Self-Attention Generative Adversarial Networks (SAGAN)

(a.k.a. Finally, Here Are Some Dogs with Separated Legs)

Han Zhang*
Rutgers University

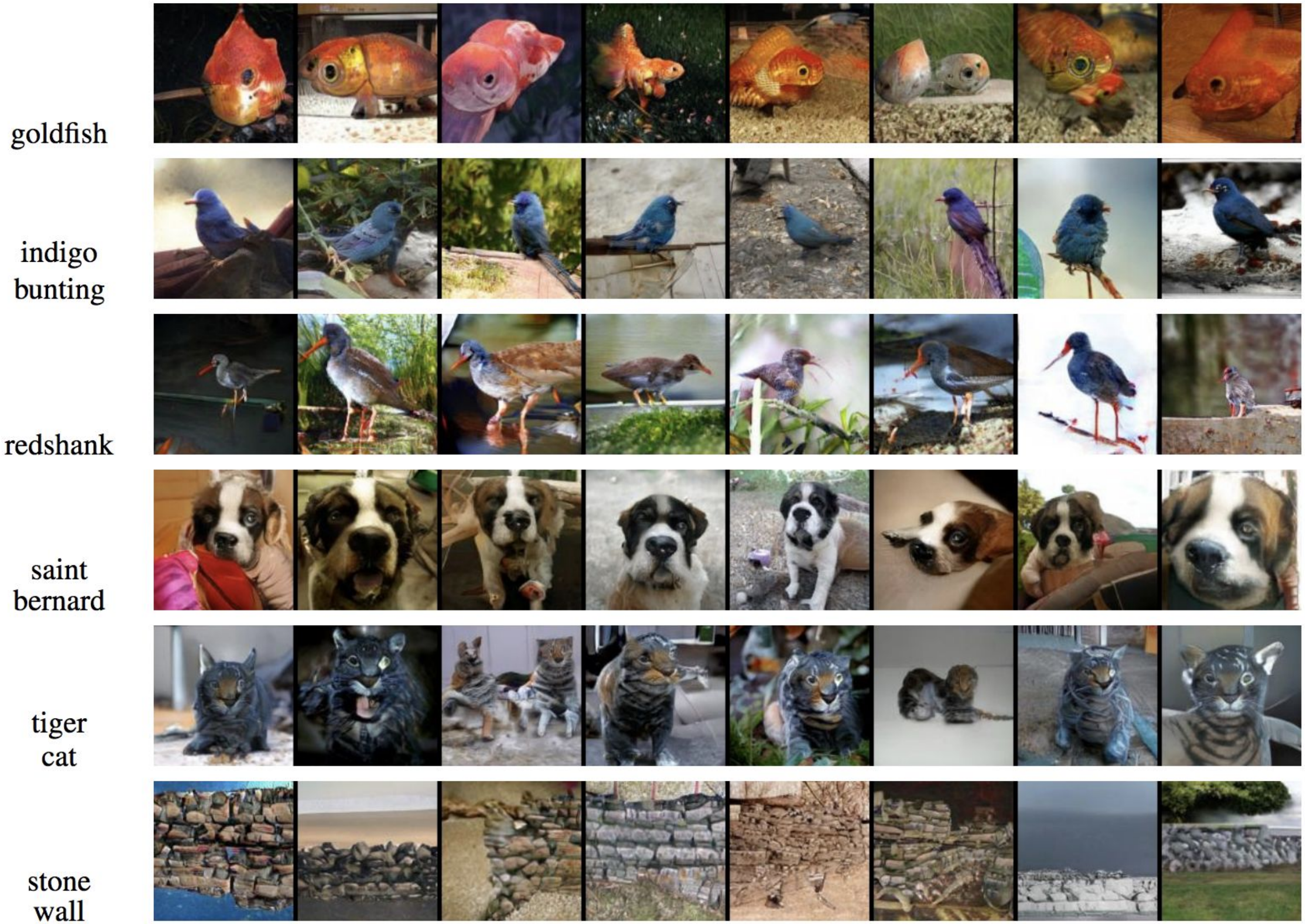
Ian Goodfellow
Google Brain

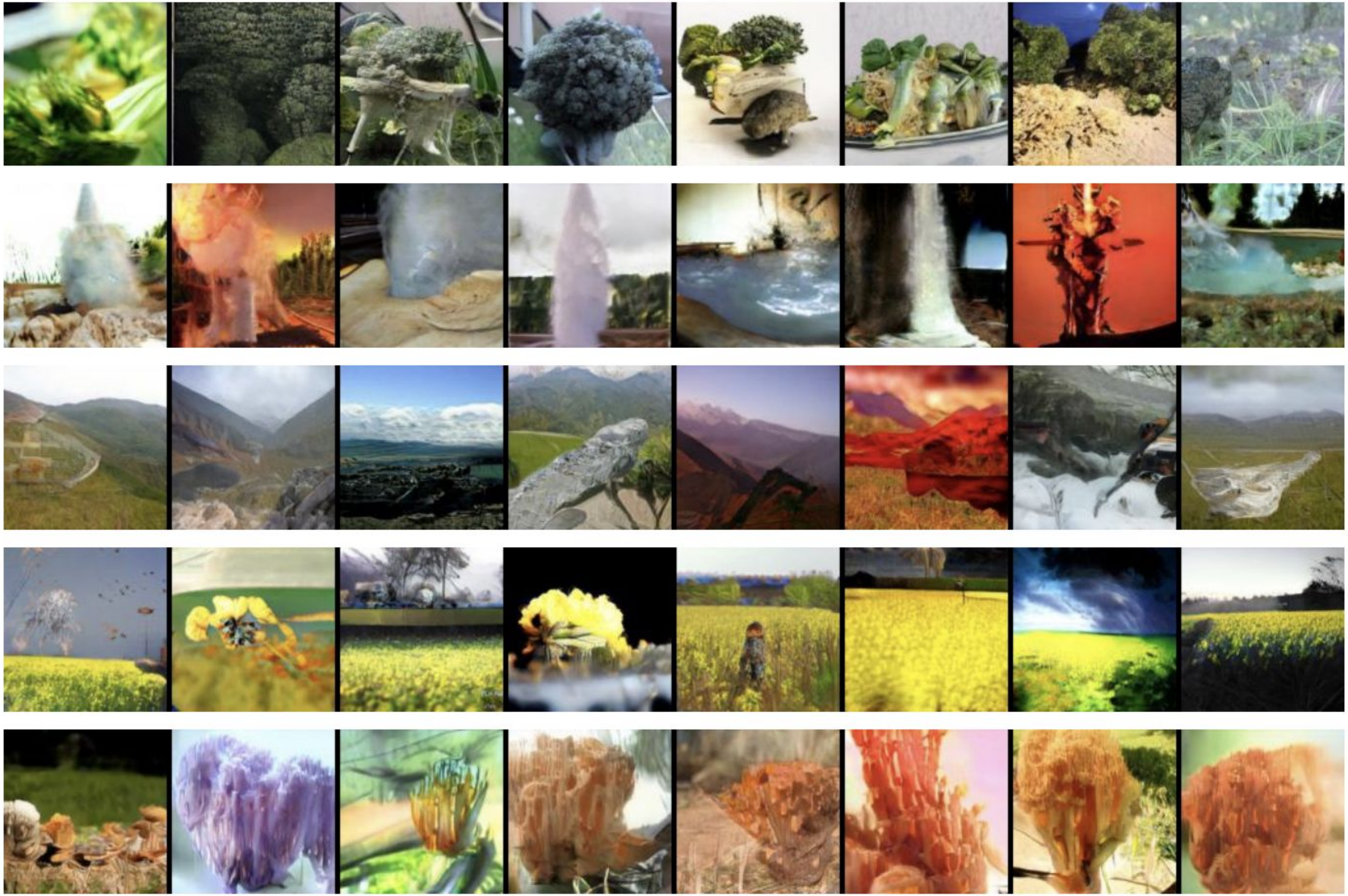
Dimitris Metaxas
Rutgers University

Augustus Odena
Google Brain

Overview

- Introduces self-attention mechanism into convolutional GAN’s
- For image generation tasks
- With spectral normalization
- Achieves state-of-the-art results
    - Frechet Inception distance on ImageNet: from 27.62 to **18.65**
    - Inception score: from 36.8 to **52.52**


goldfish
indigo bunting
redshank
saint bernard
tiger cat
stone wall



broccoli

geyser

valley

rapeseed

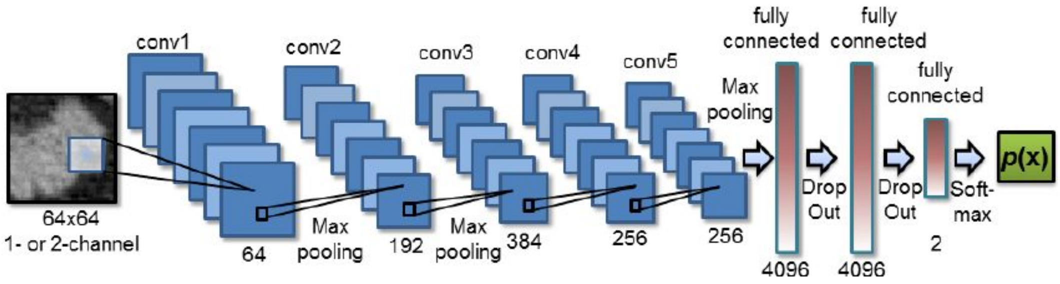
coral fungus

Problems

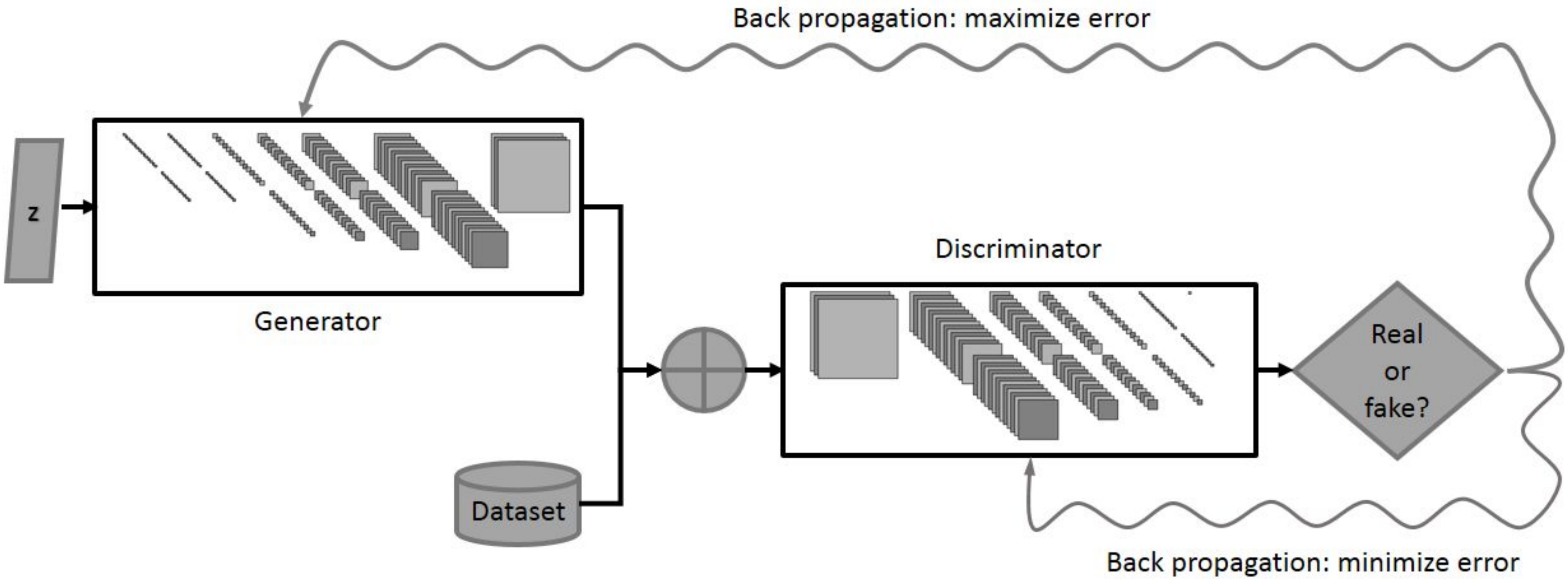
- Existing ImageNet GAN models
  - Good at generating images with few structural constraints (ocean, sky & landscape)
  - Fails at capturing geometric or structural patterns
    - Dogs with good fur, but without two distinct legs

ConvNets & long-range dependencies

- Convolution operator has a local receptive field
- Long range dependencies can only be processed after passing through several conv layers
- Possible reasons for failure to capture:
  - Model is small
  - Optimization algorithms may not be good at capturing cross-layer dependencies
  - Parameterization may be statistically brittle and prone to failure when applied to unseen inputs
- Remedies
  - Increase the size of convolution kernels
    - But it’s slower



GAN

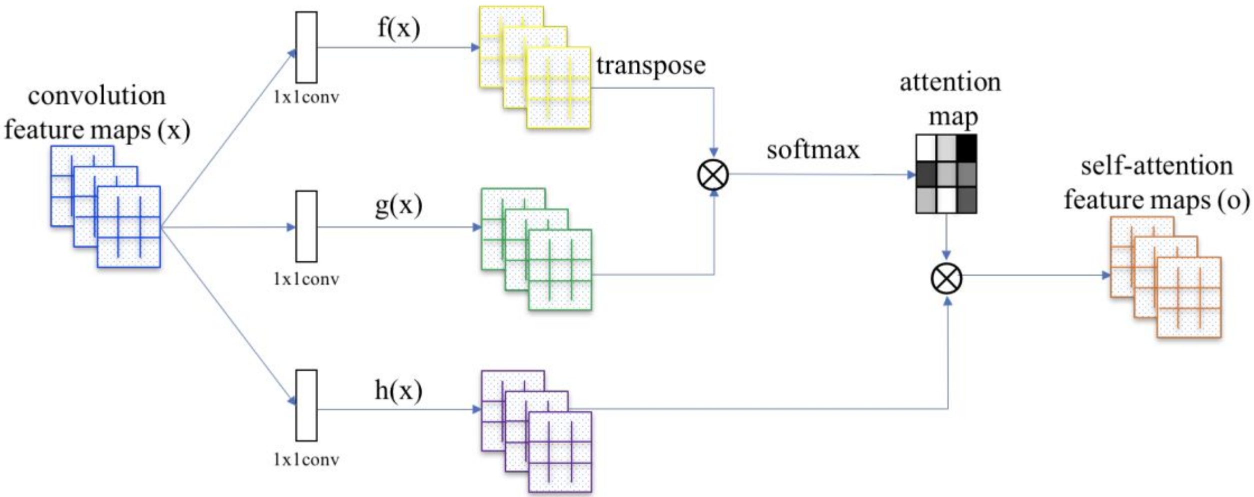


From Prof. Antonio Torralba course slides.
https://github.com/llSourcell/Pokemon_GAN/blob/master/Generative%20Adversarial%20Networks.ipynb

Self-Attention

- Achieves state-of-the-art results in machine translation
- A better balance between long-range dependency modeling and efficiency
  - Response at a position = weighted sum of features at all positions
  - Weights ⇔ attention vectors
    - Small calculation cost
- Could yield more interpretable models
  - Attention heads clearly learn to perform different tasks
- Could be *complementary* to convolutions

Visualization of most attended-regions



$$\boldsymbol{x} \in \mathbb{R}^{C \times N}$$

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{W_f x},\ \boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{W_g x}$$

$$\boldsymbol{W_g} \in \mathbb{R}^{\bar{C} \times C}, \boldsymbol{W_f} \in \mathbb{R}^{\bar{C} \times C}, \boldsymbol{W_h} \in \mathbb{R}^{C \times C}$$

$$\bar{C} = {}^{C}/_{8}$$

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N} \exp(s_{ij})}, \text{where } s_{ij} = \boldsymbol{f}(\boldsymbol{x_i})^T \boldsymbol{g}(\boldsymbol{x_j})$$

$$\boldsymbol{o} = (\boldsymbol{o_1}, \boldsymbol{o_2}, \ldots, \boldsymbol{o_j}, \ldots, \boldsymbol{o_N}) \in \mathbb{R}^{C \times N}$$

$$\boldsymbol{o_j} = \sum_{i=1}^{N} \beta_{j,i} \boldsymbol{h}(\boldsymbol{x_i}), \text{where } \boldsymbol{h}(\boldsymbol{x_i}) = \boldsymbol{W_h x_i}$$

Final output:

$$\boldsymbol{y_i} = \gamma \boldsymbol{o_i} + \boldsymbol{x_i},$$

γ is initialized as 0.

The loss

$$L_D = -\mathbb{E}_{(x,y)\sim p_{data}}[\min(0, -1 + D(x,y))] - \mathbb{E}_{z\sim p_z, y\sim p_{data}}[\min(0, -1 - D(G(z),y))],$$
$$L_G = -\mathbb{E}_{z\sim p_z, y\sim p_{data}} D(G(z), y),$$

Spectral Normalization

- Applied to both generator and discriminator
- Advantage: does not require extra hyper-parameter tuning

Imbalanced Learning Rate

- Regularized discriminator
  - Requires multiple discriminator update steps per generator update step
  - slow
- Mitigation: separate learning rates
  Two-timescale learning rate (TTUR) by Heusel *et al.*
  - Better results given the same wall-clock time

Measurement: Inception Score (IS) & Fréchet Inception distance(FID)

- Inception score
  - Computes KL divergence between conditional class distribution and marginal class dist.
  - Higher IS ⇔ better image quality
  - Widely used, thus making the work comparable
  - The higher the better

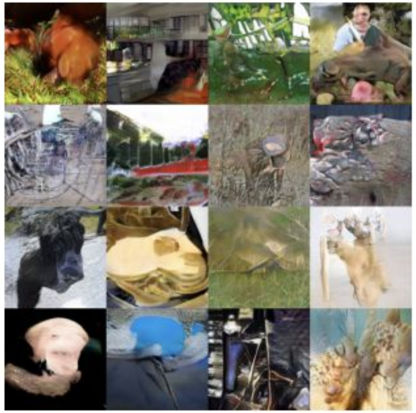
$$\mathrm{IS}(G) = \exp\left(\mathbb{E}_{\mathbf{x}\sim p_g} D_{KL}(\, p(y|\mathbf{x}) \,\|\, p(y)\,)\,\right)$$

- FID
  - More principled and comprehensive metric
  - Wasserstein-2 distance between generated and real images in feature space of Inception-v3
  - More consistent with human evaluation
  - The lower the better

Experiment

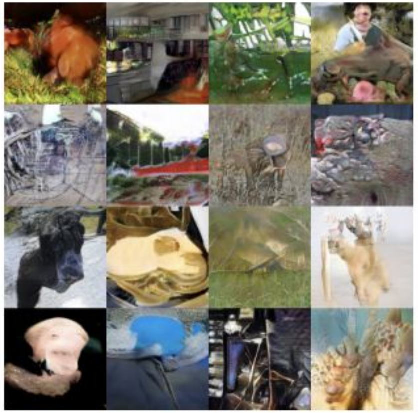
- Generate 128 x 128 images
- Spectral normalization used for both generator and discriminator
- Adam optimizer
  - Learning rate for discriminator is 0.00004
  - Learning rate for generator is 0.0001

Baseline: SN on D
(10k, FID=181.84)

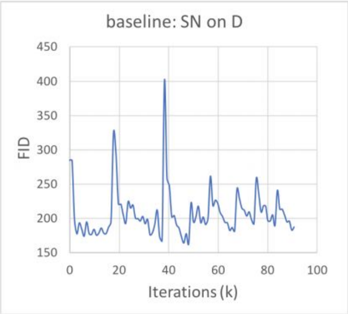


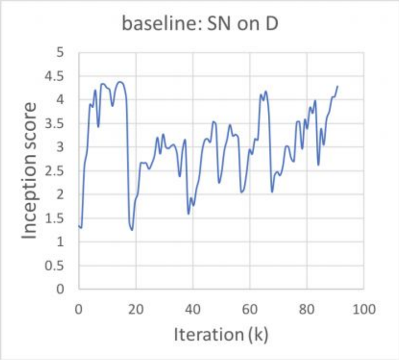
SN on G/D
(10k, FID=93.52)

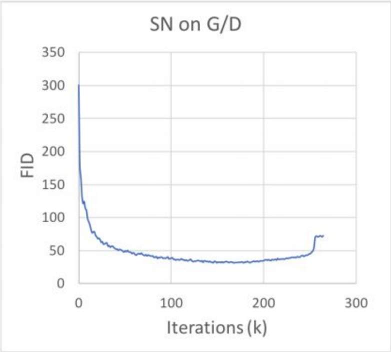
SN on G/D
(160k, FID=33.39)

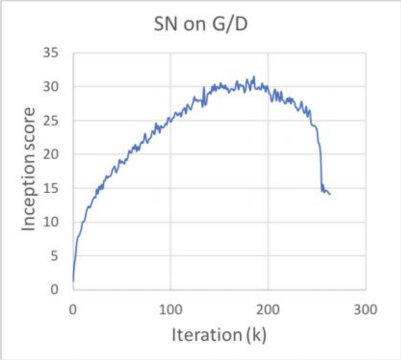


SN on G/D
(260k, FID=72.41)

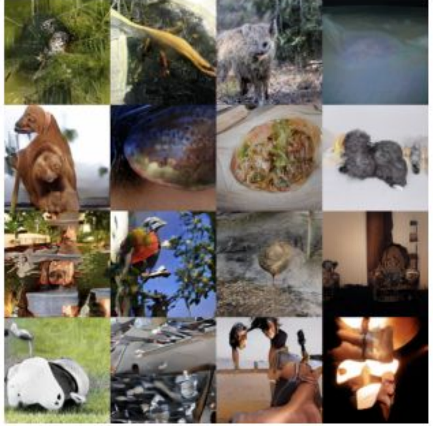




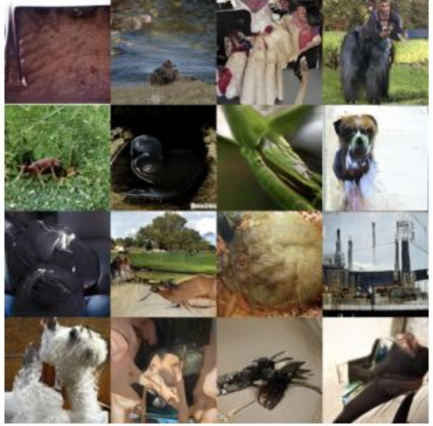




Note: Using 1 : 1 balanced updates on G & D.



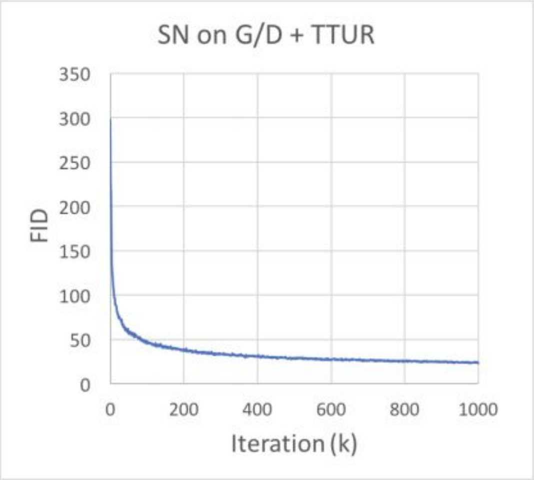
SN on G/D+TTUR
(10k, FID=99.04)

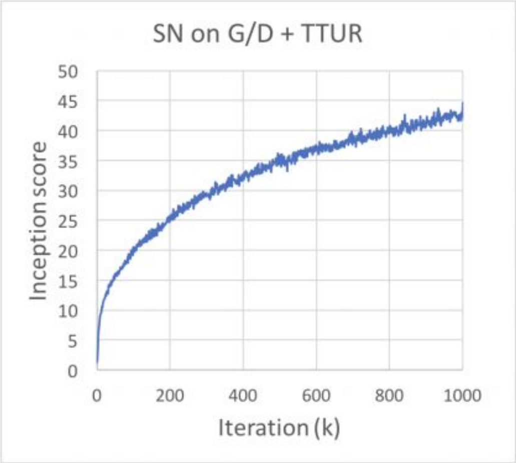


SN on G/D+TTUR
(160k, FID=40.96)

SN on G/D+TTUR
(260k, FID=34.62)

SN on G/D+TTUR
(1M, FID=22.96)





Model	no attention	SAGAN				Residual			
		$feat_8$	$feat_{16}$	$feat_{32}$	$feat_{64}$	$feat_8$	$feat_{16}$	$feat_{32}$	$feat_{64}$
FID	22.96	22.98	22.14	**18.28**	18.65	42.13	22.40	27.33	28.82
IS	42.87	43.15	45.94	51.43	**52.52**	23.17	44.49	38.50	38.96

- All models have been trained over one million iterations
- Self-attention added to different stages of the g and d.
- Attention put onto middle-to-high levels yields better results

Comparison with state-of-the-art

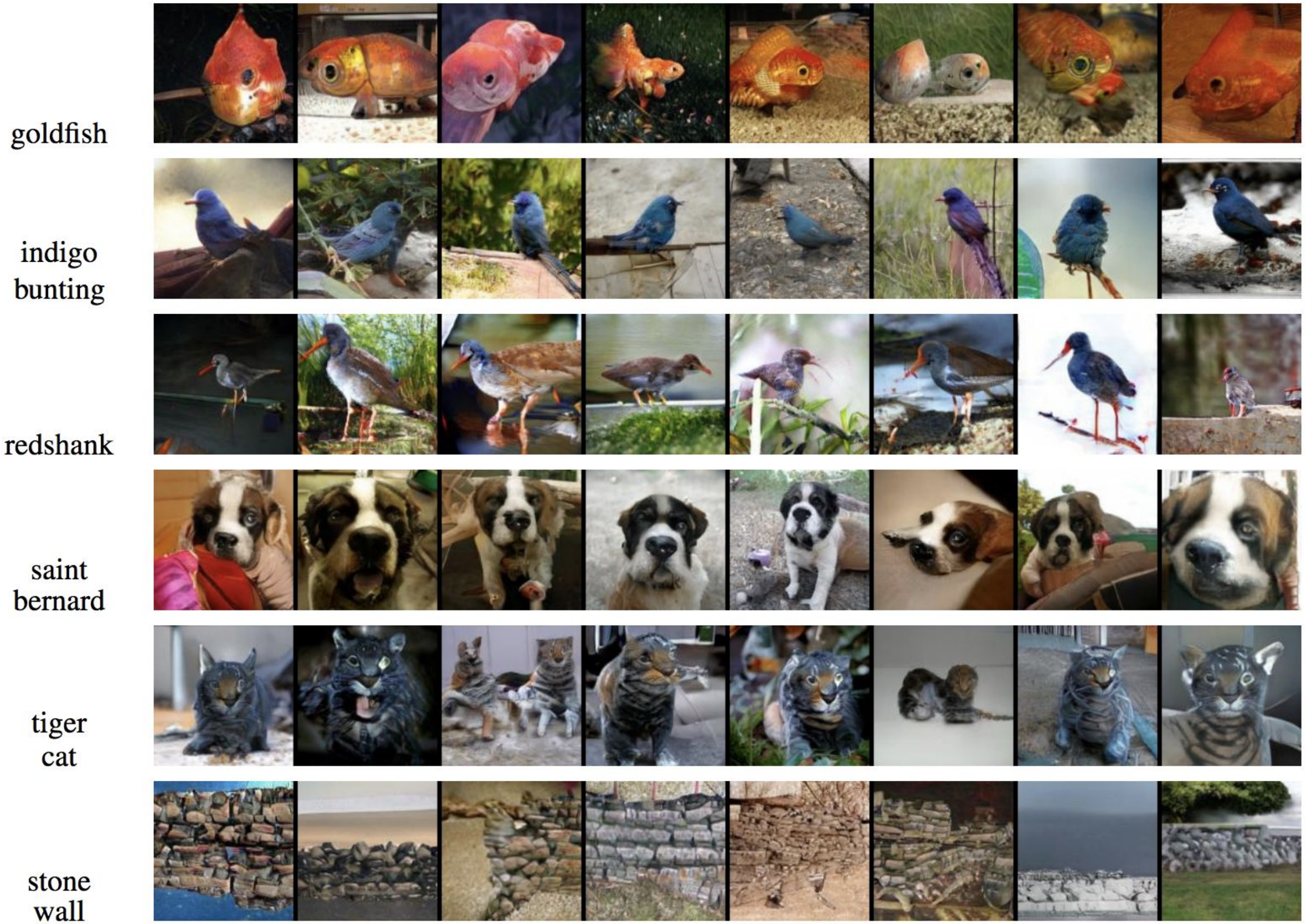
Model	Inception Score	FID
AC-GAN [31]	28.5	/
SNGAN-projection [17]	36.8	27.62*
SAGAN	**52.52**	**18.65**

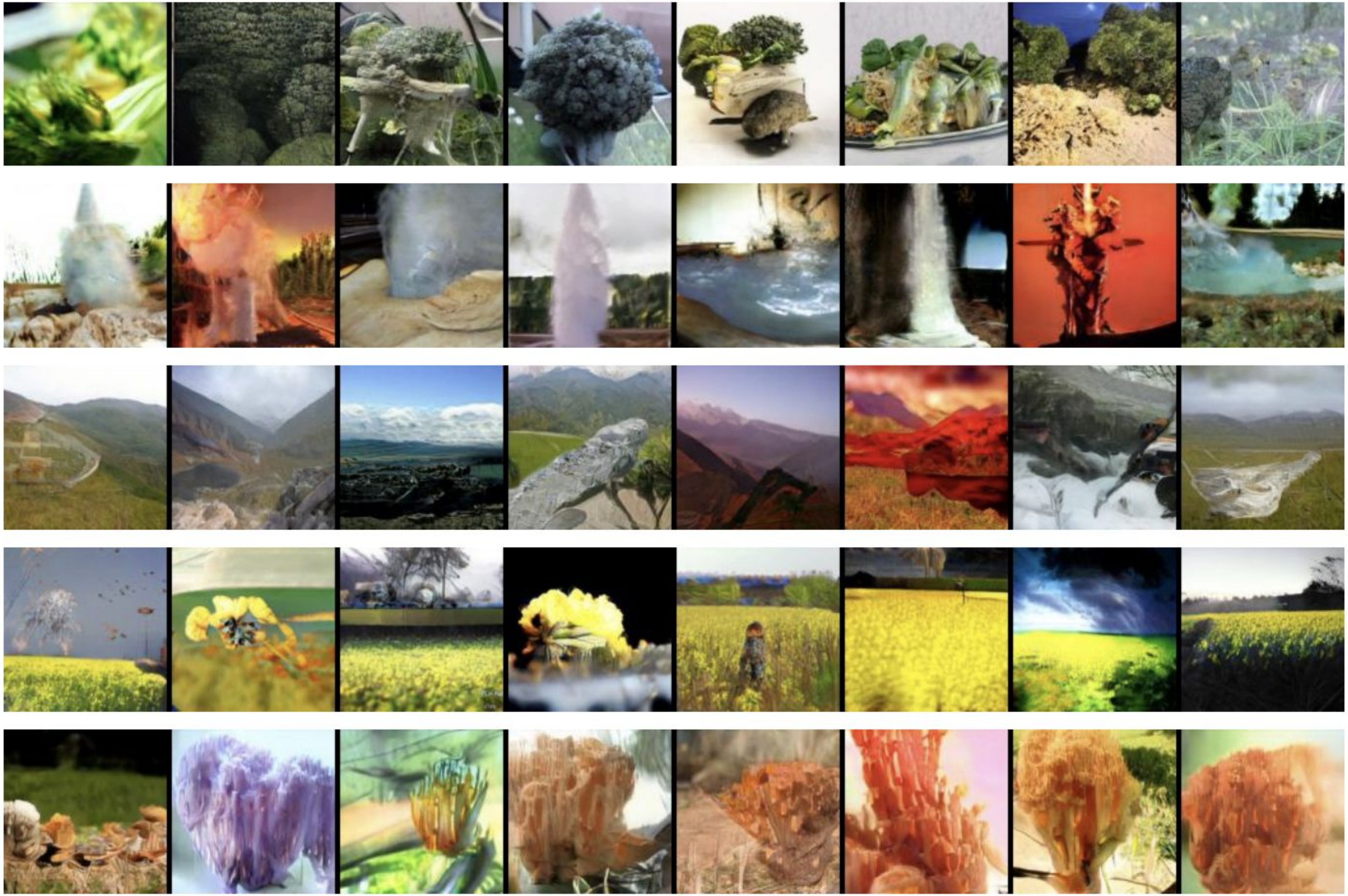
- Takeru Miyato, Masanori Koyama. ***cGANs with Projection Discriminator***. ICLR2018.

Previous State-of-the-art (Miyato, et al.)

See more at https://github.com/pfnet-research/sngan_projection/#other-materials

Previous State-of-the-art (Miyato, et al.)


goldfish
indigo bunting
redshank
saint bernard
tiger cat
stone wall



broccoli

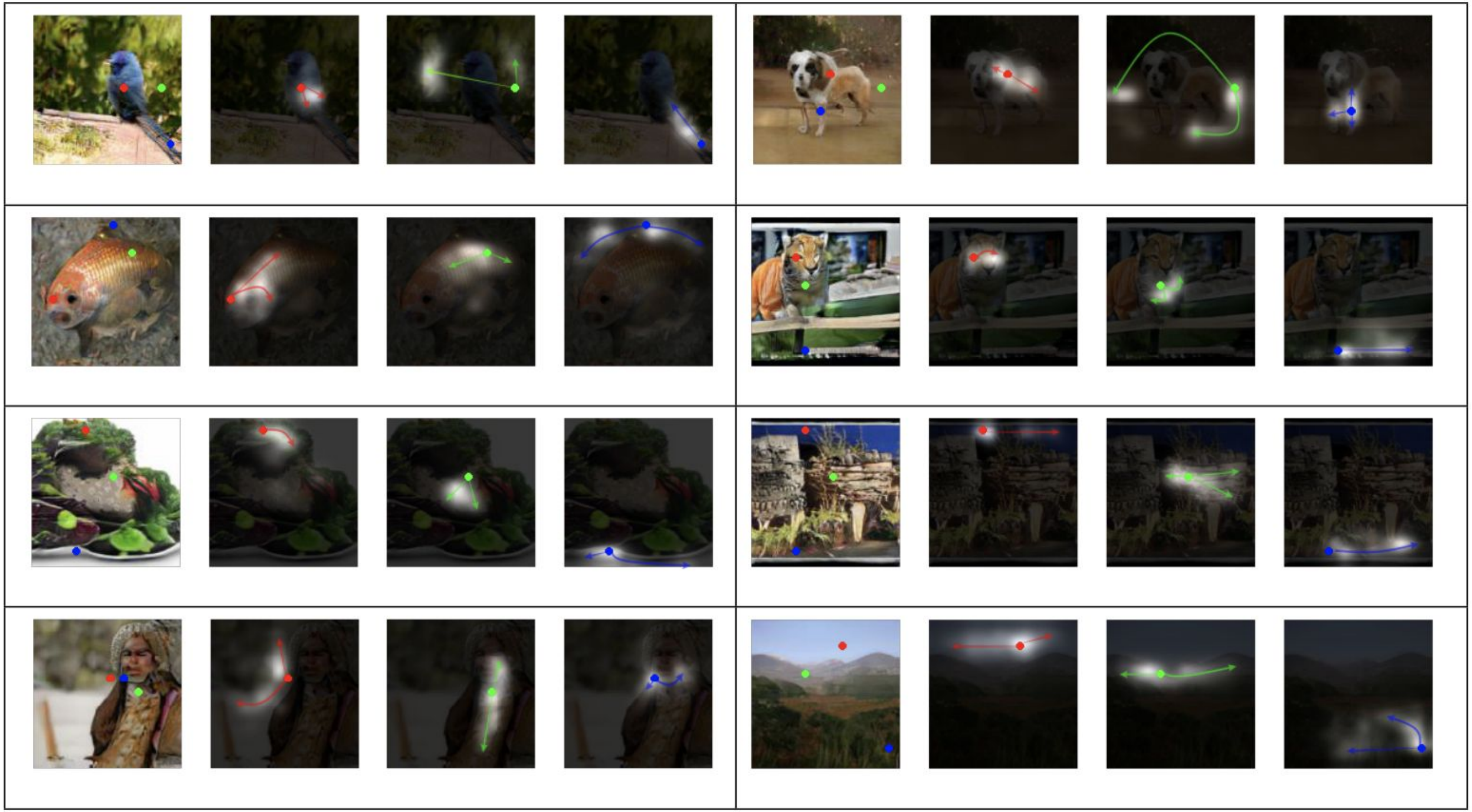
geyser

valley

rapeseed

coral fungus

Attention map visualization



References

- I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. **Generative adversarial nets**. In NIPS, 2014
- I.J. Goodfellow, **NIPS 2016 Tutorial: Generative Adversarial Networks**
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. **Attention is all you need**. arXiv:1706.03762, 2017
- Takeru Miyato, Toshiki Kataoka, Masanori Koyama, Yuichi Yoshida. ***Spectral Normalization for Generative Adversarial Networks***. ICLR2018.
- M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. **Gans trained by a two time-scale update rule converge to a local nash equilibrium**. In NIPS, pages 6629–6640, 2017.
- X. Wang, R. Girshick, A. Gupta, and K. He. **Non-local neural networks**. In CVPR, 2018
- Takeru Miyato, Masanori Koyama. **cGANs with Projection Discriminator**. ICLR2018.