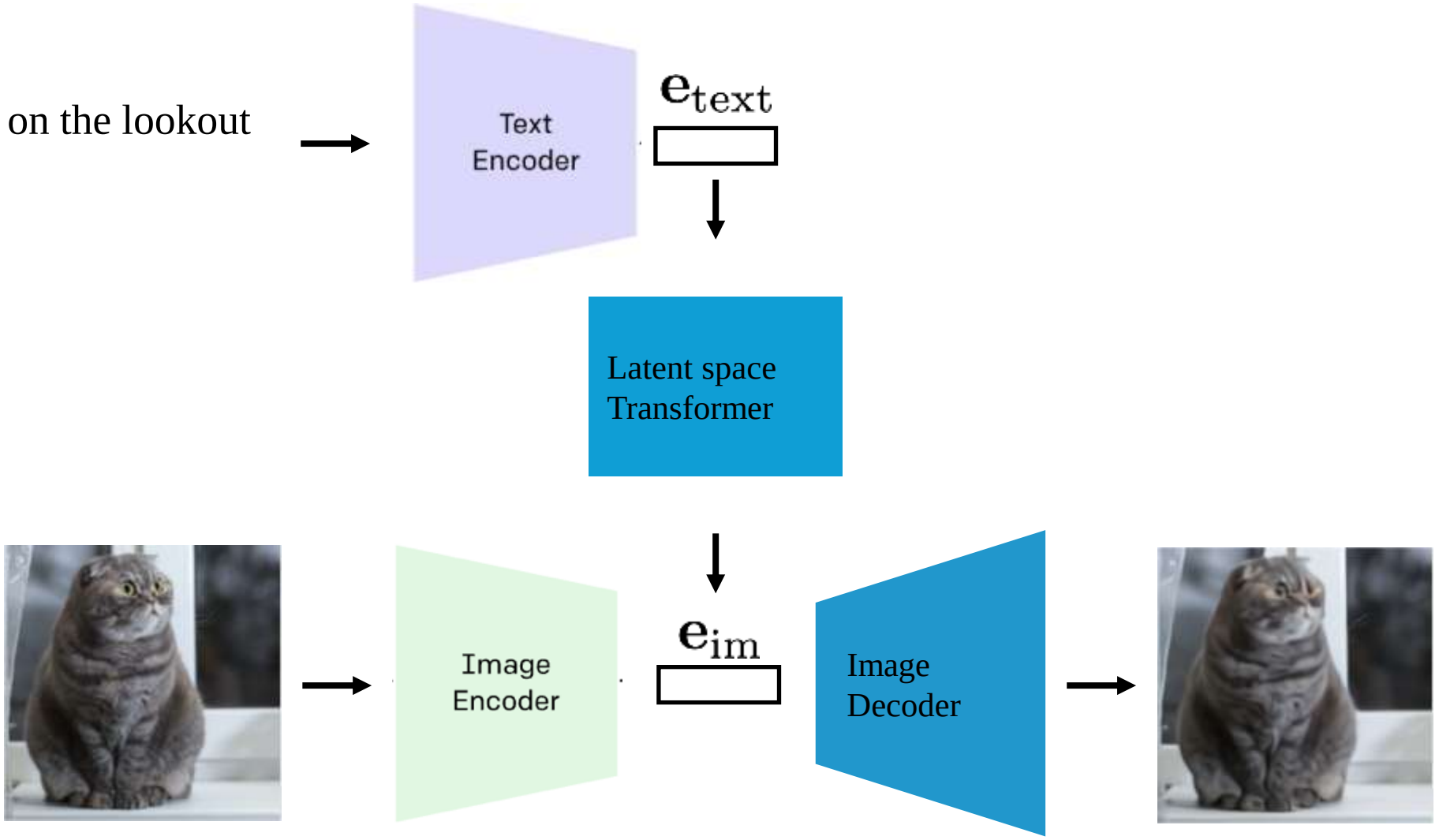
Use/adapt foundations to solve new problems



DALL-E [Ramesh et al. 2021] https://arxiv.org/pdf/2102.12092.pdf https://openai.com/blog/dall-e/

INPUT:

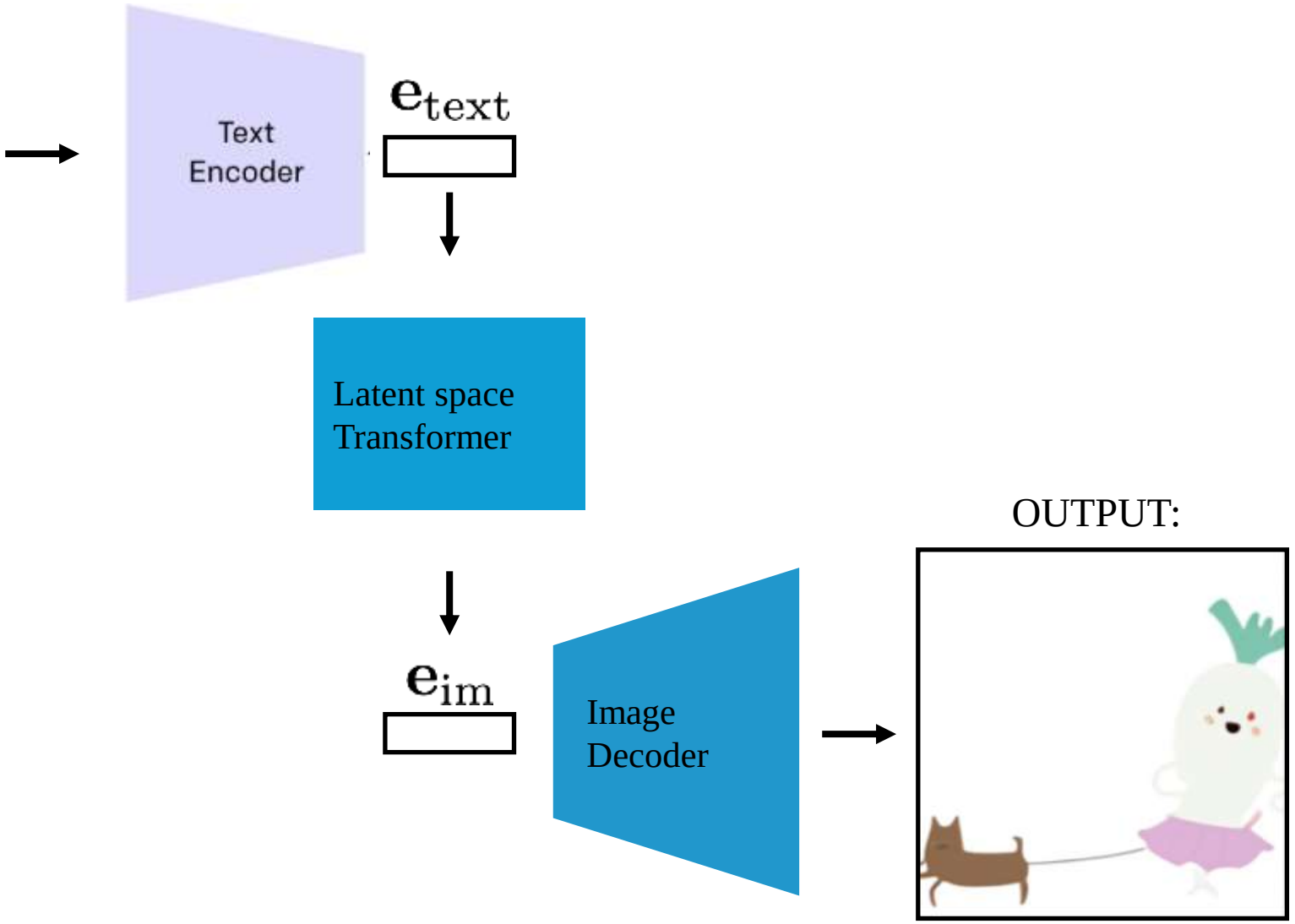
"A wide-eyed cat on the lookout for food"



Text-to-image translation

INPUT:

"An illustration of a baby daikon radish in a tutu walking a dog"



New capabilities by just asking

TEXT PROMPT

an armchair in the shape of an avocado. an armchair imitating an avocado.

AI-GENERATED IMAGES

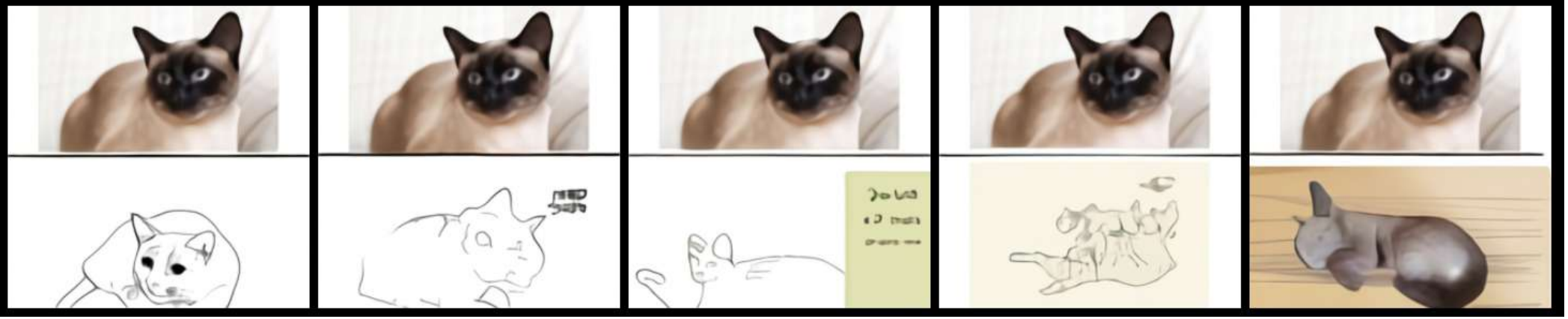
New capabilities by just

TEXT PROMPT

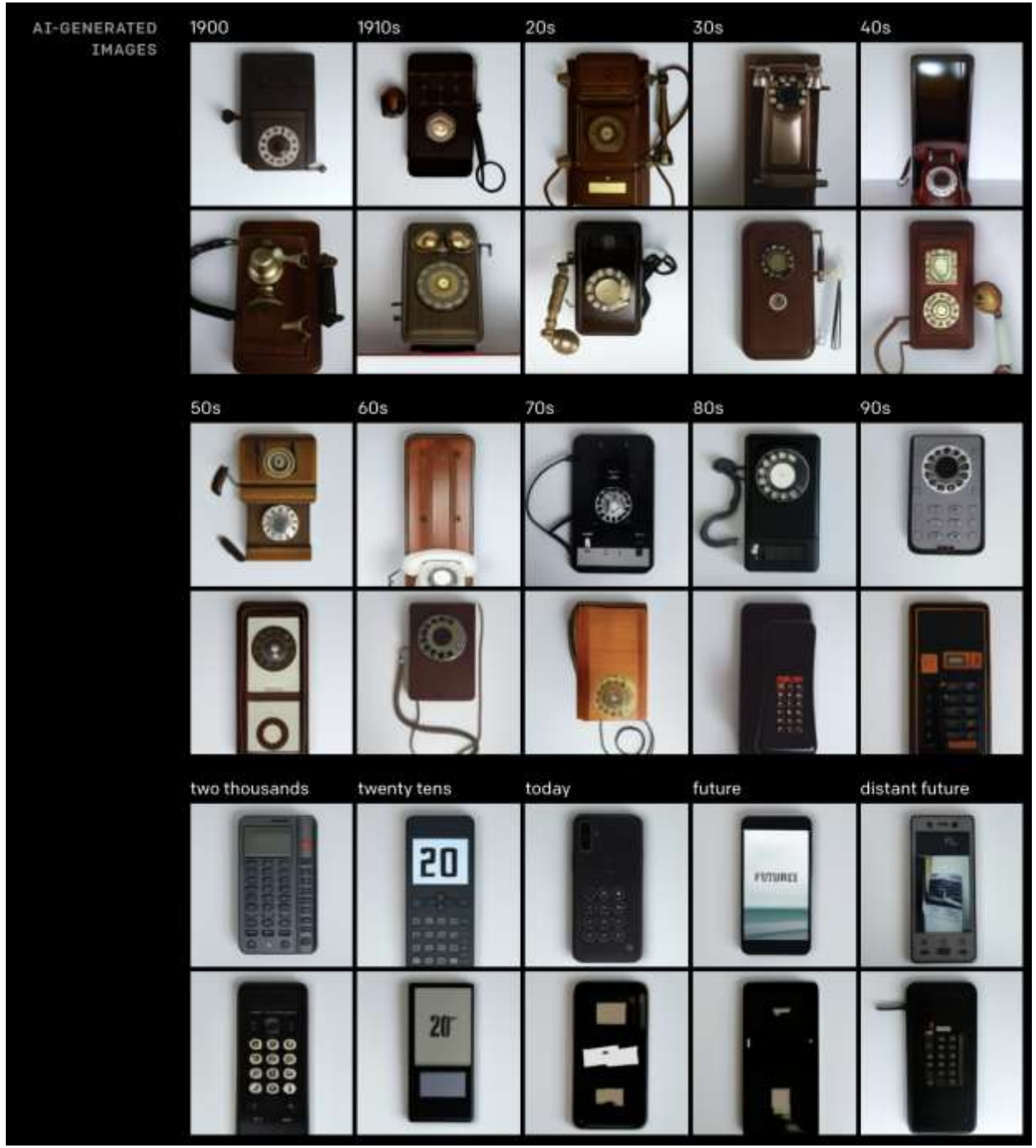
the exact same cat on the top as a sketch on the bottom

AI-GENERATED IMAGES

AI-GENERATED IMAGES

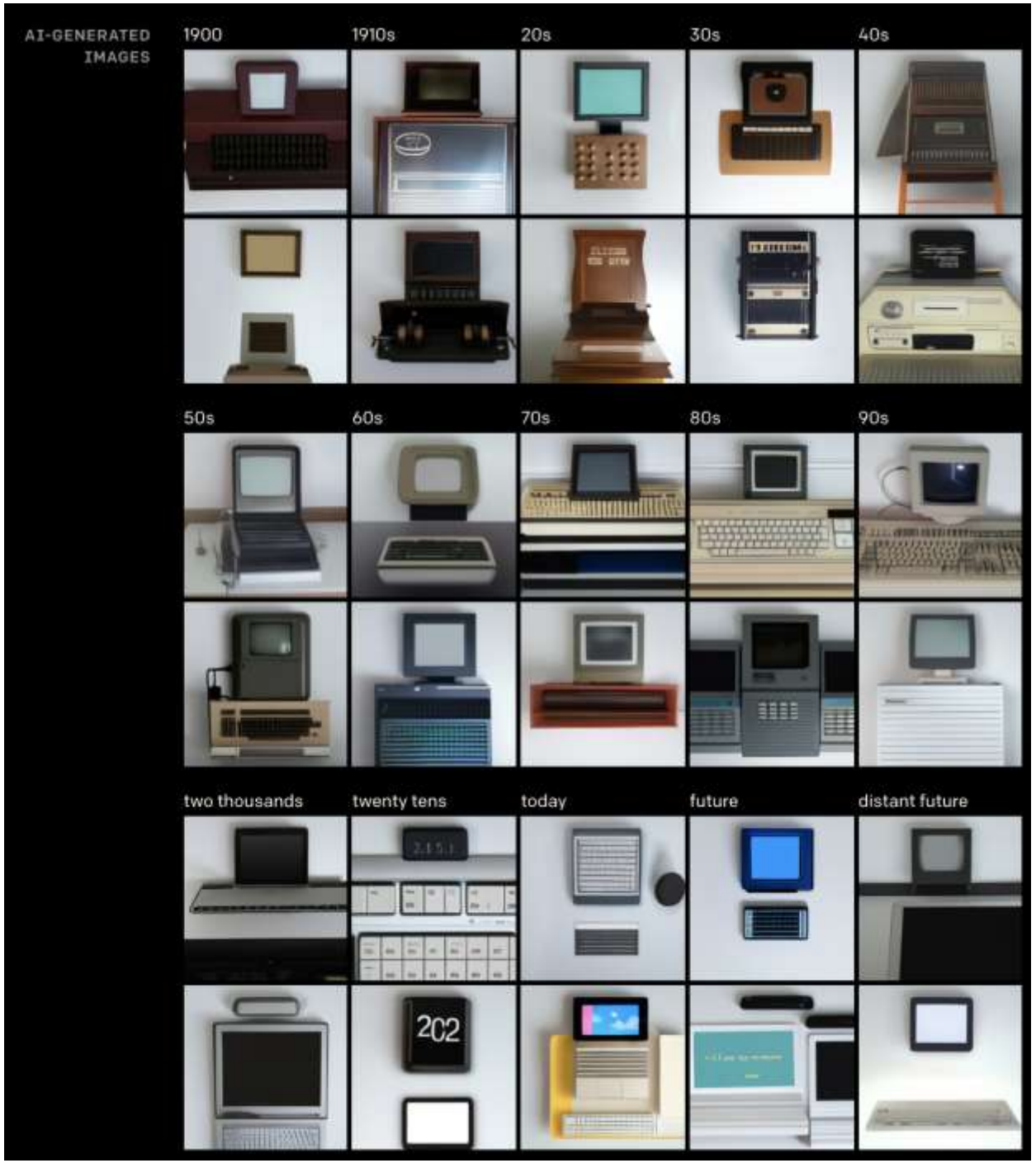


TEXT PROMPT a photo of a phone from the ...

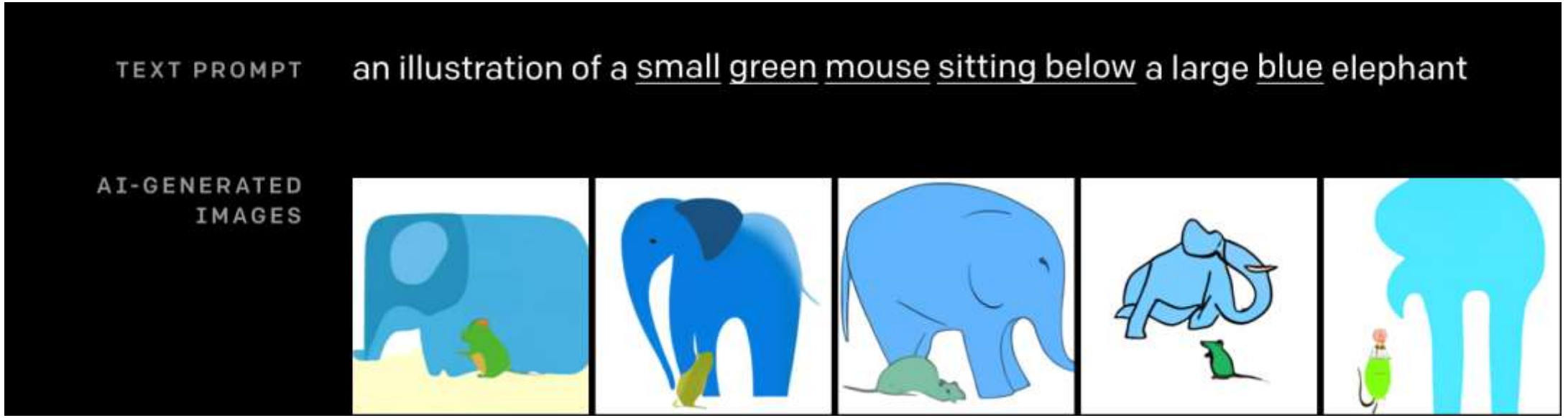


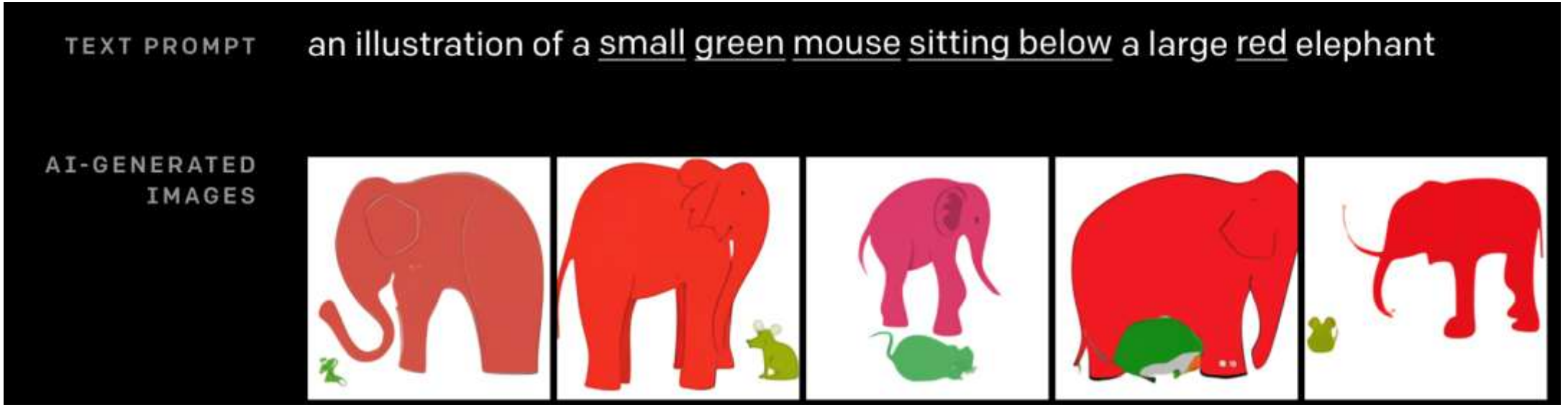
TEXT PROMPT a photo of a computer from the ...

AI-GENERATED IMAGES

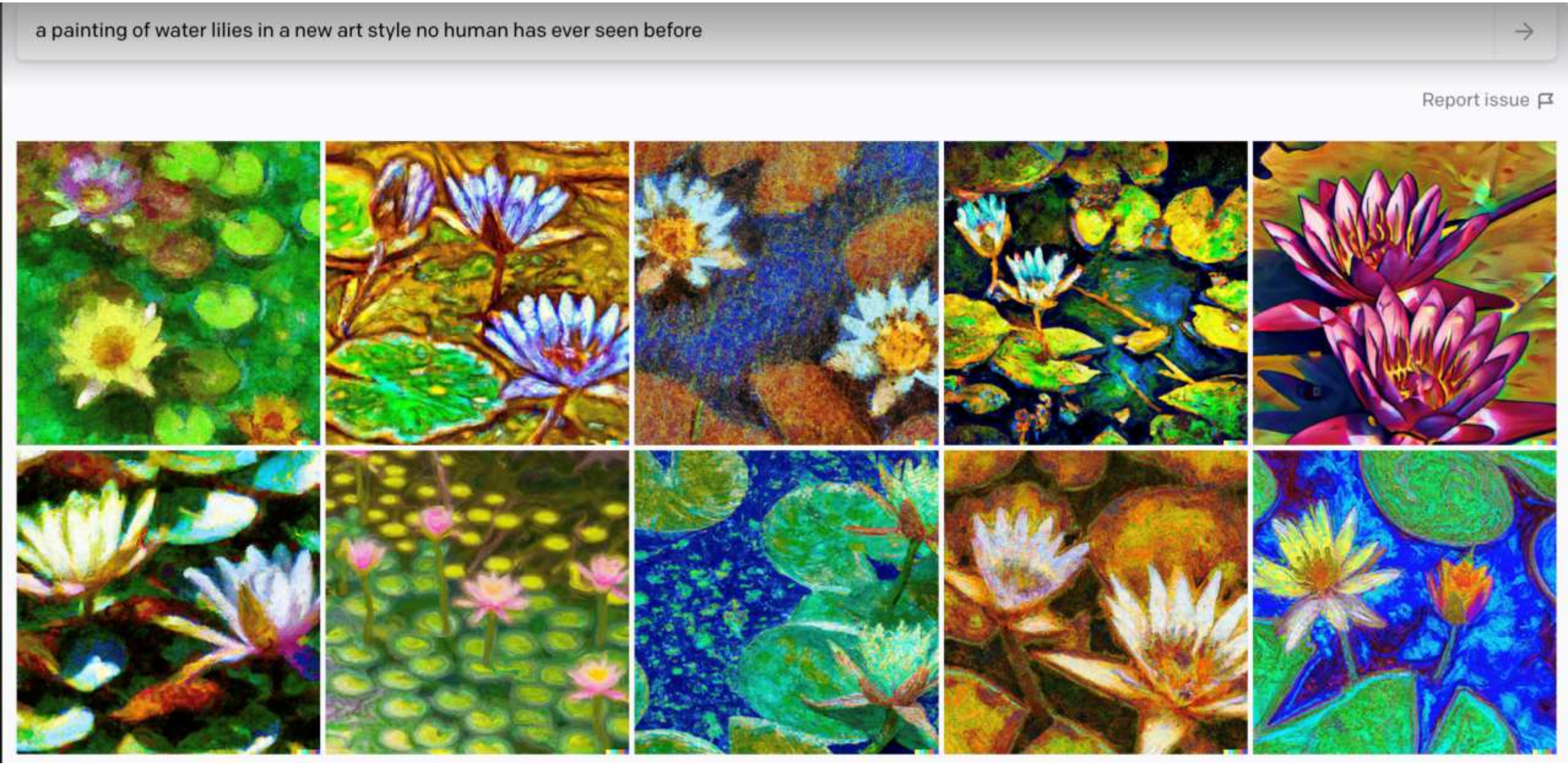


DALL-E [Ramesh et al. 2021] https://arxiv.org/pdf/2102.12092.pdf https://openai.com/blog/dall-e/

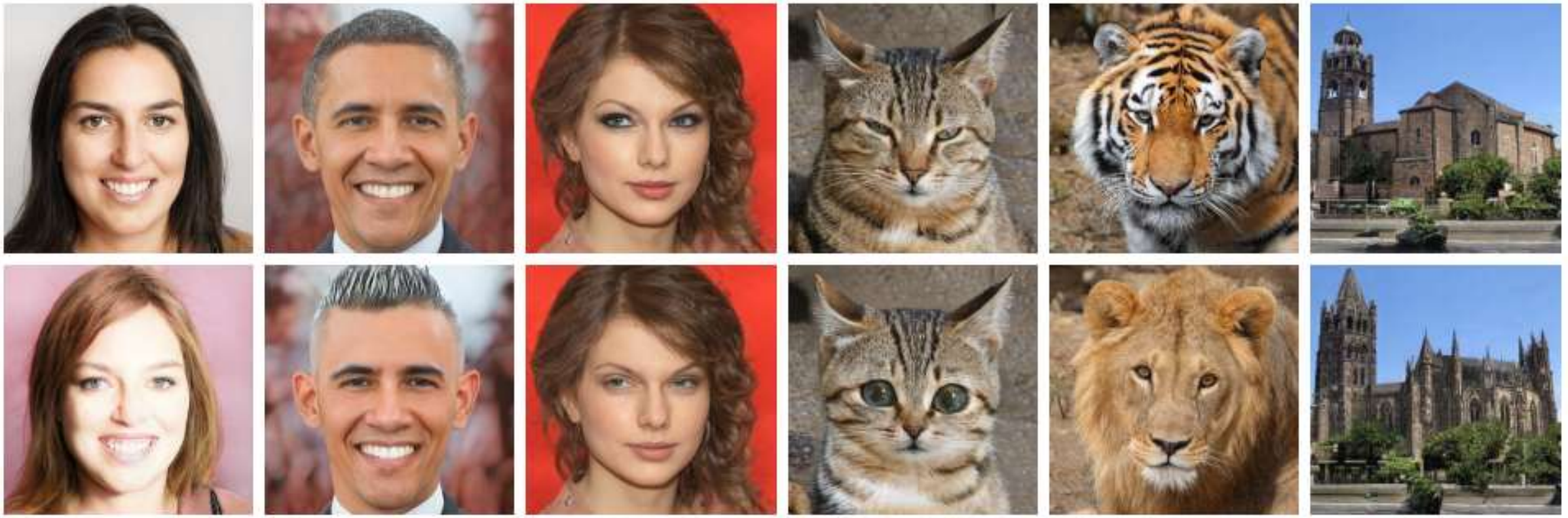




DALL-E 2


a painting of water lilies in a new art style no human has ever seen before
Report issue

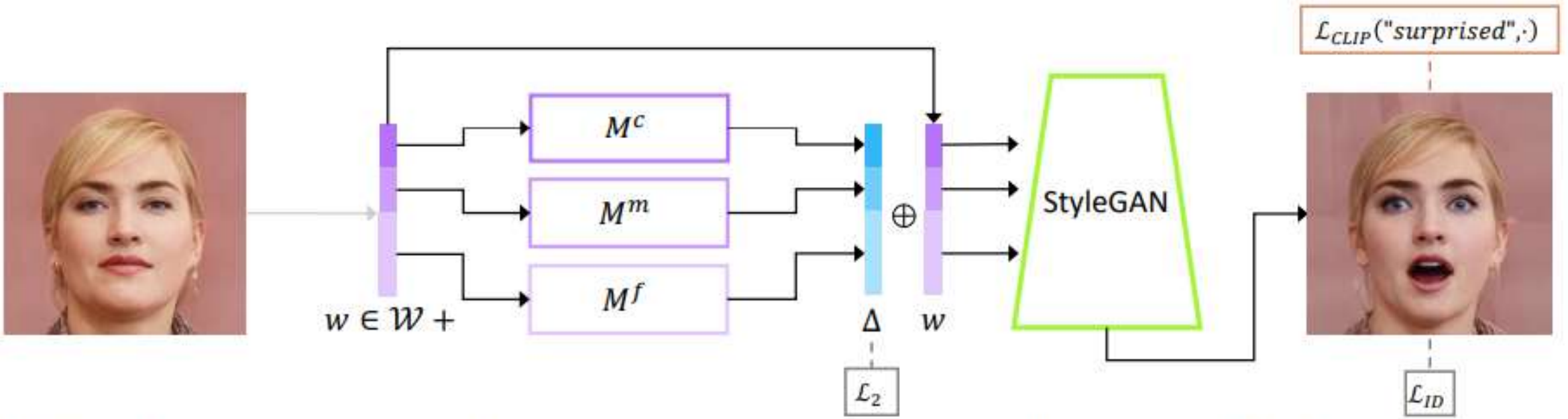
StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery



"Emma Stone" "Mohawk hairstyle" "Without makeup" "Cute cat" "Lion" "Gothic church"

https://arxiv.org/pdf/2103.17249

StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery



	pre-proc.	train time	infer. time	input image dependent	latent space
optimizer	–	–	98 sec	yes	$\mathcal{W}+$
mapper	–	10 – 12h	75 ms	yes	$\mathcal{W}+$
global dir.	4h	–	72 ms	no	$\mathcal{S}$

https://arxiv.org/pdf/2103.17249

StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery

Input | "Beyonce" (0.004, 0) | "A woman without makeup" (0.008, 0.005) | "Elsa from Frozen" (0.004, 0)

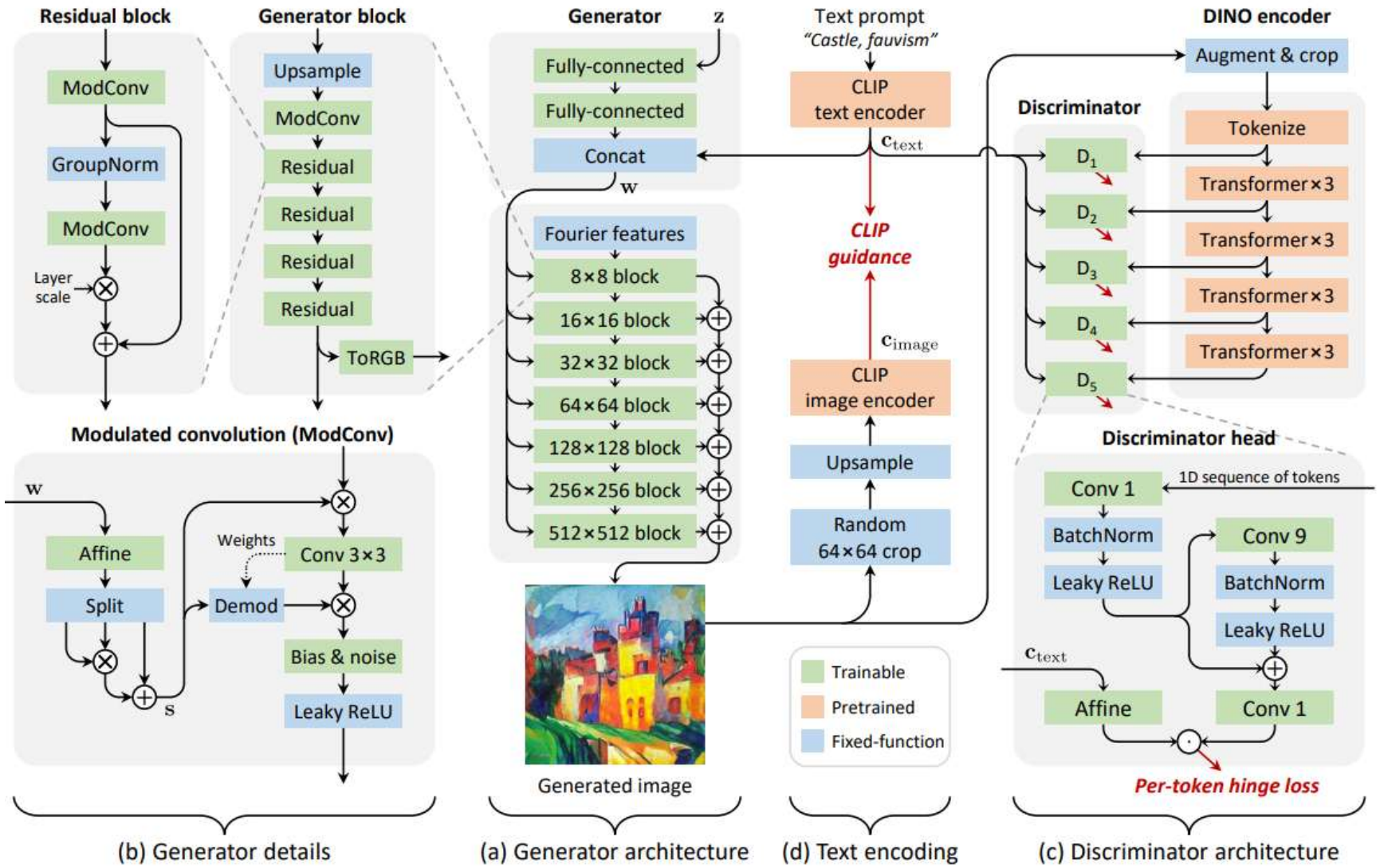
Input | "A man with a beard" (0.008, 0.005) | "A blonde man" (0.008, 0.005) | "Donald Trump" (0.0025, 0)

https://arxiv.org/pdf/2103.17249

StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery

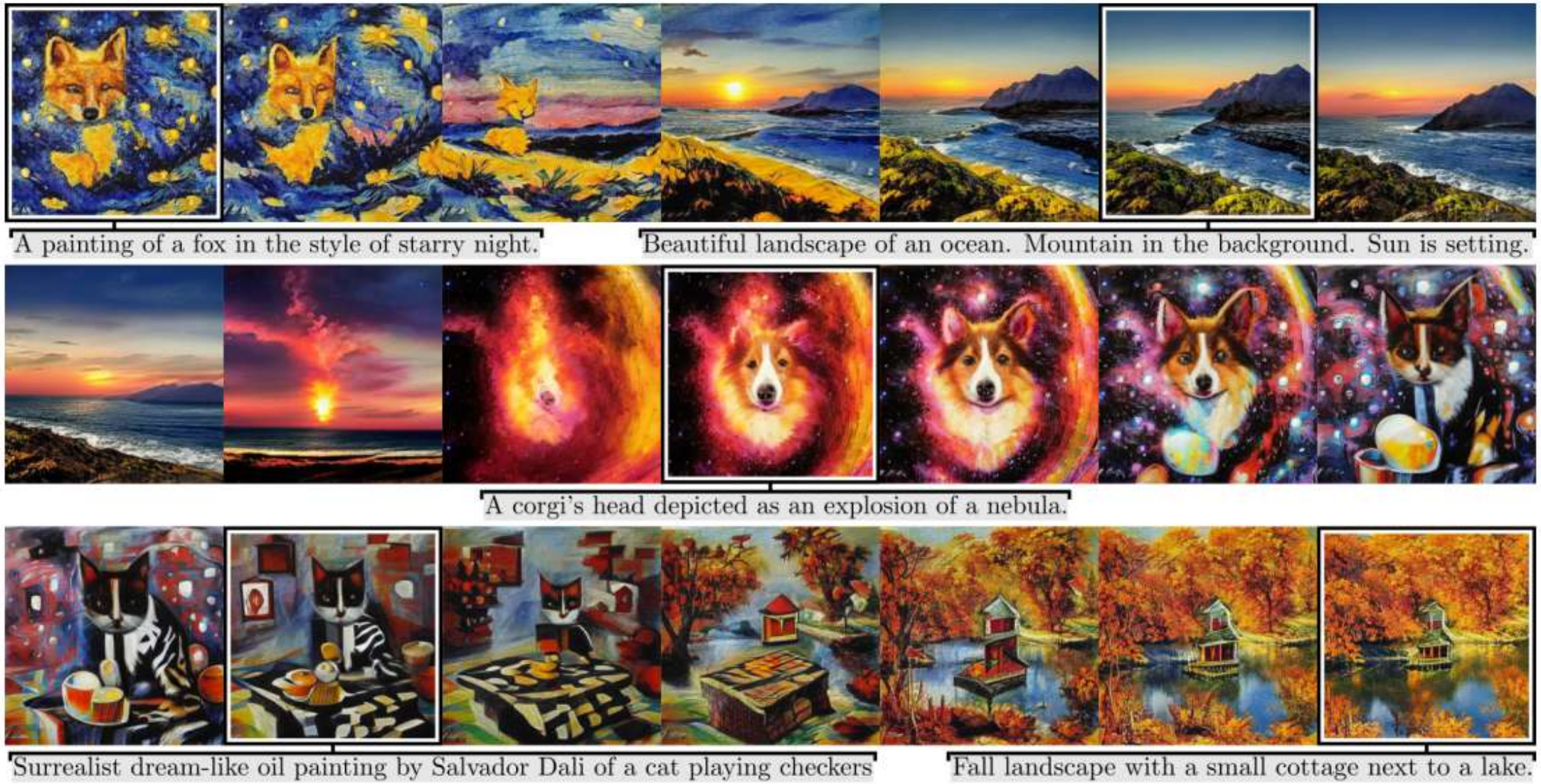
https://arxiv.org/pdf/2103.17249

StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis



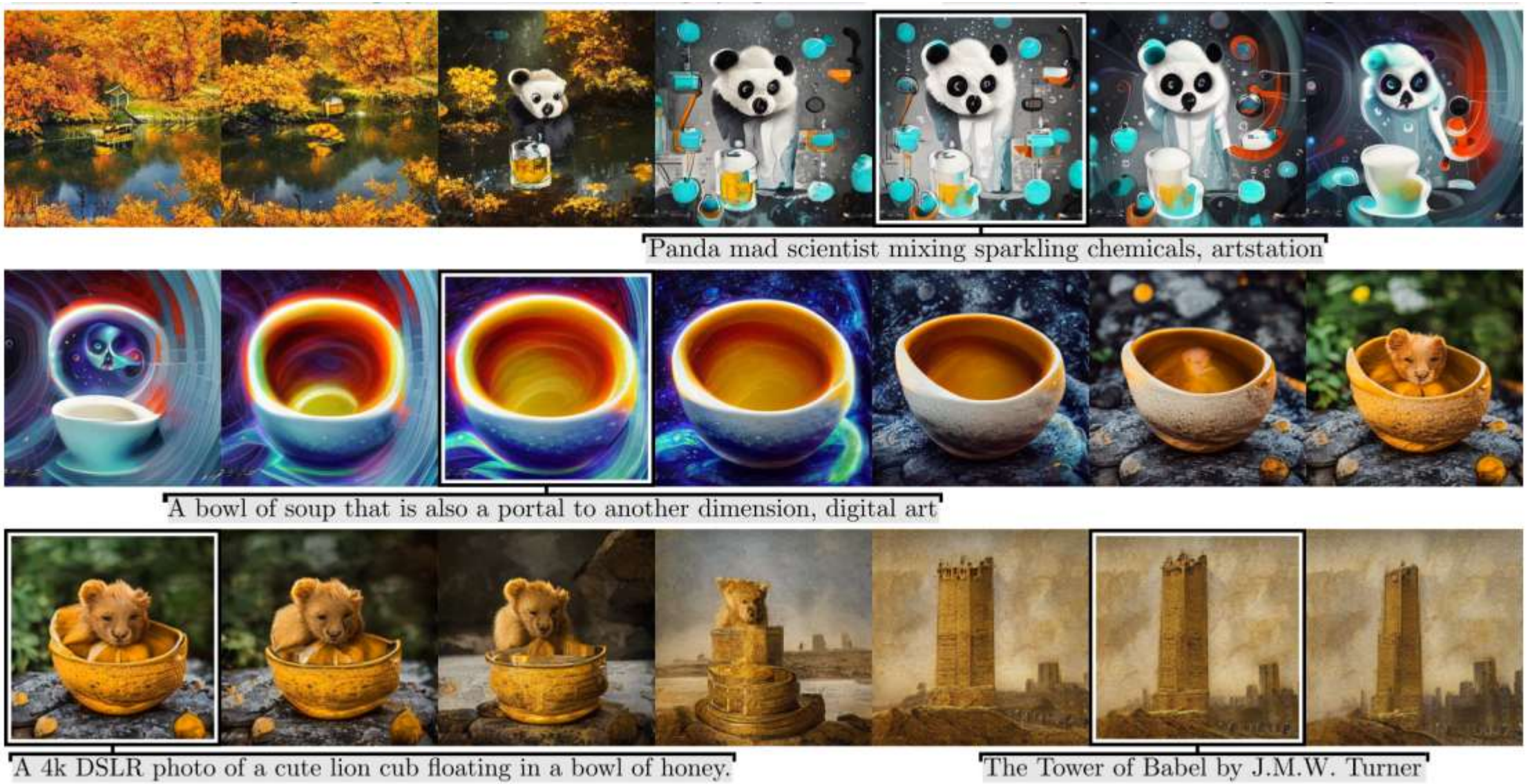
https://arxiv.org/pdf/2301.09515

StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis



https://arxiv.org/pdf/2301.09515

StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis



https://arxiv.org/pdf/2301.09515

StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis

"a red cube on a blue cube" | "astronaut, child's drawing" | "a sign that says deep learning"

Figure 9. **Failure cases.** StyleGAN-T can struggle to bind attributes to objects, and to produce coherent text.

	Zero-shot FID_{30k} ↓	CLIP score ↑
StyleGAN-XL	51.88	5.58
New generator	45.10	6.02
New discriminator	26.77	9.78
$\mathcal{L}_{\text{CLIP}}$	**20.52**	**11.72**

https://arxiv.org/pdf/2301.09515

GigaGAN: Scaling up GANs for Text-to-Image Synthesis

A portrait of a human growing colorful flowers from her hair. Hyperrealistic oil painting. Intricate details.

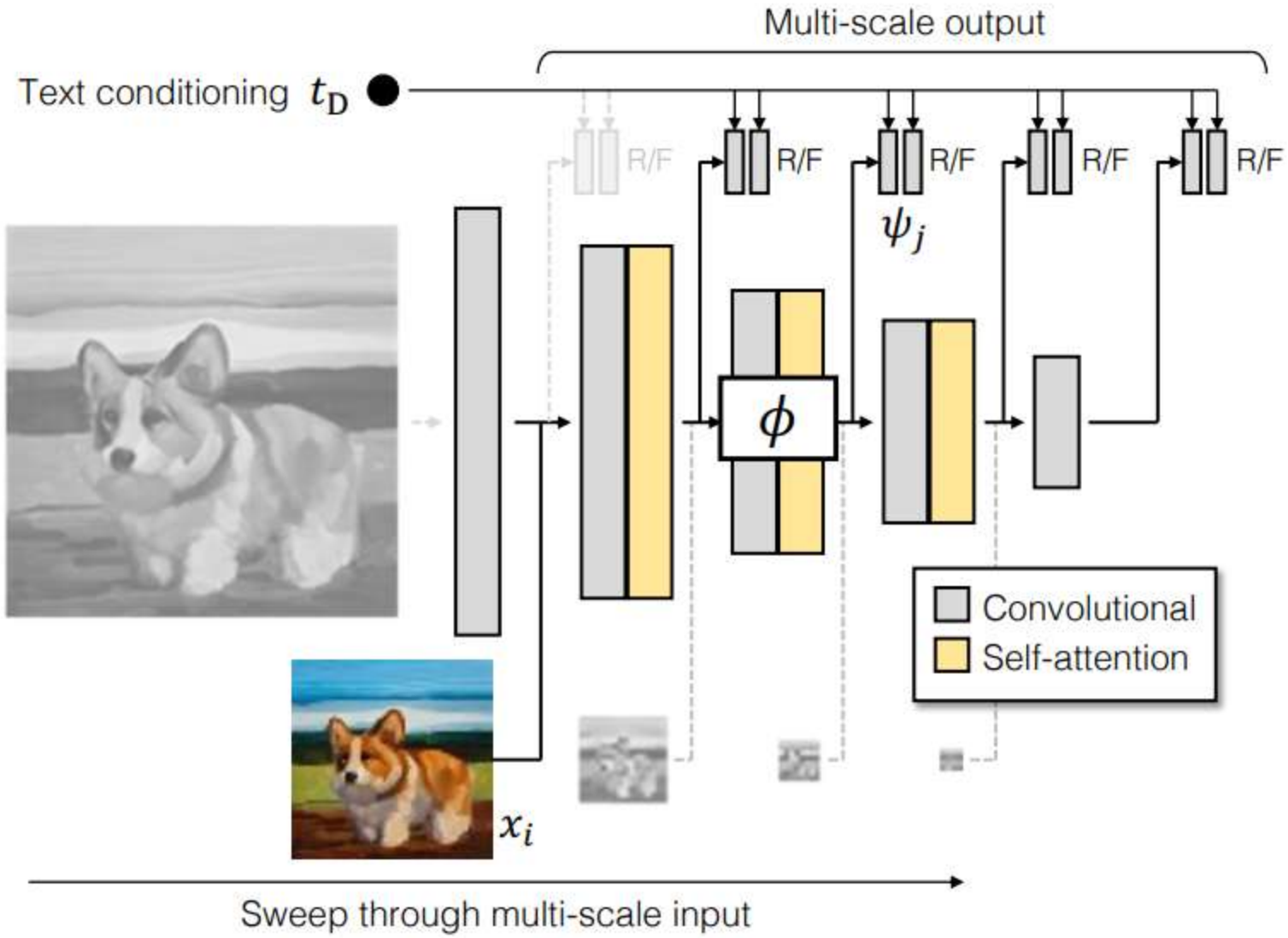
A golden luxury motorcycle parked at the King's palace. 35mm f/4.5.

a cute magical flying maltipoo at light speed, fantasy concept art, bokeh, wide sky

It is orders of magnitude faster at inference time, taking only 0.13 seconds to synthesize a 512px image. Second, it can synthesize high-resolution images, for example, 16-megapixel images in 3.66 seconds.

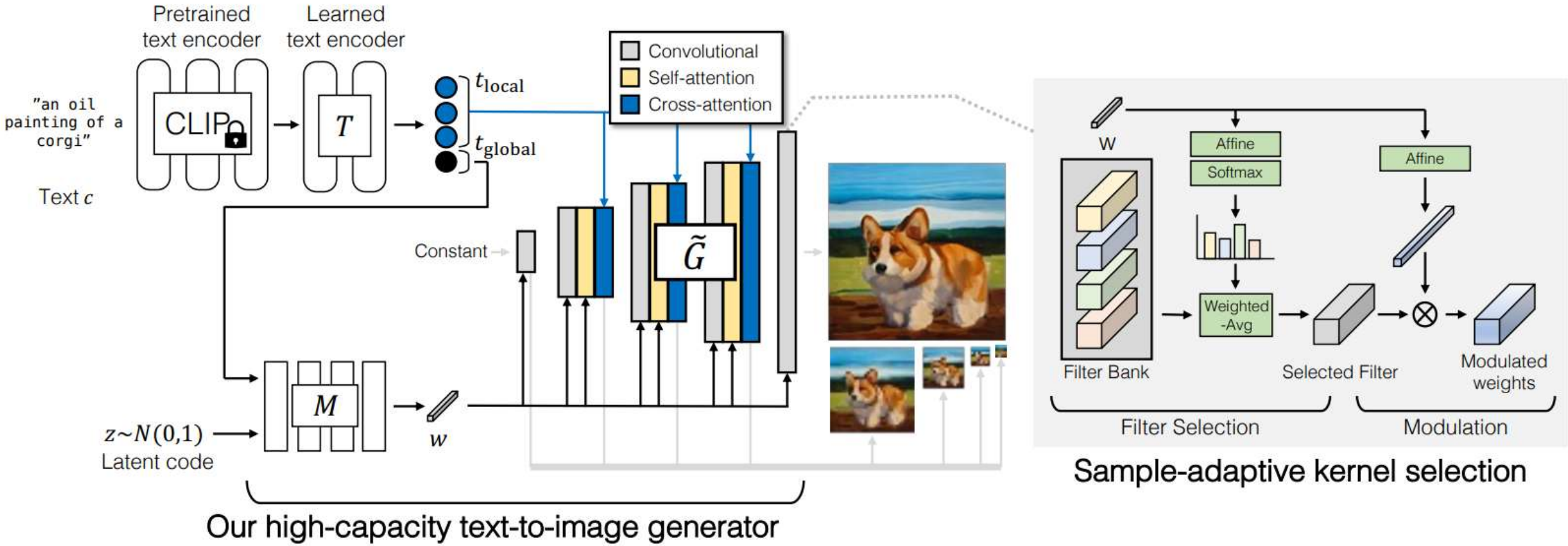
https://arxiv.org/pdf/2303.05511

GigaGAN: Scaling up GANs for Text-to-Image Synthesis



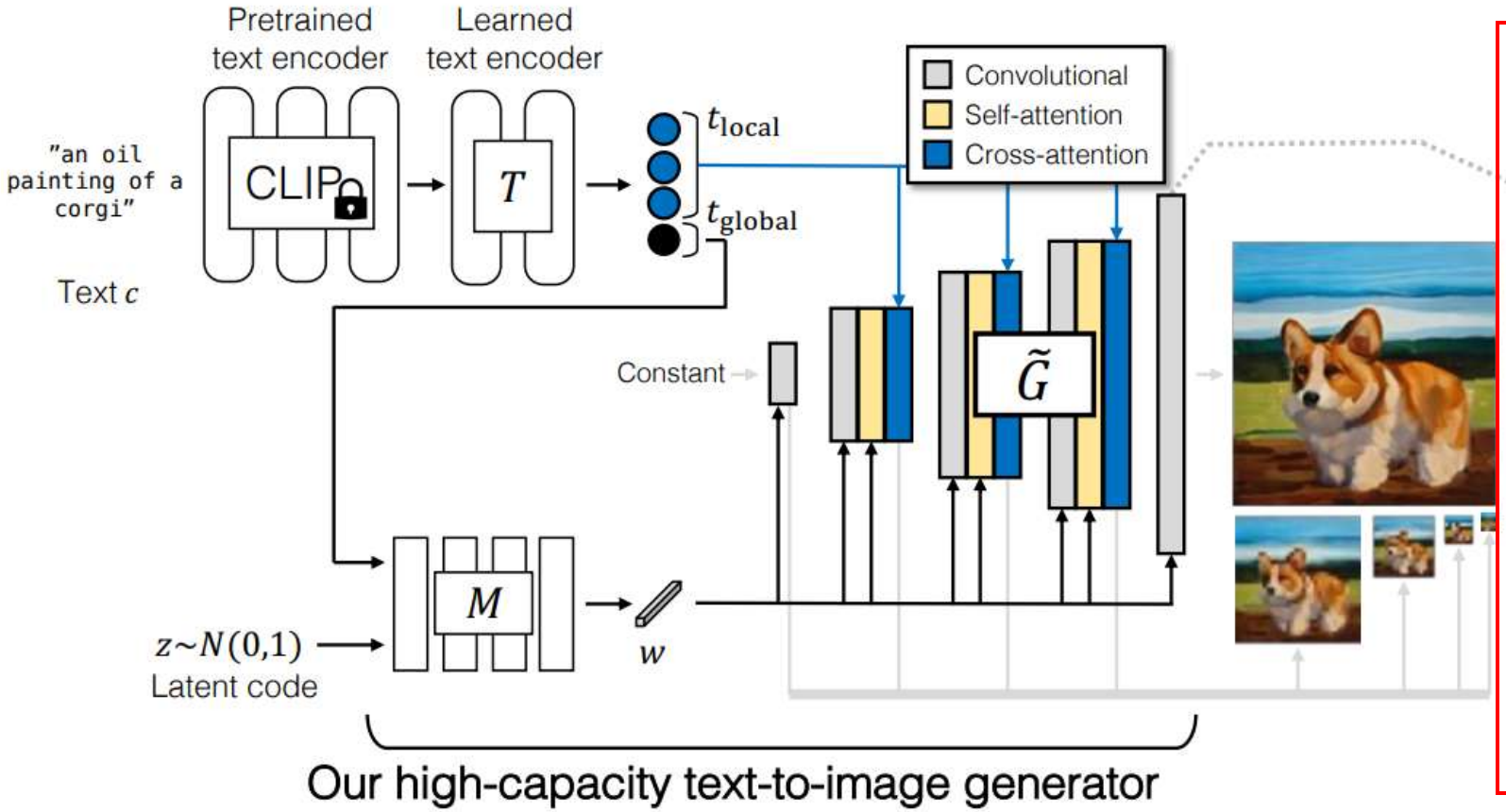
https://arxiv.org/pdf/2303.05511

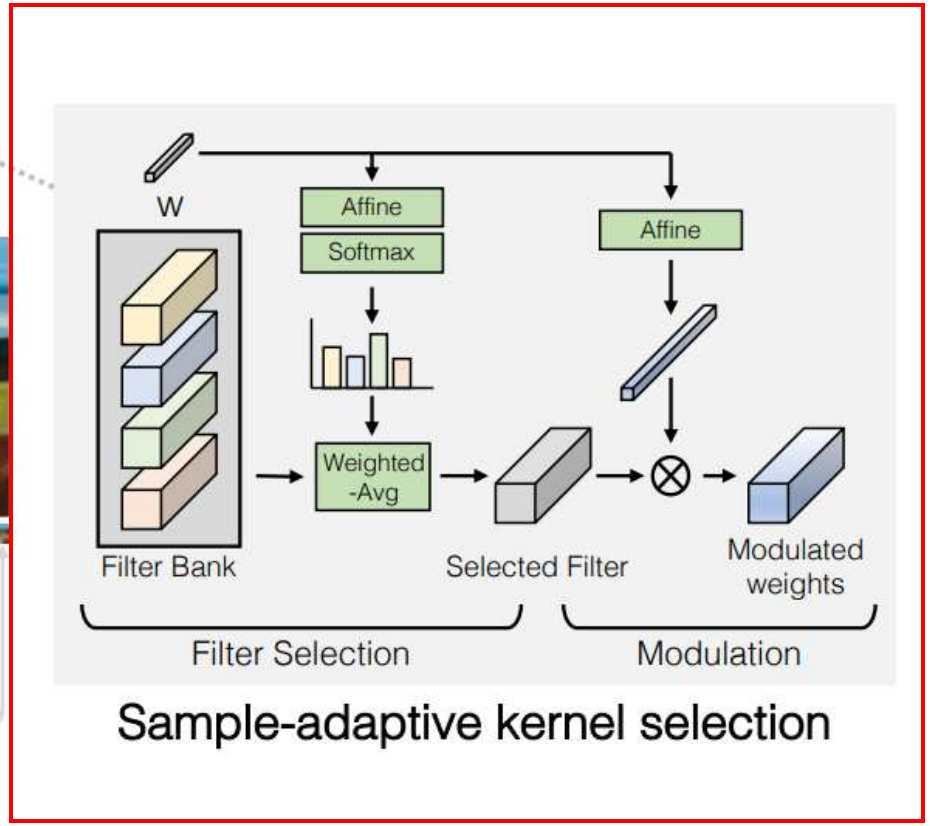
GigaGAN: Scaling up GANs for Text-to-Image Synthesis



https://arxiv.org/pdf/2303.05511

GigaGAN: Scaling up GANs for Text-to-Image Synthesis





https://arxiv.org/pdf/2303.05511

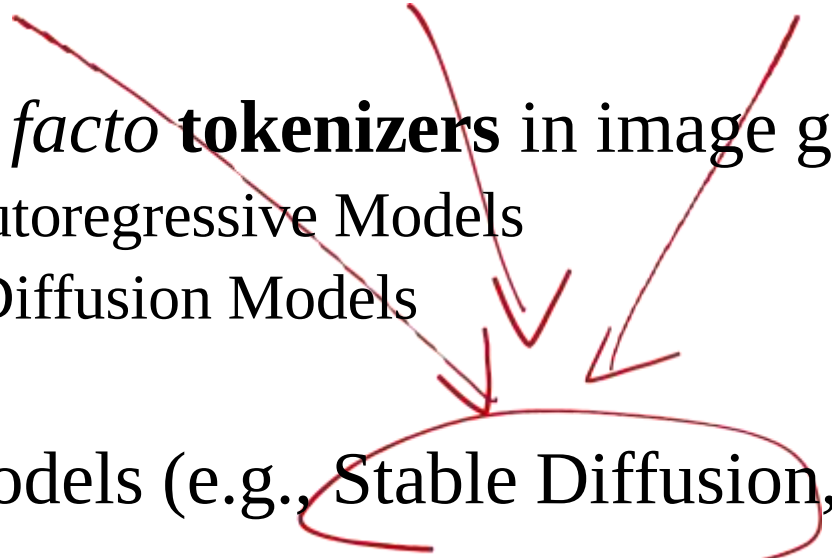
Discussion

- To be precise: **VQ-GAN** = **VQ-VAE** + Adv Loss + Perceptual Loss

- w/o VQ, it's **VAE** + Adv Loss + Perceptual Loss

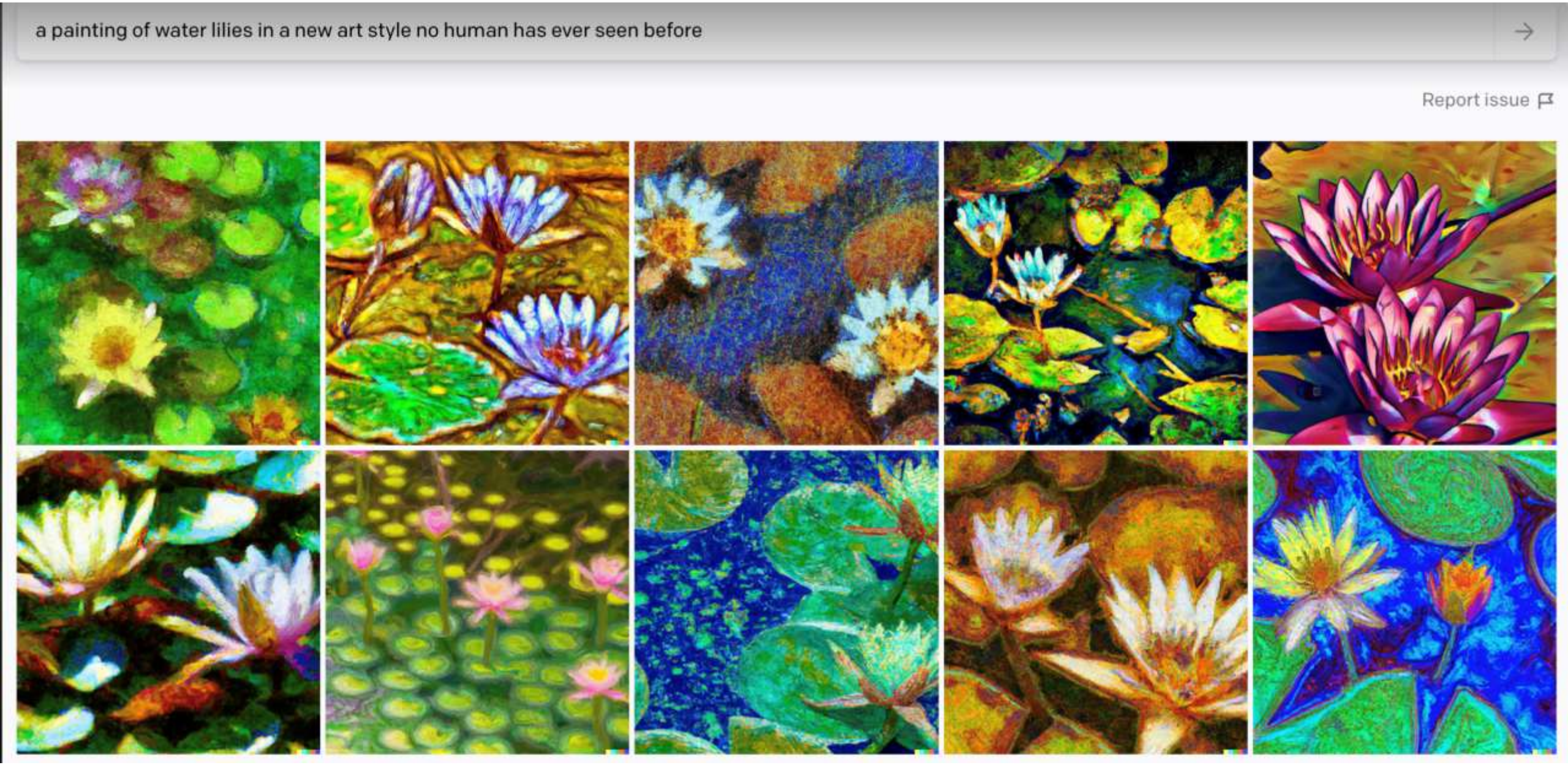
- Both are the *de facto* **tokenizers** in image generation
  - w/ VQ: e.g., Autoregressive Models
  - w/o VQ: e.g., Diffusion Models

- Commercial models (e.g., Stable Diffusion, Sora) use these tokenizers



It involves everything!

DALL-E 2


a painting of water lilies in a new art style no human has ever seen before
Report issue