# On gradient regularizers for MMD GANs

Michael Arbel[1*], Dougal J. Sutherland[1*], Mikolaj Bińkowski[2] and Arthur Gretton[1]

[1]Gatsby Computational Neuroscience Unit, University College London [2]Department of Mathematics, Imperial College London

[*]Equal contribution

Imperial College London

UCL

## Overview

- ✓ MMD-based losses for implicit generative models are effective and principled.
- × Previous approaches have bad topological properties.
- ✓ We introduce gradient-regularized MMD loss with better topology.
- ✓ New insight on the desired properties for the discriminator network.
- ✓ State-of-the-art results on $64 \times 64$ unconditional ImageNet and $160 \times 160$ CelebA.

## Integral Probability Metrics

Integral Probability Metrics (IPMs) are distances between distributions defined by a class of critic functions $\mathcal{F}$:

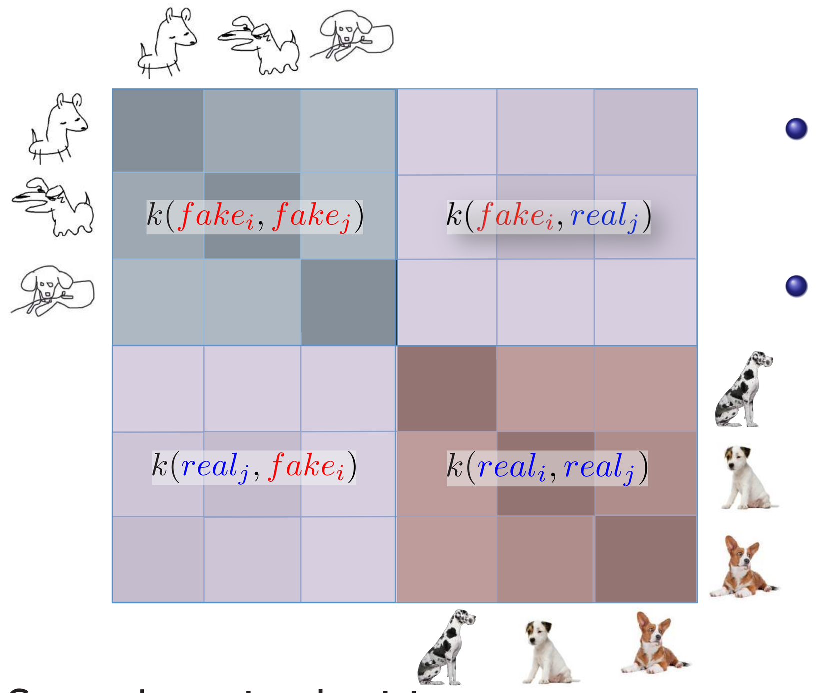
$$\mathcal{D}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)]$$

- **1-Wasserstein distance**: $\mathcal{F}$ is the set of 1−Lipschitz functions

$$\mathcal{F} = \{f : |f(x) - f(y)| \leq \|x - y\|, \forall x, y\}$$

WGANs approximate $f$ with a critic network $\phi_\psi$. Weight clipping [1] or gradient penalty [4] used to make $\phi_\psi$ approximately Lipschitz.

- **Maximum Mean Discrepancy (MMD)** has $\mathcal{F}$ a unit ball in a *Reproducing Kernel Hilbert Space (RKHS)* $\mathcal{H}$ with kernel $k$:



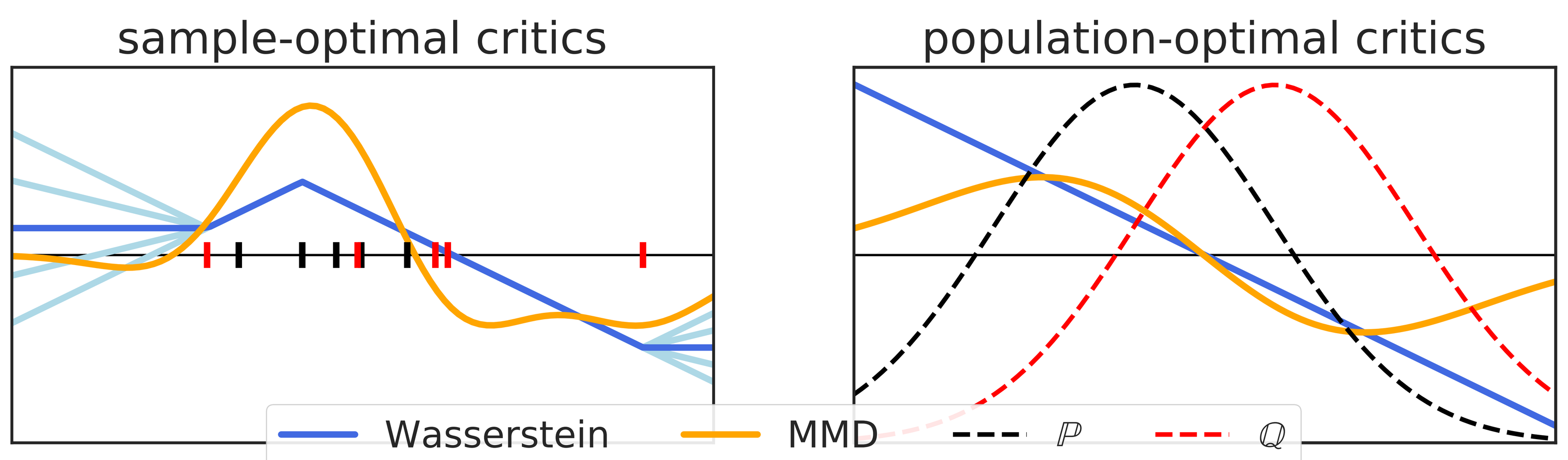

- Closed form solution:

$$f^*(t) \propto \mathbb{E}_{\mathbb{P}}[k(X,t)] - \mathbb{E}_{\mathbb{Q}}[k(Y,t)]$$

- Unbiased estimator:

$$\widehat{\mathrm{MMD}}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(real_i, real_j) + k(fake_i, fake_j) - \frac{2}{n^2} \sum_{i,j} k(real_i, fake_j)$$

Smooth optimal critic:

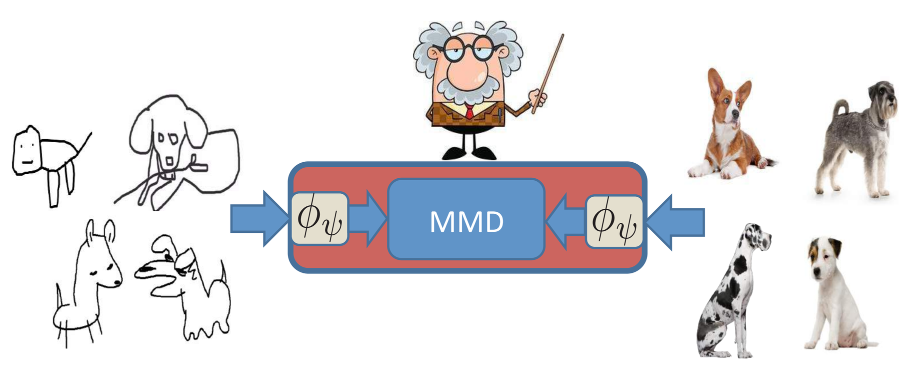



## Maximum Mean Discrepancy for GANs

MMD GANs optimize critic in kernel:

$$k_\psi(x, y) = k_{\text{base}}(\phi_\psi(x), \phi_\psi(y))$$

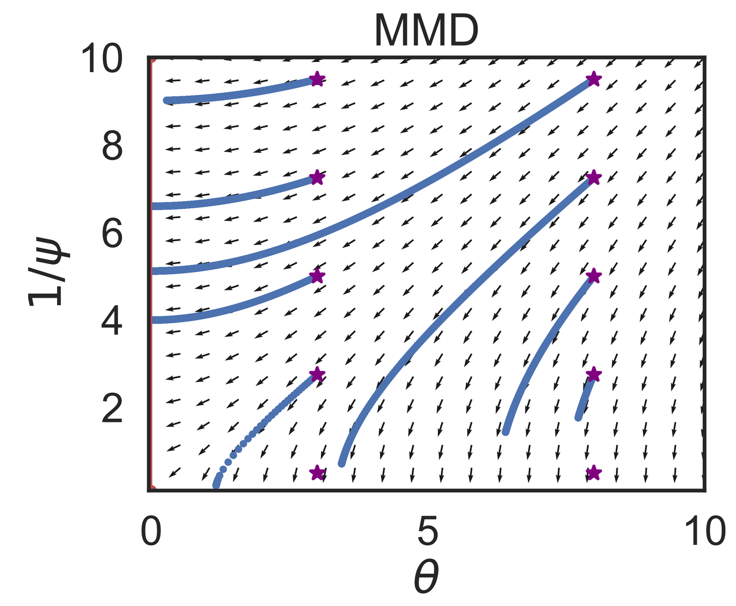
$$\inf_\theta \sup_\psi \underbrace{\mathrm{MMD}^2_{k_\psi}(\mathbb{P}, \mathbb{Q}_\theta)}_{\mathcal{D}_{\mathrm{MMD}}(\mathbb{P}, \mathbb{Q}_\theta)}$$

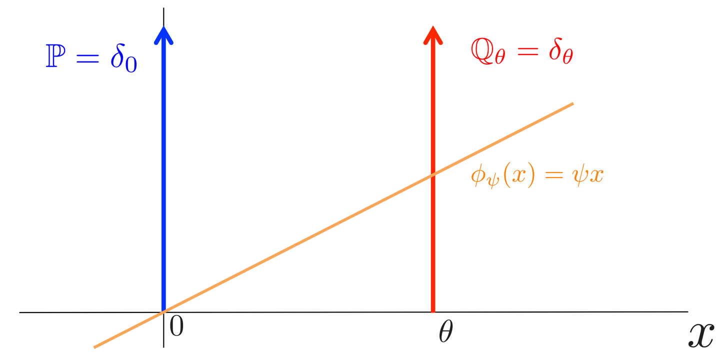



Can also use gradient penalty [2].

## Continuity under weak topology

$\mathcal{D}_{\mathrm{MMD}}$ not continuous / differentiable in general:




DiracGAN [7]

## Gradient Constrained MMD

- Adjust the radius of the RKHS ball according to the smoothness of $k$:

$$\mathcal{F}_S = \{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq \sigma_k\}$$

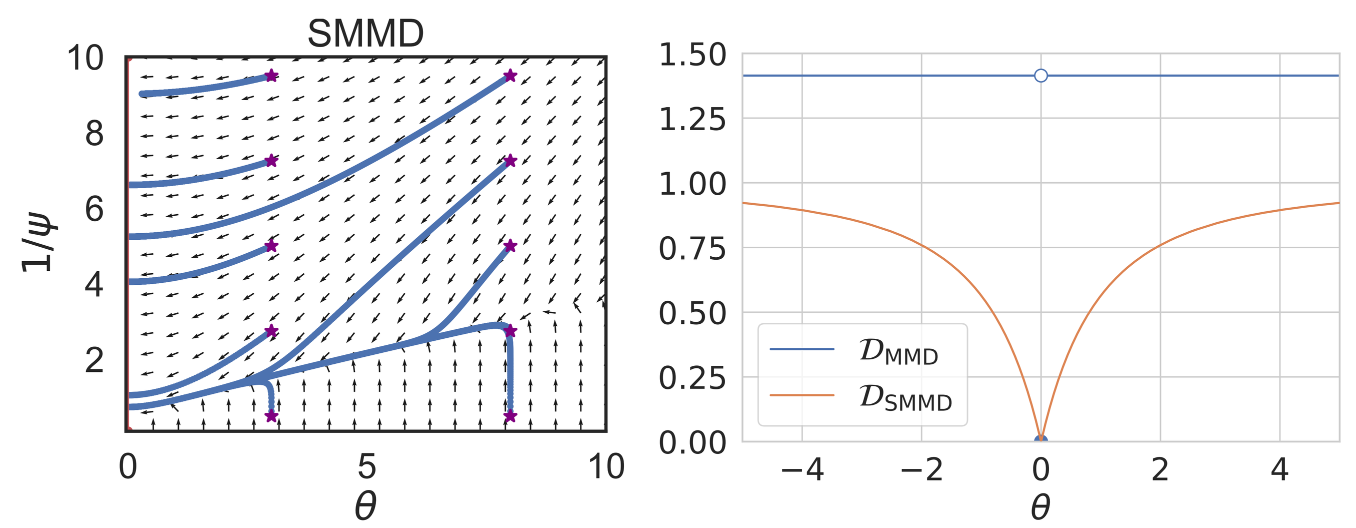
$$\mathrm{SMMD}_k(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathcal{F}_S} \mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(X)] = \sigma_k \,\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q})$$

$$\sigma_k := \left( \lambda + \mathbb{E}_{X\sim\mathbb{S}} \left[ k(X,X) + \sum_{i=1}^d \left. \frac{\partial^2 k(y,z)}{\partial y_i \partial z_i} \right|_{(y,z)=(X,X)} \right] \right)^{-\frac{1}{2}}$$

- Optimal $f^*$ satisfies $\mathbb{E}_{X\sim\mathbb{S}}[\|\nabla f^*(X)\|^2] \leq 1$
- Other possible choices for $\mathcal{F}$:

$$\mathcal{F}_{Lip} := \{f \in \mathcal{H}_k : \|f\|^2_{Lip} + \lambda\|f\|^2 \leq 1\} \quad \mathcal{F}_{GC} := \{f \in \mathcal{H}_k : \|f\|^2_{L_2(\mathbb{S})} + \|\nabla f\|^2_{L_2(\mathbb{S})} + \lambda\|f\|^2 \leq 1\}$$

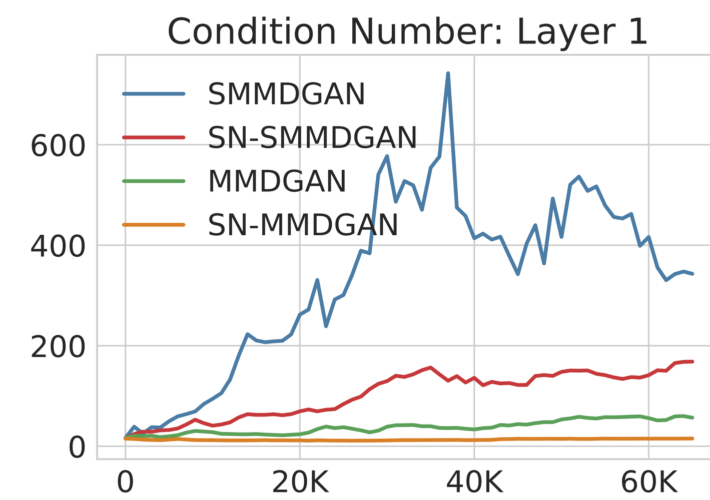
| $\mathcal{F}$ | $\mathcal{F}_{Lip}$ | $\mathcal{F}_{GC}$ | $\mathcal{F}_S$ |
|---|---|---|---|
| Effectiveness | 😎 | 🙂 | 🙂 |
| Tractability | 🙁 | 😐 | 😎 |




## Theory: Continuity under weak topology

$\mathcal{D}_{\mathrm{SMMD}}(\mathbb{P}, \mathbb{Q})$ is continuous in weak topology if:

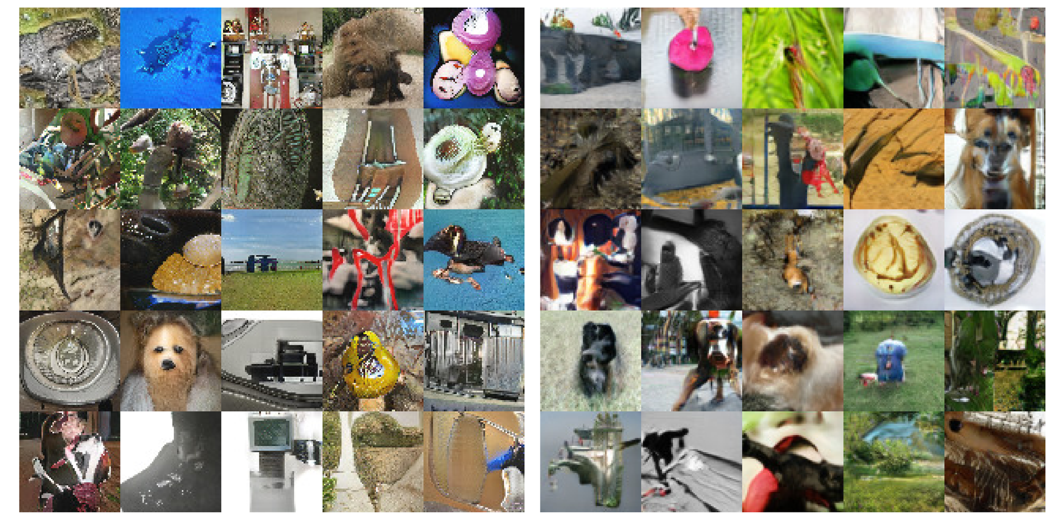
- $\mathbb{S}$ has a density (can depend on $\mathbb{P}$, $\mathbb{Q}$)
- $\phi_\psi$ is fully connected, Leaky-ReLU activations, non-increasing width
- Each layer of $\phi_\psi$ has weights with bounded condition number
- $k_{\text{base}}$ is "reasonable" (Gaussian, linear, ...)




Orthogonal Normalization [3] or Spectral Normalization [8] control the condition number in practice.

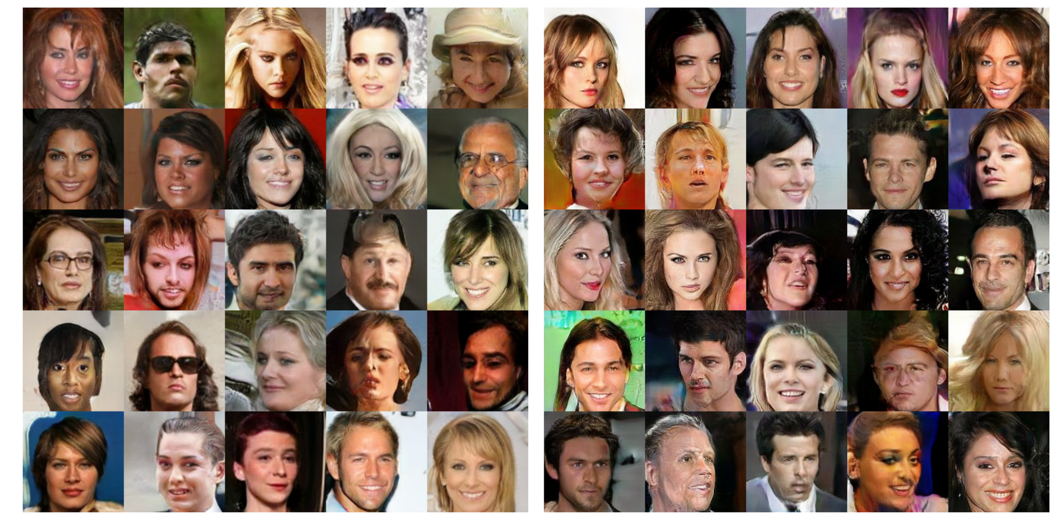
## Experimental Comparison

Scaled MMD GANs outperform other GANs (WGAN-GP, MMD-GAN, SN-GAN).



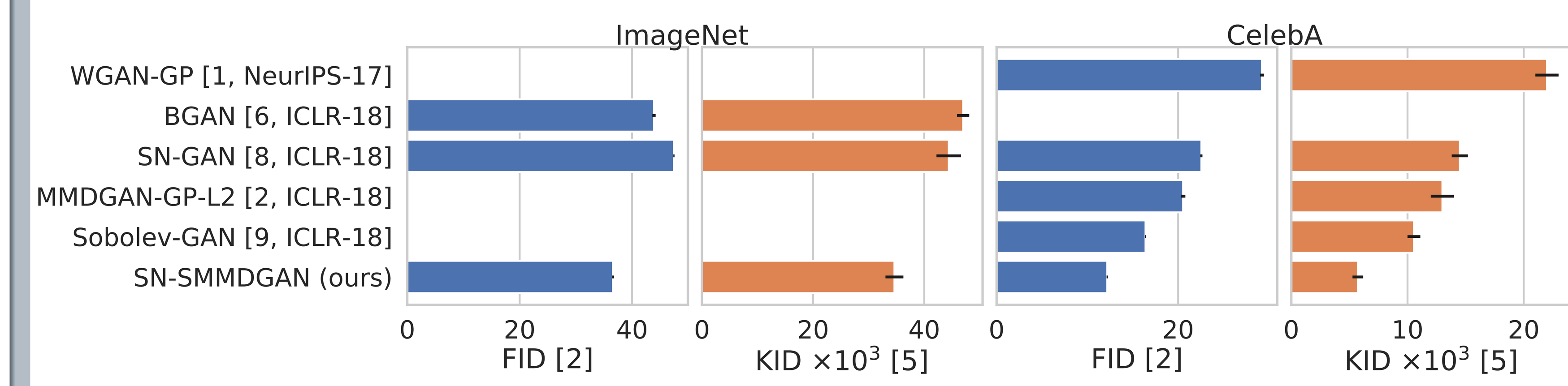
(a) Scaled MMD GAN, SN (b) SN-GAN

**ImageNet**, $64 \times 64$.
**No labels**.
Generator: 10-layer ResNet.
Critic: 10-layer ResNet.



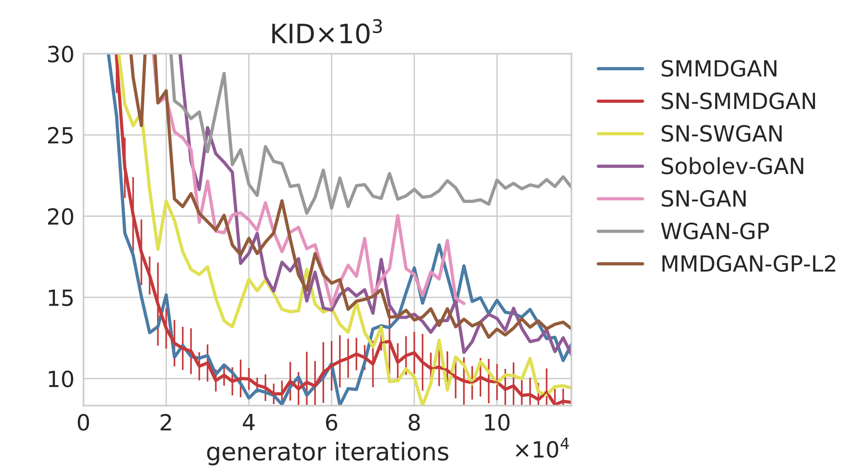
(a) Scaled MMD GAN, SN (b) MMD GAN, GP+L2

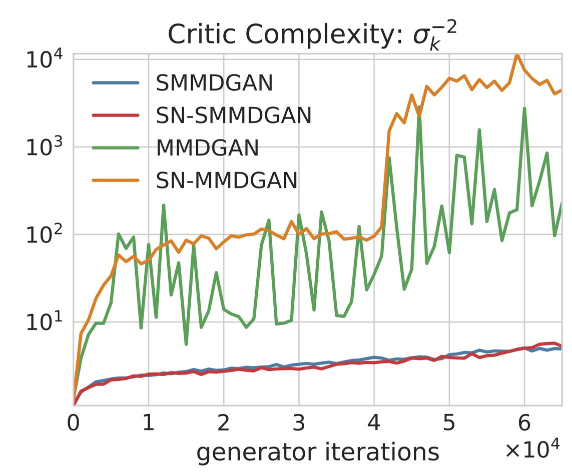
**CelebA**, $160 \times 160$.
Generator: 10-layer ResNet.
Critic: 5-layer DCGAN.




Implementation at `github.com/MichaelArbel/Scaled-MMD-GAN`

## Faster training and better complexity control




## Bibliography

M. Arjovsky, S. Chintala, and L. Bottou. "Wasserstein Generative Adversarial Networks". *ICML*. 2017.
M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. "Demystifying MMD GANs". *ICLR*. 2018.
A. Brock, T. Lim, J. M. Ritchie, and N. Weston. "Neural Photo Editing with Introspective Adversarial Networks". *ICLR*. 2017.
I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. "Improved Training of Wasserstein GANs". *NeurIPS*. 2017.
M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter. "GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium". *NeurIPS*. 2017.
R. Hjelm, A. Jacob, T. Che, A. Trischler, K. Cho, and Y. Bengio. "Boundary-Seeking Generative Adversarial Networks". *ICLR*. 2018.
L. Mescheder, A. Geiger, and S. Nowozin. "Which Training Methods for GANs do actually Converge?" *ICML*. 2018.
T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. "Spectral Normalization for Generative Adversarial Networks". *ICLR*. 2018.
Y. Mroueh, C.-L. Li, T. Sercu, A. Raj, and Y. Cheng. "Sobolev GAN". *ICLR*. 2018.

This work was funded by the Gatsby Charitable Foundation.

Contact: `michael.n.arbel@gmail.com`